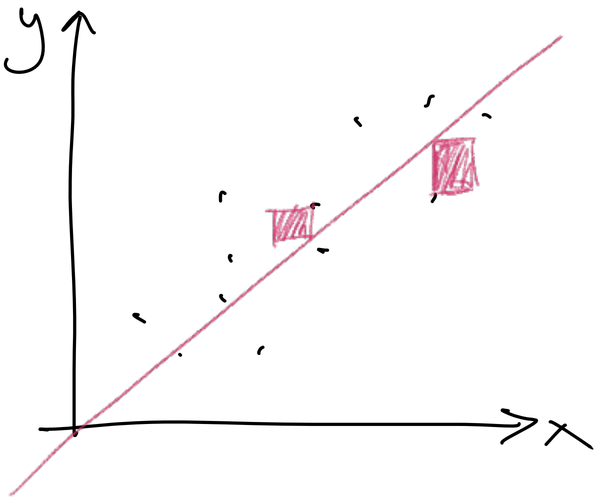
ISTT100x, uke 9 fellesmodul, 2024

Uke 1:

n observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$



$y = a + bx$

finnes ved å minimere kvadratavvik

↳ veldig generelt verktøy

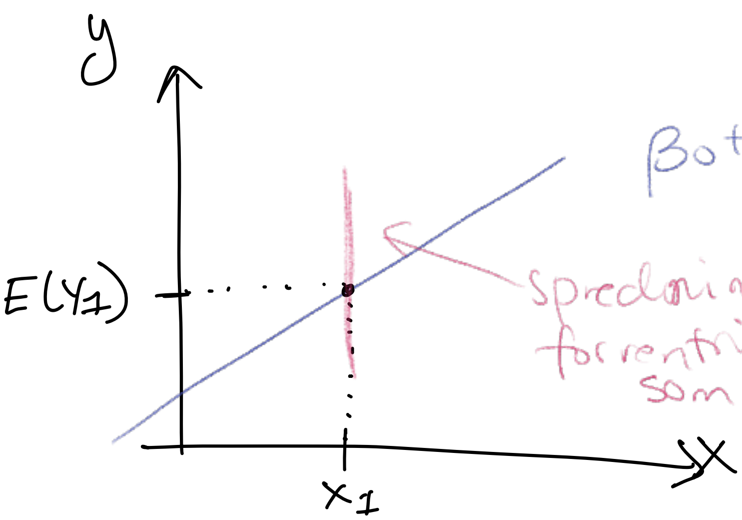
Uke 9: Y-ene er stokastiske!

n gitte/fikserte verdier av x:

$$\begin{array}{cccc} x_1, & x_2, & \ldots, & x_n \\ \downarrow & \downarrow & & \downarrow \\ Y_1 & Y_2 & & Y_n \end{array}$$

n stokastiske variabler

} Dvs. dersom vi gjentok forsøket igjen med samme verdier av x, ville vi fått ulike verdier av y-ene.



$\beta_0 + \beta_1 x$

Spredning rundt forventningsverdi som følge av en normalfordeling med standardavvik σ

Modell:

(1) $$Y|x \sim N(\underbrace{\beta_0+\beta_1 x}_{\mu},\ \underset{\sigma}{\underset{\downarrow}{\sigma}})$$

(2) $$Y = \underbrace{\beta_0+\beta_1 x}_{\substack{\text{konstantledd} \\ (E(Y))}} + \underset{\substack{\uparrow \\ \text{støyledd} \\ (\text{Var}(Y))}}{\varepsilon} \qquad \varepsilon \sim N(0, \sigma)$$

NB: 3 parametere!

β_0, β_1, σ

Som i de to forrige ukene snakker vi da om estimater og estimatorer, konfidensintervaller og hypotesetester.

Eks 1

Modell

antall bokstaver

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$$

vekt av twist

støy ledd

fordi vi vet at f.eks to twist av samme type kan ha ulik vekt.

er $\beta_1 \neq 0$?

Minste kvadratsum:

$$y = 5.42 + 0.17x$$

FIGUR

Eks: $x = 5$ (banan)

$y = 6.27$ gram

$x = 15$ (chocolate toffee)

$y = 7.97$ gram

Såhar jeg 'bevis' for påstanden min?

(Enkel) lineær regresjon

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma)$$

Stokastisk   gitt/bestemt

$i = 1, \dots, n$ uavhengige stykkevis

* 3 ukjente parametere

[FORMELARK]

Estimatorer for β_1, β_0 og σ^2:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$S^2 = \frac{1}{n-2} \sum (Y_i - \hat{Y}_i)^2$$

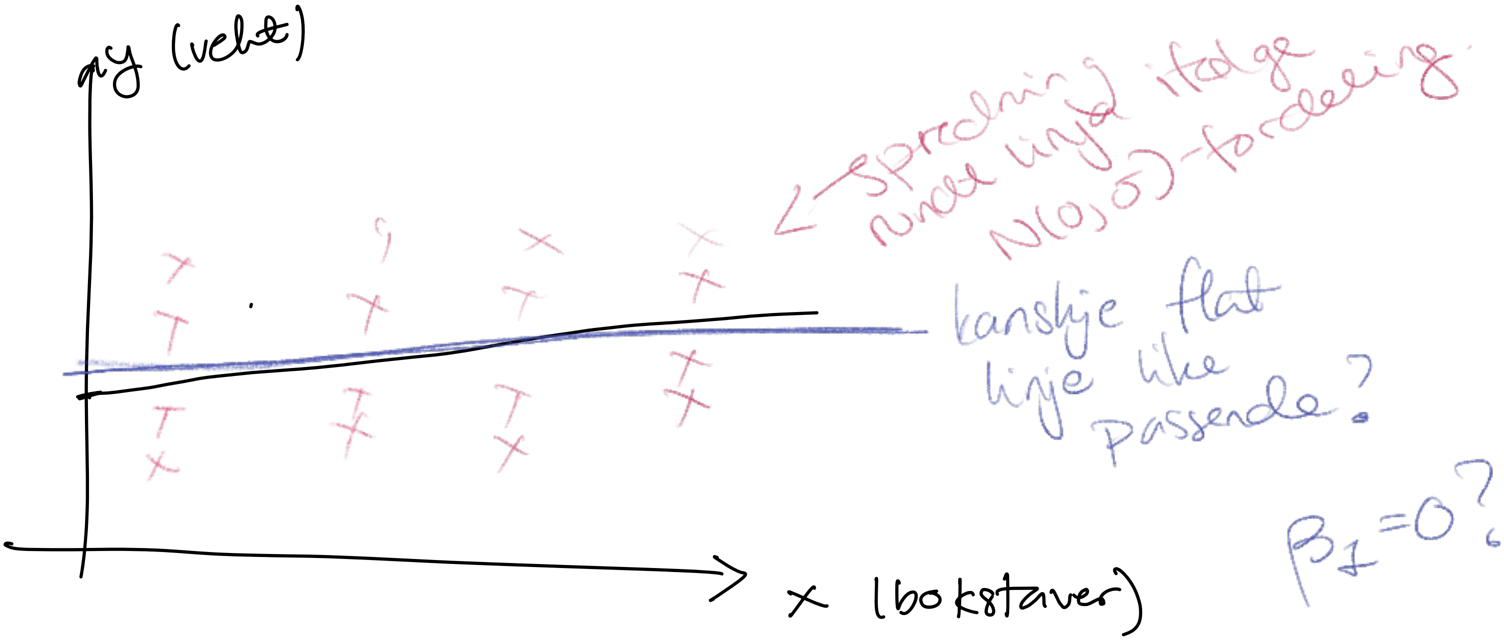
* Forventningsrette for β_1, β_0 og σ^2

Estimater basert på vårt utvalg

$$\hat{\beta}_1 = 0.17g \qquad \hat{\beta}_0 = 5.42g \qquad s^2 = 1.35\, g^2$$

$$s = 1.16\, g.$$

lite stigning   stor spredning



[FORMELARK]

$H_0: \beta_1 = 0 \qquad H_1: \beta_1 \neq 0 \qquad \alpha = 0.05$

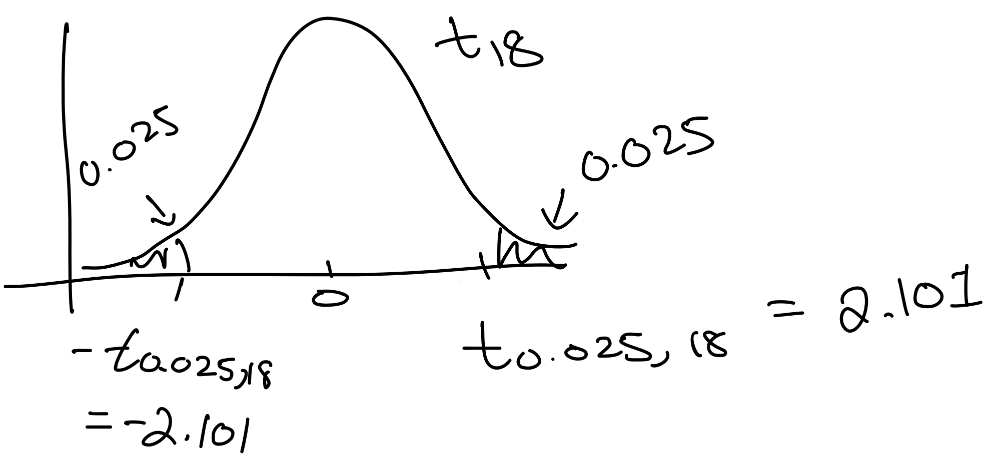
$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

H_0 sann

nb! $n-2$ i enkel lin. reg.

$$SE(\hat{\beta}_1) = S \Big/ \sqrt{\sum (x_i - \bar{x})^2}$$

$$S = \sqrt{\frac{1}{n-2} \sum (y_i - \hat{y}_i)^2}$$



Forkast dersom
$t_{obs} > 2.101$ eller
$t_{obs} < -2.101$

↑ Denne testen utføres 'automatisk' når vi tilpasser regresjonslinje med Python (og R, excel, osv.)

Python, modell-sammendrag

Finn

1) Est. β_0 og β_1

$\hat{\beta}_0 = 5.42 \qquad \hat{\beta}_1 = 0.17$

2) Standardfeil til estimatorene

$SE(\hat{\beta}_0) = 0.622$

$SE(\hat{\beta}_1) = 0.071$

3) t-test for β_1

$t_{obs} = 2.36$ ← fra $\frac{0.1676}{0.071}$

Husk ↑ Forkast H_0 siden $t_{obs} > 2.101$. (evt p-verdi < 0.05)

4) . 95% KI for β_1 $\underbrace{[0.018, 0.317]}$

innehold ikke 0,
kan dermed ikke fastslå at
$\beta_1 \neq 0$.

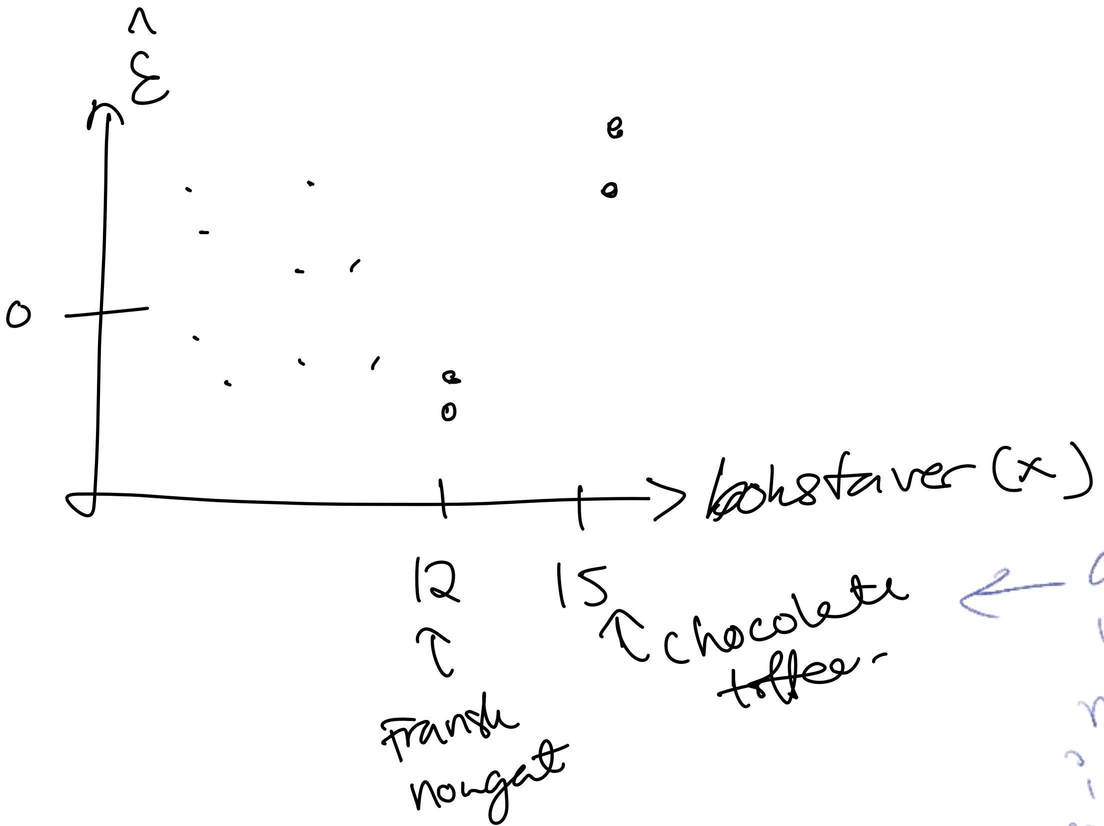
NB: Linja for 'intercept' gjelder β_0,
og selv om vi ikke har formler for
denne på formelarket må
dere likevel kunne bruke disse
resultatene.

Plott, plott, plott.

Residualer: $\hat{\varepsilon}_i = y_i - \hat{y}_i$

bør oppføre seg normalfordelte
rundt 0, og f.eks samme
variasjon uansett x-verdi.

Ser ok ut for
bokstaver < 10,
litt rart
etter dette...



← dersom utvalget mitt på 20 ikke inneholdt disse ville linja sett helt annerledes ut.

Konklusjon:

En slags trend, men dataene ser ut til å bryte med ~~fordel~~ antagelsene i linear regresjon, og skal derfor være forsiktige med å tro på hypotesetest på β_f.

Samvariasjon eller kausalitet?

Eks 2: π (lineær?)

$$V = \frac{4}{3}\pi r3$$

hva er π?!

- har målinger av radius
- har usikre målinger av volum

Regresjonsmodell

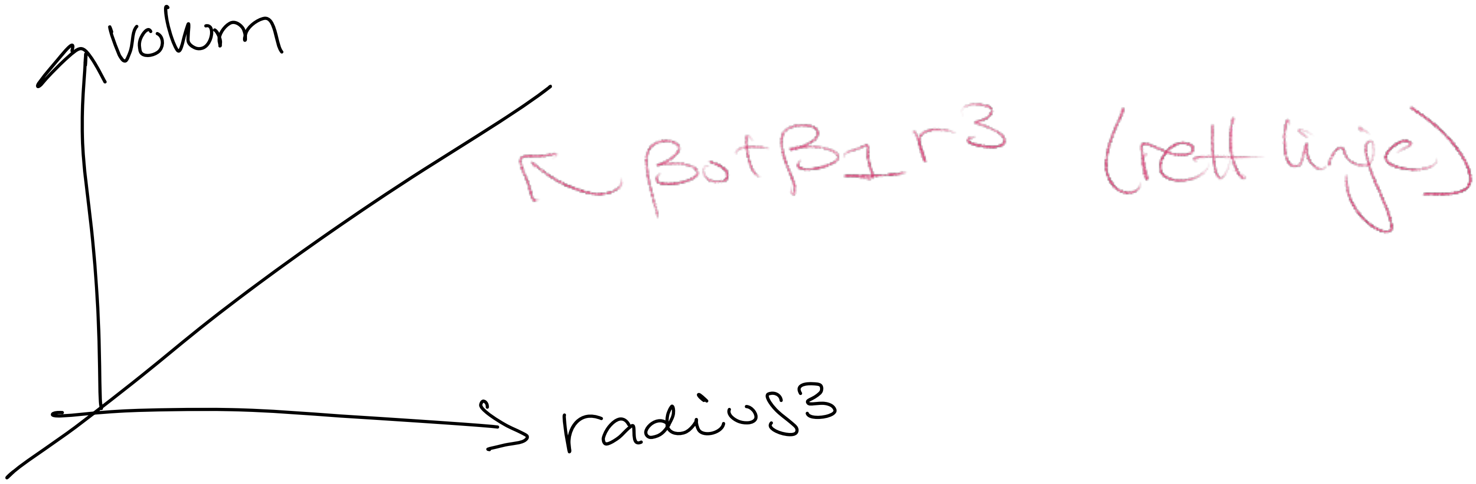
$\frac{4}{3}\pi$

$$V = \beta_0 + \beta_1 r^3 + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$$

bør være 0

ser ikke så veldig lineær ut?

NB: modellen er ikke-lineær mtp radius, men lineær i β-ene!



Modell-sammendrag

- Finn estimert regresjonslinje

  $\hat{V} = -0.42 + 4.19 r^3$

- Er skjæringspunktet β_0 signifikant forskjellig fra 0? ($\alpha = 0.05$)

  Nei p-verdi $= 0.790 > 0.05$

  95% KI $[-4.287, 3.443]$

- Hva blir estimatet vårt for π?

  $V = \frac{4}{3}\pi r^3$

$$\hat{\beta}_1 = 4.19 = \frac{4}{3}\hat{\pi}$$

$$\Downarrow$$

$$\hat{\pi} = 4.19/4/3 = 3.1425$$

ok.

Merk: også vanlig å bruke feks log-transformasjon for å skape lineære sammenhenger.