Linear regression

Har i hørt "Lineær regression" før?
→ hvad betyder det?
→ hvad bruges det til?
Vi anvender ordene:
→ Datasæt
→ Måle punkter

hvad?

Datasæt → vores data, som består af ALLE vores målepunkter, som bliver opstillet i en tabel
Målepunkter → punkter som vi måler og de består af en uafhængig (X) og afhængig (y)

Eksempel 7.16.1.1: Datasæt med 4 målepunkter

Datasættet i tabellen indeholder fire målepunkter. De uafhængige variable står på rækken over de afhængige variable. Således har målepunkt nr. 3 værdien $(X, Y) = (5{,}74; 2{,}04)$.

Dette er hvad vi kalder vores værdier ved en regression

Uafhængig variabel (baggrundsvariabel)

Måling nr.	1.	2.	3.	4.
X	0,78	2,72	5,74	7,48
Y	0,44	5,52	2,04	6,06

Afhængig variabel (responsvariabel)

Regression kan vi bruge til at se om der er en sammenhæng (f.eks. en lineær sammenhæng) mellem vores data/punkter.
→ laves ved brug af CAS (vi ser en video om hvordan man gør i excel)

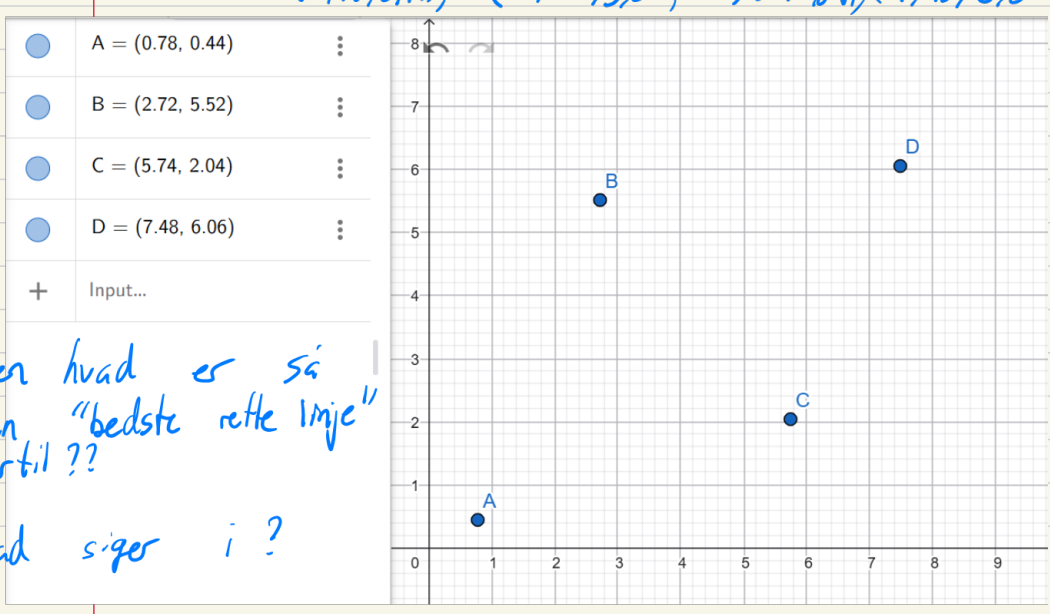
Ideen er at man indtegner sine punkter i et koordinat system.

Eksempel: videre arbejde med 7.16.11

Vi indtegner vores punkter i et koordinat-system.

Måling nr.	1.	2.	3.	4.
X	0,78	2,72	5,74	7,48
Y	0,44	5,52	2,04	6,06

(0,78; 0,44), (2,72; 5,52), (5,74; 2,04), (7,48; 6,06)



Men hvad er så den "bedste rette linje" hertil??

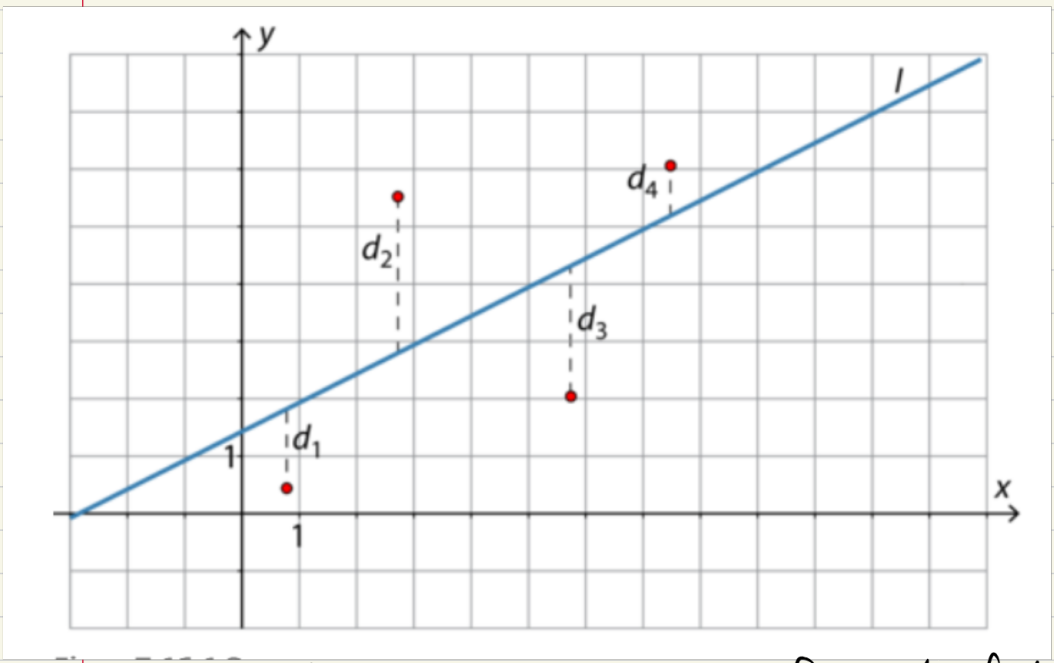
Hvad siger i?

Regressionslinje

Regressionslinjen kan opfattes som en "gennemsnitslinje" mellem målepunkterne. Regressionslinjen er tilpasset, så *summen* af de *kvadrerede* lodrette *afstande* mellem målepunkterne og regressionslinjen er mindst mulig.

I eksemplet med 4 målepunkter er regressionslinjen således den linje, hvor summen af de kvadrerede lodrette afstande, $d_1^2 + d_2^2 + d_3^2 + d_4^2$, er den mindst mulige.

Hvis man laver den samme beregning med de samme målepunkter, men på alle andre rette linjer, vil summen $d_1^2 + d_2^2 + d_3^2 + d_4^2$ være større.



Jeres CAS program sørger for at finde den mindste gennemsnitlige afstand

Den lineære regression går derfor ud på at opstille en ligning for den rette linje, der "passer bedst" til målepunkterne.

Bemærk

Lineær regression kræver, at der er *mindst* tre målepunkter i datasættet.

Det falder uden for dette kapitel at redegøre for metoden bag lineær regression.

Lineær regression foretages oftest med et CAS-værktøj.

Nu skal vi se film

https://matb-htx.systime.dk/?id=1433#c13579

Nu skal i selv prøve

Øvelse 7.16.1.1: Lineær regression

Med udgangspunkt i datasættet fra eksempel 7.16.1.1 på denne side skal ligningen for regressionslinjen beregnes.

Uafhængig variabel (baggrundsvariabel)

Måling nr.	1.	2.	3.	4.
X	0,78	2,72	5,74	7,48
Y	0,44	5,52	2,04	6,06

Afhængig variabel (responsvariabel)

Sammenhæng mellem uafhængig og afhængig variabel

Indtil videre har vi bare set hvordan man kan lave en regression ud fra en lineær sammenhæng, men:

Regressionsligningen siger ikke noget om, hvor god en lineær sammenhæng der er mellem den uafhængige variabel (x) og den afhængige variabel (y). Vi kan jo lave en regressionlinje, uanset hvordan målepunkterne ligger spredt.

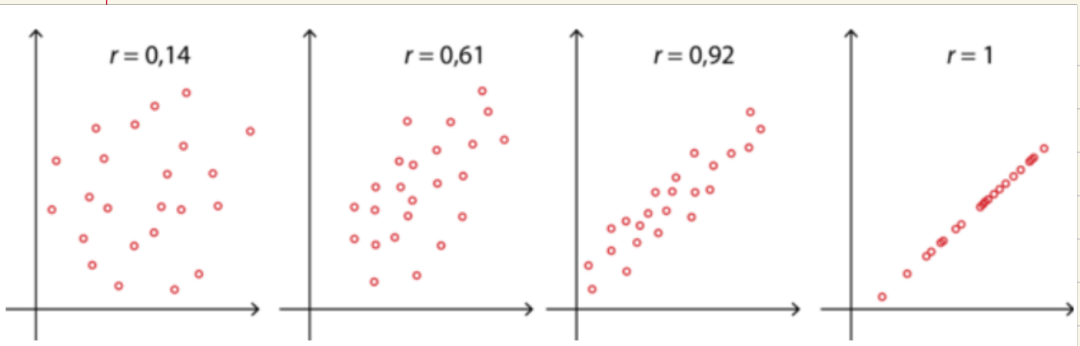
Vi bruger derfor en *korrelationskoefficient*, som beskriver i hvor høj grad, der er lineær sammenhæng. Korrelationskoefficienten betegnes med r eller R. Derfor anvendes af og til ordet *"r-værdi"*. Værdien r^2 kaldes også *forklaringsgrad*. r^2 "forklarer" noget om, hvor god en lineær sammenhæng der er mellem punkterne i datasættet.

Jo tættere r^2 er på 1 jo bedre sammenhæng
→ hvis $r^2 = 1$ så passer den perfekt

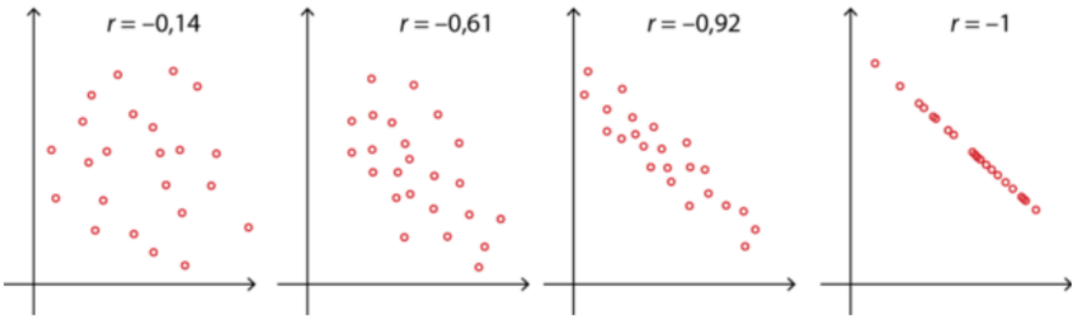
Hvis $r = 0$, er der ingen lineær sammenhæng. Hvis r er lille, er der kun ringe lineær sammenhæng. Hvis r ligger tæt på 1 eller -1, er der en god lineær sammenhæng. Følgende kan anvendes som tommelfingerregel:

- Når forklaringsgraden $r^2 > 0{,}99$, er der tale om en god lineær sammenhæng.
- Når $0{,}95 \leq r^2 \leq 0{,}99$, er der tale om en "rimelig" lineær sammenhæng.

Når alle målepunkter ligger på regressionslinjen, er $r = 1$ eller -1.



Figur 7.16.2.1



Figur 7.16.2.2

Man kan lave regression til alle vores funktionstyper, men vi kigger kun på "Lineær regression"

Men vi bruger r^2 til at bestemme om der er sammenhæng eller ikke i forhold til den funktionstype vi kigger på

Til hver øvelse skal i angive r^2 værdi og forklare om det er en god eller dårlig lineær regression

Øvelser

Øvelse 1: sammenhæng mellem temperatur og solgt is

Temperatur (°C)	Antal is solgt
10	25
12	31
14	40
15	43
17	50
19	57
21	63

1. Tegn datapunkterne i et koordinatsystem.
2. Bestem regressionslinjen (mindste kvadraters metode).
3. Angiv ligningen på formen:

$$y = ax + b$$

4. Bestem hældningskoefficienten og forklar, hvad den fortæller om is-salg.
5. Bestem hvor mange is der forventes solgt ved **18°C**.

6 hvor varmt er der hvis der bliver solgt 47 is.

Øvelse 2: Transporttid og afstand

Distance (km)	Transporttid (min)
2	5
4	9
6	14
8	18
10	23
12	27

1. Placer datapunkterne i et koordinatsystem.
2. Find den lineære regressionslinje.
3. Hvad er den gennemsnitlige hastighed ifølge modellen?
4. Beregn den forventede transporttid for **15 km**.

Øvelse 3: porto og vægt

Vægt (g)	Porto (kr.)
50	20
100	24
150	28
200	32
250	37
300	41
350	46

1: Lav en lineær regression for data til højre
2: Er det en god lineær regression
3: For hver gram vores pakke vokser, forklar hvor meget portoen stiger, og hvordan i har aflæst det
4: Udregn, hvor meget det koster at sende en pakke der vejer 400 g. og en anden pakke der koster 75 g.
5: Hvor tung kan vores pakke være, hvis man betaler 35 kroner i porto.

Øvelse 4: Bilens brændstofforburg og hastighed

Hastighed (km/t)	Forbrug (L/100 km)
30	5.2
40	5.6
50	5.8
60	6.1
70	6.4
80	6.8
90	7.3

1: Lav en lineær regression for data til højr

2: Er det en god lineær regre

3: Hvor meget stiger vores forbr

med per 10 km/t vi ændre far

4: Hvor stort er vores forbrug