Statistisch modelleren

1. Scatterplot

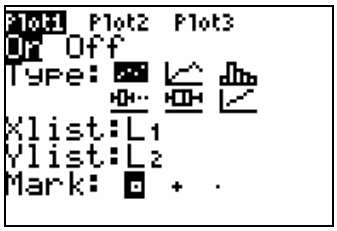
Een efficiënte manier om een eerste indruk te krijgen van een eventueel verband tussen twee kwantitatieve grootheden is het tekenen van een puntenwolk (spreidingsdiagram of scatterplot).

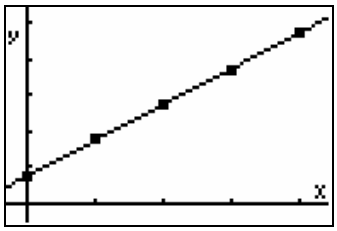
Veronderstel dat een TV-technicus €15 vraagt om naar je huis te komen en dat per uur hij werkt aan je TV € 20 extra kost. Noem y de totale kostprijs voor een reparatie aan huis en x het aantal werkuren. De relatie of verband tussen x en y is: $y = 15 + 20x$.

Om de totale kostprijs te berekenen, pluggen we het aantal gewerkte uren in de bovenstaande formule en zo bekomen we het resultaat. We noemen x de onafhankelijke variabele en y de afhankelijke variabele. y afhankelijk van x, $y(x)$.

# uren (x) L1	prijs (y) L2
0	15
1	35
2	55
3	75
4	95

Tabel 1



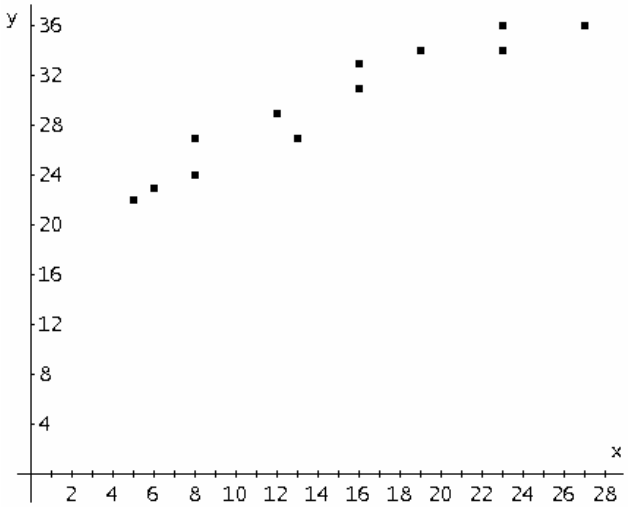


Tussen x en y is er een perfect lineair verband. Alle data liggen precies op één rechte. Bovendien geldt dat indien x vermeerdert ook y vermeerdert. We spreken in dit geval over een positief verband. Indien y zou verminderen als x vermeerdert, spreken we over een negatief verband.

In een tweede voorbeeld bekijken we de gegevens die een elektronicafirma verzamelde over het salaris van technicus om zich een idee te vormen van welk salaris men best betaalt in functie van het aantal jaren ervaring.

# jaren ervaring (x) L1	Salaris in € 1000 (y) L2
12	29
16	31
6	23
23	34
27	38
8	24
5	22
19	34
23	36
13	27
16	33
8	27

Tabel 2



In dit geval toont de scatterplot geen perfect lineair verband. Toch laat deze plot vermoeden dat er een benaderd lineair verband is omdat alle datapunten zich in de omgeving van een rechte bevinden.

Teken op de bovenstaande grafiek een rechte die volgens jou het best aansluit bij de puntenwolk.

Leid van de zo juist getekende rechte het functievoorschrift af, de rechte die volgens jou het best het verband aangeeft tussen de verzamelde data.

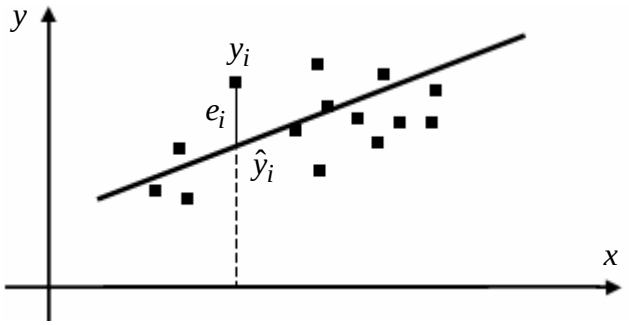
.....
.....
.....
.....
.....

2. Lineaire regressie

Wat is de rechte die het best aansluit bij de puntenwolk $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$?

Veronderstel even dat het lineair model gegeven is door $y = ax + b$. Voor een datapunt (x_i, y_i) noteren we de voorspelde waarde met $\hat{y}_i = ax_i + b$.

Het verschil tussen de geobserveerde waarde en de voorspelde waarde noemen we de fout van de voorspelling of het residu, e_i, behorende bij x_i: $e_i = y_i - \hat{y}_i$.



We zoeken een rechte die op een bepaalde manier het geheel van al deze verticale afwijkingen minimaliseert.

Een criterium voor het minimaliseren noemt men het *kleinste kwadraten criterium*: nl. bepaal a en b zodanig dat $\sum_{i=1}^{n} e_i^2$ minimaal is.

Deze beste rechte noemt men de kleinste kwadraten rechte of de *lineaire regressielijn* van y op x.

Men kan aantonen dat volgens het kleinste kwadraten criterium voor de beste rechte $y = ax + b$ door de punten $(x_1, y_1), ..., (x_n, y_n)$ geldt dat:

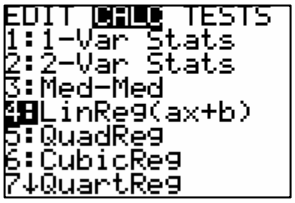
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{en} \quad b = \bar{y} - a\bar{x} \quad \text{met } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ en } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

3. Berekening van de lineaire regressielijn voor tabel 1.2 met de TI 83/84 Plus

a. Voer de data van tabel 1.2 in via de STAT-editor in de lijsten L_1 en L_2.

b. Maak een scatterplot met L_1 op de x-as en L_2 op de y-as.

c. Voer een lineaire regressie uit met het commando STAT<CALC> 4:LinReg(ax+b).



Vul dit commando als volgt aan op het rekenscherm: LinReg(ax+b), L_1,L_2.

Wat is de beste rechte volgens het kleinste kwadraten criterium?

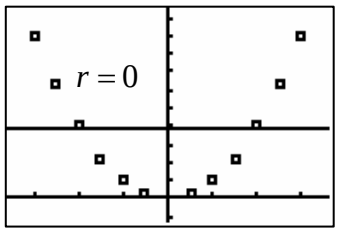
.....
.....
.....

Vergelijk de regressielijn met het resultaat dat je eerst handmatig tekende.

d. Om de lineaire regressielijn toe te voegen aan de scatterplot, kan je het functie-voorschrift uit punt c invoeren via Y= maar ook automatisch indien je Y_1 toevoegt aan het commando LinReg: LinReg(ax+b), L_1,L_2,Y_1.

Y_1 vind je via VARS<Y-VARS> 1:Function.

4. Correlatie

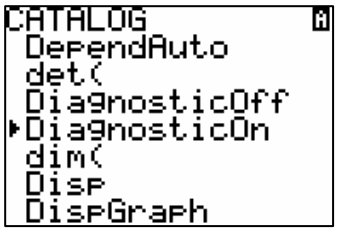


Een lineaire regressielijn kan in principe voor iedere puntenwolk berekend worden. Vanzelfsprekend heeft deze regressielijn alleen maar zin als de scatterplot data een lineair verband aangeeft. Hoe meer de data aansluiten bij de regressielijn hoe beter het model overeenkomt met de realiteit. De regressiecoëfficiënt

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

is een getal tussen 1 en -1 dat de graad van het lineair verband aangeeft. Hoe dichter de correlatiecoëfficiënt bij 1 of -1 hoe sterker het lineair verband. Voor $r > 0$ spreekt men van een positief verband en indien $r < 0$ over een negatief verband.

Ook deze coëfficiënt kan berekend worden met de TI-83/84 Plus. Eerst moeten we hiervoor de optie Diagnostics aanzetten via: 2nd[CATALOG]. Indien we dan opnieuw de regressielijn berekenen voor tabel 1.2 zien we dat ook $r = .9738972232$ verschijnt en kunnen we spreken van een sterk lineair verband.

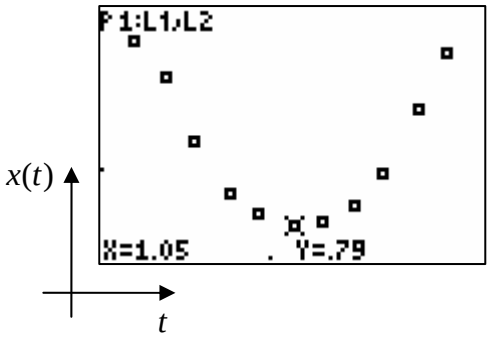


5. Niet-lineaire regressie

De hieronder afgebeelde data horen bij een botsende bal. Daar deze beweging gebeurt in het zwaartekrachtveld is er een kwadratisch verband tussen tijd en afstand.

Tijd t	Afstand $x(t)$
0.67	1.46
0.75	1.33
0.82	1.09
0.9	0.91
0.97	0.83
1.05	0.79
1.12	0.8
1.2	0.86
1.27	0.98
1.35	1.21
1.42	1.42

Tabel 3



Het model dat we hiervoor opstellen is van de vorm: $x(t) = A(t-B)^2 + C$. Probeer eerst manueel een zo goed mogelijk kwadratisch model, $x(t) = A(t-B)^2 + C$, op te stellen voor deze data.

a. Bepaal eerst een waarde voor B en C.
(Hint: Wat is het verband tussen de coördinaten van de top en B en C?)
.....
.....

b. Bepaal de coëfficiënt A zodat de parabool volgens jou zo goed mogelijk aansluit bij de data.
.....
.....

c. Vergelijk de waarde van A met de gravitatieconstante $g = 9.81\frac{m}{s^2}$. .....

Ook in dit geval kunnen we een beste parabool zoeken volgens het kleinste kwadraten criterium. Wat is het resultaat als we op deze data een kwadratische regressie, STAT<CALC> 5:QuadReg, uitvoeren met de TI-83/84 Plus?
.....
.....

Vergelijk dit kwadratisch model met het manueel gevonden resultaat.

De kwaliteit van een niet-lineair regressiemodel wordt weergegeven door de determinatiecoëfficiënt R^2 (DiagnosticOn). Hoe dichter R^2 bij 1, hoe beter het regressiemodel. Voor een lineaire regressie geldt dat $R^2 = r^2$.