

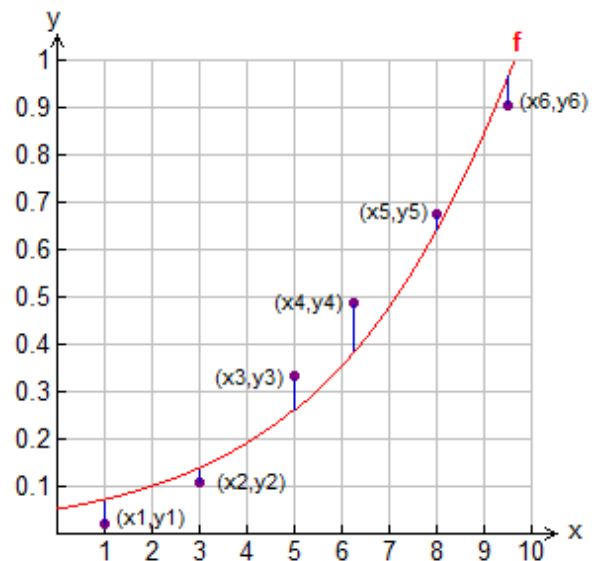
Regression

Når vi har et sæt af datapunkter/sammenhørende værdier $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$, vil vi gerne bestemme en model, der passer bedst muligt til disse punkter.

For at gøre det benytter vi dels en formodning om, hvilken model der er mest passende og dels regression.

Formodningen om, hvilken modeltype, der kan anvendes, bygger ofte – som det var tilfældet med logistisk vækst – på antagelser om væksthastighed og en opstilling af en differentialligning ud fra denne antagelse.

Regression benyttes derefter til at opstille lige netop den funktion inden for den aktuelle modeltype, der passer bedst muligt til det givne datasæt.



Mindste kvadraters metode

"Bedst muligt" betyder i denne sammenhæng, at den totale lodrette afstand fra punkterne til grafen er mindst mulig – og ikke den egentlige (vinkelrette) afstand til grafen.

I stedet for at minimere selve total-afstanden, minimeres den totale kvadrat-afstand, da kvadrater er nemmere at håndtere regneteknisk end numeriske værdier.

Teorien er udviklet af C. F. Gauss i 1820-erne.

Når man har bestemt typen, går problemet altså ud på at bestemme konstanterne i den valgte modeltype

$$f(x) = a \cdot x + b, f(x) = b \cdot a^x, f(x) = b \cdot x^a, f(x) = \frac{c}{1 + a \cdot e^{-b \cdot x}}, f(x) = a \cdot x^2 + b \cdot x + c \text{ eller andre, så}$$

summen af kvadraterne på afstandene mellem $f(x_i)$ og y_i , dvs. summen af alle

$(f(x_i) - y_i)^2$ er minimal.

Dvs. at a og b fastlægges, så summen af residualerne i anden bliver mindst mulig.

Eks. Modellen er lineær $f(x) = a \cdot x + b$

Data:

x	y
1	7698
4	7685
5	7632
8	7610
18	7565
23	7495

Kvadratsummen KS afhænger af de to konstanter a og b , og skal minimeres i forhold til de to.

$$KS = (f(1) - 7698)^2 + (f(4) - 7685)^2 + (f(5) - 7632)^2 + osv. = \dots$$

De to konstanter a og b skal nu ændres, så KS bliver mindst mulig.

I TI-Nspire kan det lettest gøres ved at lave to lister L_x og L_y , lave punktplot (L_x, L_y) og tilpasse funktionen $f(x) := a \cdot x + b$, så godt som muligt ved at ændre på a og b og samtidig holde øje med kvadratsummen $KS := \text{sum}((L_x - f(L_y))^2)$.

Det kræver en hel del tålmodighed og prøven frem og tilbage ved skiftevis at ændre på a og b (fx med en skyder), men man kan komme ganske tæt på.

TIP! For at få en ide om størrelsesordenen af a og b , kan man tegne en linje på øjemål og bestemme ligningen for den.

Man kan udlede nedenstående formler for a og b i en lineær regression

$$a := \left(\frac{-n \cdot x_m \cdot y_m + \text{sum}(L_y \cdot L_x)}{\text{sum}(L_x^2) - n \cdot x_m^2} \right) \text{ og } b := -a \cdot x_m + y_m, \text{ hvor } x_m \text{ er middelværdien af 1. koordinaterne til datapunkterne og } y_m \text{ er middelværdien af 2. koordinaterne}$$