Bonus #2 Fredag 1/12

Regresjon (igjen)

Hva er en regresjonsmodell:
Det er (den formelen, de spesifikasjonene, de formuleringene) som spesifiserer sammenhengen mellom en responsvariabel (som vi kaller y) og en eller flere forklaringsvariabler/prediktorer/kovariater x

Hva det er riktigst/mest hensiktsmessig å kalle x, avhenger av hva vi har tenkt å bruke regresjonsmodellen til:
- Estimere/gi et tall for sammenhengen mellom x og y : $\hat{\beta}$ $(\hat{\beta}_1)$

eller - predikere y på best mulig måte.

Uansett hva formålet med regresjonsanalysen er, vil en enkel, lineær regresjonsmodell se slik ut:

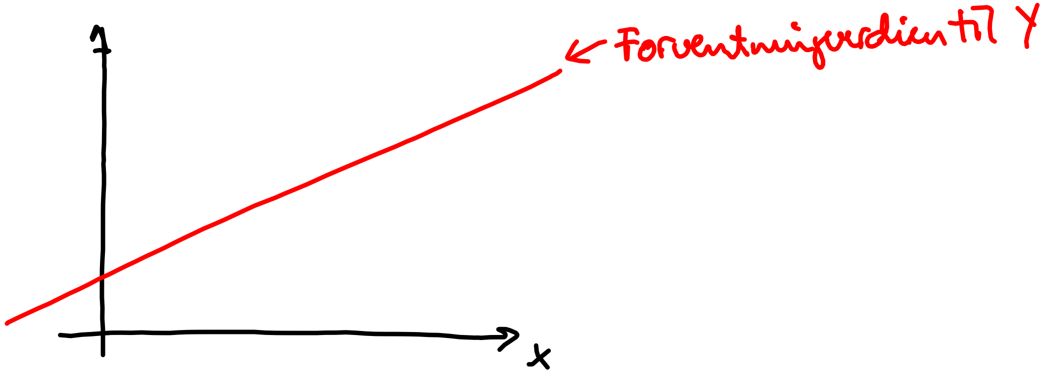
enkel betyr én forklaringsvar/prediktor/kovariat

Ch 10

(ofte har ~~out~~ jeg blandet y og Y: altså observasjoner (y) og stokastiske variabler (Y) når jeg har satt opp regresjonsmodeller for dere. La oss være litt mer formelle nå :)

$$Y = \underbrace{\beta_0 + \beta_1 x}_{\text{Faste ledd (hvis vi vet hva x er)}} + \underbrace{\varepsilon}_{\text{Stokastisk}}, \quad \varepsilon \sim N(0, \sigma)$$

$$E(Y) = E(\beta_0 + \beta_1 \cdot x + \varepsilon) = \beta_0 + \beta_1 x + 0 = \beta_0 + \beta_1 x$$

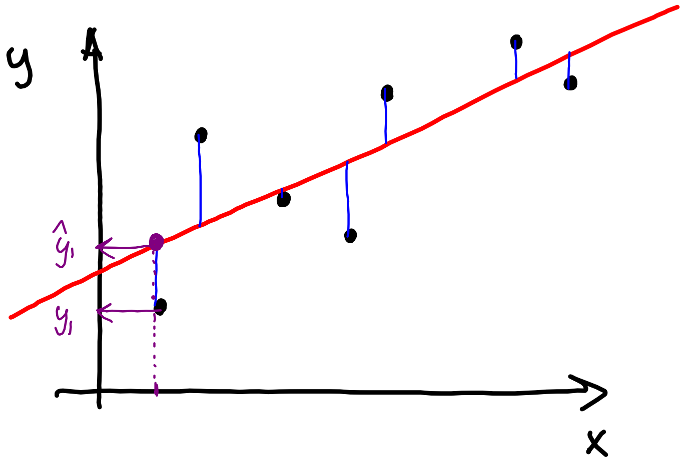


Så begynner vi å observere: For hver x_i observerer vi y_i (men ikke alle ligger nøyaktig på linja: de er e_i unna linja)

Modellen kan derfor også se slik ut:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \quad , i = 1, \ldots, n$$

betegner individene i studien

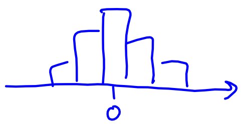


• observasjonene: y_i

— forventet verdi for y: $\hat{y}$

| Residualer: Disse kan brukes til å beregne et estimat for σ.

Residualene skal helst, hvis vi plotter dem i et histogram, være omtrent normalfordelte omkring 0:



Binomisk fordeling:

Vi har* n uavhengige forsøk
med * 2 mulige utfall: S / F,
og * P(S) = p er lik i hvert forsøk

og hvis

X = # suksesser på n forsøk,
så er $X \sim Bin(n, p)$
$X \sim bin(n, p)$
$X \sim binomisk(n, p)$

I en binomisk fordeling er
$E(X) = np$
$Var(X) = np(1-p)$

H-2016 Eksamen 1c

Vi har * n = 200 uavh. hypotesetester
med * 2 mulige utfall: Forkast H_0/Behold H_0
og * $P(\text{Forkaste } H_0) = P(\text{Fork } H_0 \mid H_0)$
$= 0.05$

premissene i oppgaven

Hvis X = # ~~Forkastede H_0~~
Feilaktig forkastede H_0,

så er $X \sim bin(200, 0.05)$

$E(X) = n \cdot p = 200 \cdot 0.05 = \underline{10}$

H-2008 2

Virkestoff i blod måles i to grupper

a) H_0: Lik konsentrasjon i de to gruppene,

$\mu_1 = \mu_2$, $\mu_1 - \mu_2 = 0$

H_1: Ikke lik konsentrasjon i de to gruppene:

$\mu_1 \neq \mu_2$, $\mu_1 - \mu_2 \neq 0$

Forutsetter at fordelingene i de to gruppene er relativt normalfordelte, så to-utvalgs t-test kan brukes (deskriptiv stat tyder på at dette er ok, selv om jeg helst skulle sett histogrammer i de to gruppene for å avgjøre n-fordelingsantakelsen. Dette er viktig fordi antallet i hver gruppe er såpass lite at vi neppe kan forvente at CLT slår inn ennå.)

b) Testobservator:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\underbrace{SE(\bar{X}_1 - \bar{X}_2)}_{S_{pooled}}} \sim T_{df} \quad \text{Ch 7}$$

$df = n_1 + n_2 - 2$

Antar at $\sigma_1 = \sigma_2$ og at de to gruppene representerer målinger fra uhv $N(\mu_1, \sigma_1)$ og $N(\mu_2, \sigma_2)$

Velg riktig nevner fra Ch 7 og tilhørende df.