Hoofdstuk 9

Regressie, correlatie en modelvorming

9.1 Lineaire regressie

9.1.1 Inleidend voorbeeld

De punten (1,3), (2,1) en (3,5) liggen niet op één rechte.

Hoe kunnen we de rechte vinden die het "best" aansluit bij die punten ?

Plaats de coördinaten in de lijsten **L1** en **L2** zoals hieronder aangegeven en druk **STAT<CALC> 4:LinReg(ax+b)**.

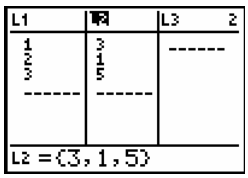
Op het basisscherm verschijnt er het commando **LinReg(ax+b)**.

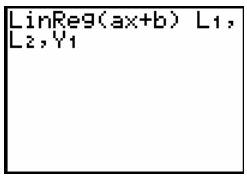
Vul dit aan met **2nd[L1]**, **2nd[L2]**, **VARS<Y-VARS> 1:Function 1:Y1**.

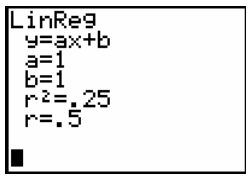
Door het toevoegen van Y1 wordt de vergelijking van de beste rechte weggeschreven in Y1.

Het drukken op **ENTER** levert de beste rechte $y = x + 1$.

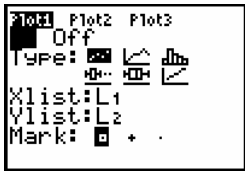
De extra gegevens r en r^2 verkrijg je indien je eerst **2nd[CATALOG] DiagnosticOn** hebt uitgevoerd vanuit het basisscherm.

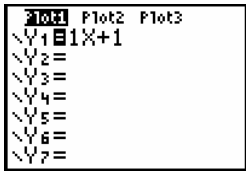


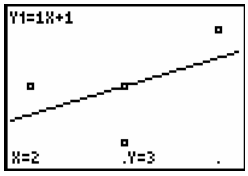




Definieer **Plot1** zoals hieronder en controleer even de functies (**Y=**) die getekend zullen worden. **ZOOM 9:ZoomStat** geeft de grafieken van de puntenwolk en de beste rechte. Via **TRACE** kunnen we de rechte doorlopen om eventueel waarden van y horende bij x te voorspellen.



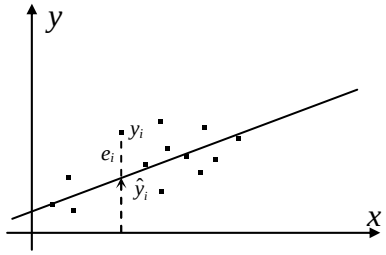




9.1.2 Hoe wordt de beste rechte berekend ?

Gegeven een puntenwolk van n punten, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ die min of meer een lineaire trend vertonen (zie figuur).

Beschouw een rechte $y = ax + b$ "door" deze punten waarmee we de grootheid y wensen te voorspellen bij gegeven x.



We definiëren voor elk punt het *residu* e_i (zie figuur), met $\hat{y}_i = ax_i + b$, als volgt :

$$\begin{aligned} e_i &= \text{observatie} - \text{voorspelling} \\ &= y_i - \hat{y}_i \\ &= y_i - (ax_i + b) \end{aligned}$$

Merk op dat een residu positief is wanneer het punt boven de rechte gelegen is en negatief wanneer het onder de rechte gelegen is.

Om de "beste" rechte door de puntenwolk te zoeken gebruiken we het *kleinste kwadraten criterium* nl. bepaal a en b zodanig dat $\sum_{i=i}^{n} e_i^{\,2}$ minimaal is.

Deze beste rechte noemt men de kleinste kwadraten rechte of de *lineaire regressie* van y op x (met y als afhankelijke en x als onafhankelijke veranderlijke).

Om de beste a en b te bepalen gebruiken we de afwijkingen $u_i = x_i - \overline{x}$ en $v_i = y_i - \overline{y}$ t.o.v. de rekenkundige gemiddelden $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ en $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Voor de residu's geldt :

$$\begin{aligned} e_i &= y_i - ax_i - b \\ &= (y_i - \bar{y}) - a(x_i - \bar{x}) - (b - \bar{y} + a\bar{x}) \\ &= v_i - au_i - (b - \bar{y} + a\bar{x}) \end{aligned}$$

De kwadratensom der residu's wordt :

$$\begin{aligned} \sum_{i=i}^{n} e_i^{\,2} &= \sum_{i=1}^{n} ((v_i - au_i) - (b - \bar{y} + a\bar{x}))^2 \\ &= \sum_{i=1}^{n} (v_i - au_i)^2 + \sum_{i=1}^{n} (b - \bar{y} + a\bar{x})^2 \quad \text{(aangezien } \sum_{i=1}^{n} u_i = 0 \text{ en } \sum_{i=1}^{n} v_i = 0\text{)} \\ &= \sum_{i=1}^{n} (v_i - au_i)^2 + n(b - \bar{y} + a\bar{x})^2 \end{aligned}$$

Nu is $\sum_{i=i}^{n} e_i^{\,2}$ geschreven als een som van twee positieve uitdrukkingen, deze som is minimaal als beide uitdrukkingen minimaal zijn. We bepalen eerst a zodat de eerste uitdrukking, $\sum_{i=1}^{n} (v_i - au_i)^2 = \sum_{i=1}^{n} u_i^{\,2}.a^2 - 2\sum_{i=1}^{n} u_i v_i.a + \sum_{i=1}^{n} v_i^{\,2}$, minimaal wordt.

Dit is een kwadratische uitdrukking in a die minimaal is als $a = \dfrac{\sum_{i=1}^{n} u_i v_i}{\sum_{i=1}^{n} u_i^{\,2}}$.

Als we $b = \bar{y} - a\bar{x}$ stellen, is $n(b - \bar{y} + a\bar{x})^2 = 0$ en de tweede uitdrukking minimaal.

<u>Samenvattend</u>

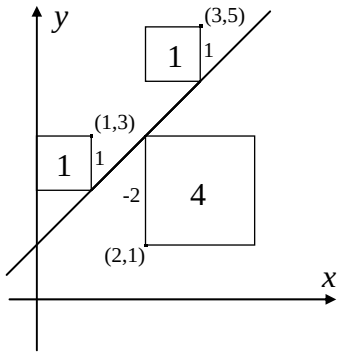
Voor de beste rechte $y = ax + b$ door de punten $(x_1, y_1), ..., (x_n, y_n)$ geldt :

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

We berekenen manueel de beste rechte door de punten (1,3), (2,1) en (3,5) van het inleidende voorbeeld met de onderstaande tabel.

	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
	1	3	-1	0	0	1
	2	1	0	-2	0	0
	3	5	1	2	2	1
$\sum_i$	6	9	0	0	2	2
	$\overline{x} = 2$	$\overline{y} = 3$				$a = 1$ & b = 1

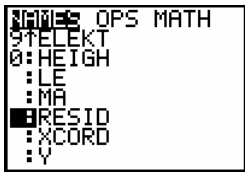
Klaarblijkelijk is $y = x + 1$ de beste rechte. De som van de kwadraten der residu's, $\sum_{i=1}^{3} e_i^{\,2} = 6$, is minimaal voor deze rechte.

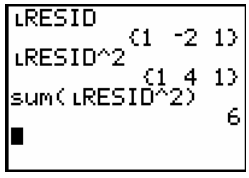


Dit is te interpreteren als een som van oppervlakten van vierkanten die minimaal is (zie hiernaast).

De **TI-83** maakt na een regressieberekening (zie 9.1.1) automatisch de lijst van de residu's aan. Deze lijst vind je met **2nd[LIST] RESID**.

Voor het berekenen van de som van de kwadraten der residu's gebruik je **2nd[LIST]**<MATH> **5:sum(**.



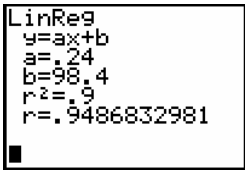


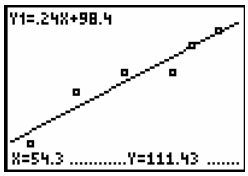
9.2 Correlatie

De lengte l van een metalen staaf wordt gemeten bij verschillende temperaturen.

t (°C)	20	30	40	50	60
l (mm)	102	107	109	109	113

In de fysica gebruikt men als model het lineair verband $l = a.t + b$. We zoeken de beste rechte uitgaande van de data (reken manueel na met een tabel zoals in de vorige paragraaf. Vervang hiertoe t door x en l door y.

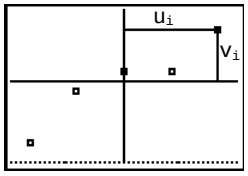




Wat is de betekenis van de *correlatiecoëfficiënt* r ?

Hiertoe tekenen we eerst de rechten $y = \bar{y}$ (met $\bar{y} = 108$) en $x = \bar{x}$ (met $\bar{x} = 40$). Teken de verticale rechte met de volgende opdracht in het basisscherm : **2nd[DRAW] 4:Vertical 40** .

Vervolgens tekenen we de verticale afwijking $v_i = y_i - \bar{y}$ en de horizontale afwijking $u_i = x_i - \bar{x}$ van een punt (x_i, y_i) t.o.v. het zwaartepunt $(\bar{x}, \bar{y})$ van de puntenwolk met **2nd[DRAW] 2:Line(.**



Voor bovenstaande puntenwolk spreken we over een *positieve correlatie* tussen de grootheden x en y aangezien de meeste punten (u_i, v_i) in het eerste of derde kwadrant gelegen zijn t.o.v. het assenstelsel met oorsprong $(\bar{x}, \bar{y})$. De beste rechte heeft een positieve richtingscoëfficiënt.

Merk op dat de beste rechte $y = ax + b$ steeds door het zwaartepunt $(\bar{x}, \bar{y})$ gaat aangezien $\bar{y} = a.\bar{x} + b$. Controleer dit door met je rekentoestel het snijpunt te zoeken van de beste rechte met de rechte $y = \bar{y}$.

Indien de meeste punten gelegen zijn in het tweede of vierde kwadrant is er sprake van een negatieve correlatie. De beste rechte heeft een negatieve richtingscoëfficiënt.

Indien de punten lukraak verdeeld zijn over de vier kwadranten, bestaat er geen *lineair* verband tussen de grootheden x en y. Een beste rechte berekenen en tekenen is nog steeds mogelijk, maar weinig zinvol.

In onze figuur is het duidelijk dat $\sum_{i=1}^{n} u_i v_i$ positief is, aangezien de bijdragen van de punten in het eerste en derde kwadrant tot die som steeds positief zijn.

De som $\sum_{i=1}^{n} u_i v_i$ is echter afhankelijk van de gekozen eenheden voor x en y, zodat we tot de volgende definitie komen van de correlatiecoëfficiënt, onafhankelijk van de gekozen eenheden voor x en y : $r = \dfrac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\left(\sum_{i=1}^{n} u_i^{\,2}\right).\left(\sum_{i=1}^{n} v_i^{\,2}\right)}}$.

Bovenstaande formule is symmetrisch in x en y en maakt geen onderscheid tussen de onafhankelijke en afhankelijke veranderlijke.

We berekenen $r = \dfrac{240}{\sqrt{1000 \cdot 64}} \approx 0.95$ voor de gegeven data m.b.v. onderstaande tabel. Vergelijk het resultaat met dat van het rekentoestel:

x_i	y_i	$u_i = x_i - 40$	$v_i = y_i - 108$	$u_i.v_i$	$u_i^{\,2}$	$v_i^{\,2}$
20	102	-20	-6	120	400	36
30	107	-10	-1	10	100	1
40	109	0	1	0	0	1
50	109	10	1	10	100	1
60	113	20	5	100	400	25
			$\sum_i$	240	1000	64

Stellen we $\vec{u} = (u_1, u_2, ..., u_n)$ en $\vec{v} = (v_1, v_2, ..., v_n)$, dan volgt uit de ongelijkheid $|\vec{u}.\vec{v}| \le ||\vec{u}||.||\vec{v}||$ van Cauchy-Schwartz dat $-1 \le r \le 1$.

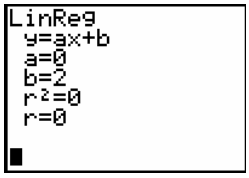
Hierbij kan gelijkheid enkel optreden als $\vec{v} = k\,\vec{u}$ of $v_i = k\,u_i$ voor elke *i*. De punten liggen op één rechte door het zwaartepunt $(\bar{x}, \bar{y})$ met vergelijking $v = k\,u$ t.o.v. het (u, v)-assenstelsel met dit zwaartepunt als oorsprong.

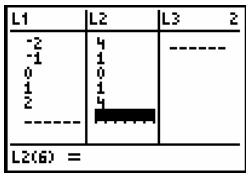
De correlatiecoëficiënt r is een maat voor het *lineaire* verband tussen de grootheden x en y. Hoe dichter de absolute waarde van r bij 1 gelegen is, hoe beter het lineaire verband.

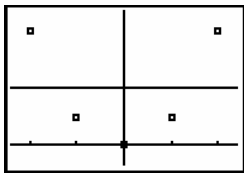
Voor positieve r is er een positieve correlatie en voor negatieve r een negatieve correlatie.

Voor waarden van r dicht bij 0 is er nagenoeg geen *lineair* verband. Er kan echter wel een ander functioneel verband zijn tussen x en y !

Beschouw hiertoe volgende data, gelegen op de parabool $y = x^2$. De beste rechte is $y = 2$, maar deze benadering heeft uiteraard geen zin.







Er bestaat een verband tussen de richtingcoëfficiënt a van de vergelijking van de beste rechte $y = ax + b$ en de correlatiecoëfficiënt r met :

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} u_i v_i}{\sum_{i=1}^{n} u_i^{\,2}} \quad \text{en} \quad r = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\left(\sum_{i=1}^{n} u_i^{\,2}\right) \cdot \left(\sum_{i=1}^{n} v_i^{\,2}\right)}}.$$

Er geldt immers :

$$a = \frac{\sum_{i=1}^{n} u_i v_i}{\sum_{i=1}^{n} u_i^{\,2}} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\left(\sum_{i=1}^{n} u_i^{\,2}\right) \cdot \left(\sum_{i=1}^{n} v_i^{\,2}\right)}} \cdot \frac{\sqrt{\sum_{i=1}^{n} v_i^{\,2}}}{\sqrt{\sum_{i=1}^{n} u_i^{\,2}}} = r \cdot \frac{\sqrt{\frac{\sum_{i=1}^{n} v_i^2}{n-1}}}{\sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n-1}}} = r \cdot \frac{s_y}{s_x}.$$

Er geldt ook dat $r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right)$ met $\frac{x_i - \bar{x}}{s_x}$ de gestandaardiseerde waarde van x_i.

Tenslotte geven we nog een nuttige formule voor r die het overbodig maakt de afwijkingen $x_i - \bar{x}$ en $y_i - \bar{y}$ t.o.v. de gemiddelden te berekenen. Door gebruik te maken van de output van het `2-Var Stats`-commando kan je hiermee r berekenen.

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

9.3 Modelvorming

9.3.1 De determinatiecoëfficiënt

De determinatiecoëfficiënt R^2 is een maat voor de kwaliteit van een regressiemodel dat *niet* noodzakelijk lineair is.

Als introductie nemen we de data (1,3), (2,1) en (3,5) die we reeds vroeger hebben gebruikt, met als beste rechte door deze drie punten de rechte $y = x + 1$.

Beschouw de onderstaande tabel met $\hat{y}_i = x_i + 1$. Dit is de voorspelde waarde van y m.b.v. het regressiemodel.

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	3	0	1
2	1	4	4
3	5	4	1
	$\sum_i$	8	6

Als we enkel de data y_i zouden kennen, is $\bar{y}$ de beste voorspelling voor elke y_i.

Voor de totale variatie van de data y_i t.o.v. $\bar{y}$ nemen we $\sum_{i=1}^{n}(y_i - \bar{y})^2$ als maat.

Deze waarde geeft aan in welke mate de puntenwolk verticaal afwijkt van de horizontale rechte $y = \bar{y}$.

Met het regressiemodel kunnen we y beter voorspellen daar de som van de kwadraten der residu's (of de variatie van de data y_i t.o.v. het regressiemodel) $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ kleiner dan of gelijk is aan $\sum_{i=1}^{n}(y_i - \bar{y})^2$ (waarom ?). Er geldt dat :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

totale variatie = verklaarde variatie + onverklaarde variatie

De determinatiecoëfficiënt R^2 is de fractie van de variatie van de data y_i t.o.v. $\bar{y}$ die verklaard wordt door het regressiemodel :

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In het vorige voorbeeld is $R^2 = 1 - \frac{6}{8} = \frac{1}{4}$.

M.a.w. er wordt slechts 25% van de variatie van de data y_i t.o.v. $\bar{y}$ verklaard door het regressiemodel. Voor de correlatiecoëfficiënt $r = 0.5$ geldt dat $r^2 = R^2$.

Het is algebraïsch algemeen te bewijzen dat voor een *lineair* regressiemodel geldt dat $r^2 = R^2$. Derhalve noemt men r^2 ook de lineaire determinatiecoëfficiënt.

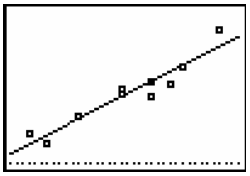
Vermeld steeds r^2 bij een lineaire regressie als maat voor hoe succesvol het lineaire model is om *y* te voorspellen.

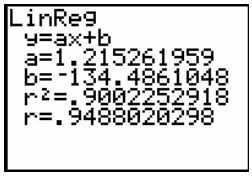
9.3.2 De residuplot

Als voorbeeld beschouwen we de lengte x in cm en de massa y in kg van 10 lukraak gekozen studenten :

x_i	163	185	180	175	168	175	191	180	160	183
y_i	60	90	78	81	71	79	104	84	64	83

We brengen deze data in de lijsten **L1** en **L2** en tekenen de puntenwolk. Zo te zien is een lineair model een goede benadering.

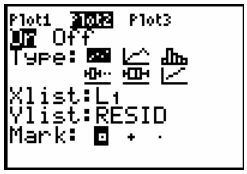


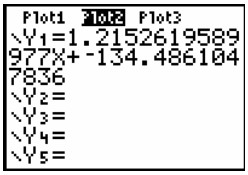


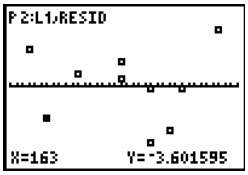
Bij lineaire regressie is de som van de residu's steeds nul (Bewijs !) :

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$$

Voor een goed model moet de *residuplot* , d.i. een grafiek van de residu's uitgezet tegen de data x_i , een lukrake verdeling tonen t.o.v. de x-as die hier de regressielijn voorstelt.

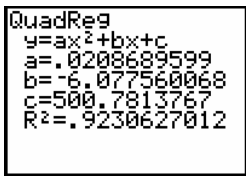


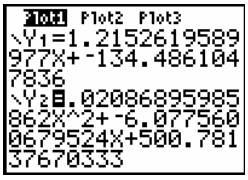


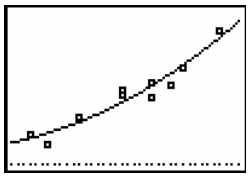


De determinatiecoëfficiënt r^2 is 0.9. Er wordt 90% van de variatie van de data y_i t.o.v. $\bar{y}$ verklaard door het lineaire regressiemodel.

Reken na, met **sum**(**LRESID**²), dat de som van de kwadraten der residu's, $\sum_{i=1}^{10} {e_i}^2$, gelijk is aan 143.7. We onderzoeken of het kwadratisch model beter is.

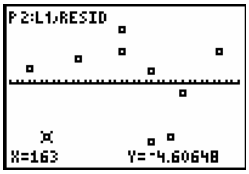


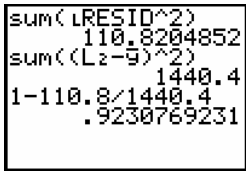




We maken nog een residuplot en berekenen de som van de kwadraten der residu's. Tevens rekenen we na dat de determinatiecoëfficiënt $R^2 = 0.923$.

Het kwadratisch model is iets beter dan het lineair model.





Onderzoek van de andere modellen die de **TI-83** berekent, geeft :

Model	R^2	r^2	r
LinReg(ax+b)		0.900	0.949
QuadReg (ax^2+bx+c)	0.923		
CubicReg (ax^3+bx^2+cx+d)	0.938		
QuartReg ($ax^4+bx^3+cx^2+dx+e$)	0.963		
LnReg ($a+b.\ln(x)$)		0.892	0.945
ExpReg ($a.b^x$)		0.915	0.957
PwrReg ($a.x^b$)		0.911	0.955

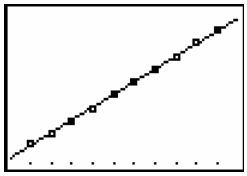
De logaritmische regressie LnReg wordt berekend met lineaire regressie van de data $\left(\ln(x_i), y_i\right)$, de exponentiële regressie ExpReg met lineaire regressie van de data $\left(x_i, \ln(y_i)\right)$ en de machtsregressie PwrReg met lineaire regressie van de data $\left(\ln(x_i), \ln(y_i)\right)$.

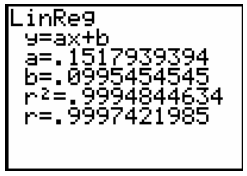
Dat er lineaire regressie werd gebruikt zie je aan het verschijnen van een correlatiecoëfficiënt en de notatie van de determinatiecoëfficiënt. Verklaar de transformaties van de data.

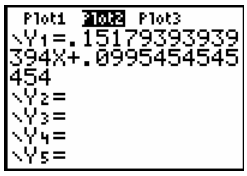
Als laatste voorbeeld beschouwen we een aantal gelijkmatige veelhoeken met zijden gelijk aan de lengte-eenheid en de corresponderende straal van de omgeschreven cirkel.

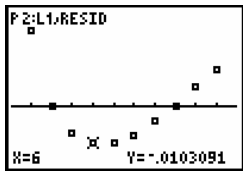
# zijden	3	4	5	6	7	8	9	10	11	12
straal	0.577	0.707	0.851	1.000	1.152	1.306	1.462	1.618	1.775	1.932

We brengen deze data in de lijsten **L1** en **L2** en tekenen de puntenwolk. Een lineair model blijkt uitstekend te zijn (zie r^2). De residuplot werpt echter een andere blik op de zaak.





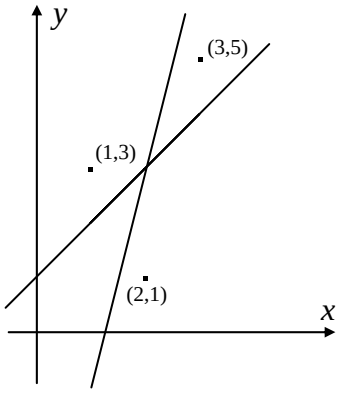




De residu's vormen een duidelijk patroon. Dit wijst op het bestaan van een beter model. We laten het aan de lezer over om dit model te bepalen.

9.4 Opdrachten

1. We beschouwen opnieuw de data (1,3), (2,1) en (3,5). Lineaire regressie van y op x leverde de rechte $y = x + 1$. Reken na dat lineaire regressie van x op y een *andere* rechte $x = \frac{1}{4}y + \frac{5}{4}$ oplevert.



In beide gevallen is $r = 0.5$ en $r^2 = 0.25$. De determinatiecoëfficiënt interpreteert de kwaliteit van beide rechten. 25% van de variatie in de data van ene veranderlijke t.o.v. hun gemiddelde wordt verklaard door het lineaire verband met de andere veranderlijke.

Naarmate $|r|$ dichter bij 1 komt te liggen zal het verschil tussen de twee rechten kleiner worden.

2. Onderstaande tabel geeft de snelheid en het verbruik van een wagen. Hoe verandert het verbruik y van die wagen in functie van de snelheid x ?

km/h	10	20	30	40	50	60	70	80
l /100 km	21	12	10	8	7	5.9	6.3	6.95

km/h	90	100	110	120	130	140	150
l /100 km	7.57	8.27	9.03	8.87	10.79	11.77	12.83

Teken de puntenwolk en bepaal de beste rechte door die punten. Zou jij deze rechte gebruiken om het verbruik bij een bepaalde snelheid te voorspellen ? Observeer de determinatiecoëfficiënt en de residuplot. Suggereer een beter model.

3. De onderstaande tabel bevat de volgende informatie over enkele convexe veelhoeken. x is het aantal zijden van de veelhoek en y het aantal verschillende diagonalen die in de veelhoek kunnen getekend worden.

Aantal zijden	4	5	6	7	8	9
Aantal diagonalen	2	5	9	14	20	27

Gebruik exponentiële regressie $y = a.b^x$ als model voor deze data. Bepaal de determinatiecoëfficiënt en de residuplot.

Een exponentieel model is goed als de tabel van de verhoudingen $\frac{y_{i+1}}{y_i}$ nagenoeg constant is (bij equidisdante x_i). Ga na dat dit hier niet het geval is.

We vinden een beter model door gebruik te maken van de techniek der voorwaartse differenties : $\Delta y_i = y_{i+1} - y_i$, $\Delta^2 y_i = \Delta y_{i+1} - \Delta y_i$, …

x_i	y_i	Δy_i	$\Delta^2 y_i$
4	2	3	1
5	5	4	1
6	9	5	1
7	14	6	1
8	20	7	-------
9	27	-------	-------

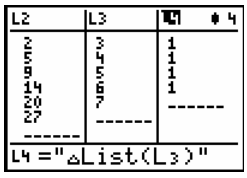
De tweede voorwaartse differenties zijn constant. Dit wijst op een model van de tweede graad. Zoek dit model en stel vast dat het door alle punten gaat. Bereken de som van de kwadraten der residu's.

Vind hetzelfde resultaat sneller met behulp van de combinatieleer !

Men kan aantonen dat, uitgaande van *equidistante* punten x_i, de n-de voorwaartse differenties van een n-de graadsveelterm constant zijn. Omgekeerd geldt ook dat als de n-de voorwaartse differenties constant zijn er juist één n-de graadsveelterm bestaat door de punten (x_i, y_i).

Men kan de voorwaartse differenties berekenen met de **TI-83** met het commando :

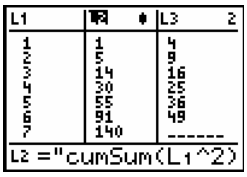
`2nd[LIST]<OPS> 7:ΔList(`.

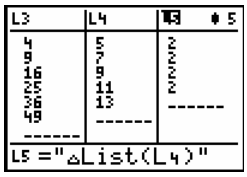


4. Zoek een formule voor de som van de eerste n kwadraten uitgaande van een tabel voor enkele waarden van n. Onderzoek de voorwaartse differenties.

n	1	2	3	4	5	6	7
$f(n) = \sum_{k=1}^{n} k^2$	1	5	14	30	55	91	140

Met de **TI-83** kun je deze tabel als volgt genereren :





De derde differenties zijn constant. Vind de derdegraadsveelterm door deze data en bewijs door inductie dat deze formule geldig is voor elke n.

5. Teken de volgende data en vind een geschikt model.

x_i	0	1	2	3	4	5	6	7	8	9	10
y_i	0.2	3.6	7.5	11.5	15	17	20.4	22.7	25.9	27.6	30.2

6. Teken de volgende data en vind een geschikt model.

Tip : bekijk de tabel van de verhoudingen $\frac{y_{i+1}}{y_i}$.

x_i	0	1	2	3	4	5	6	7	8	9	10
y_i	100	104	108	112	117	122	127	132	137	142	148

7. a) Een lineaire regressie levert een correlatiecoëfficiënt $r = 0$.
Wat is de vergelijking van de beste rechte ?
Geef een voorbeeld van data met $r = 0$.

b) Wat is de waarde van de correlatiecoëfficiënt van data gelegen op een rechte evenwijdig met de *x*-as of *y*-as ?

8. Zoek een formule voor de coëfficiënt *a* in het regressiemodel $y = a\,x$ (rechte door de oorsprong) voor gegeven data (x_i, y_i) met $i = 1, 2, \ldots, n$.

a) Wanneer we met een kracht *F* trekken aan een veer met veerconstante *k*, rekt de veer uit over een afstand *x* met $F = k\,x$. Bepaal de veerconstante (in N/m) van een veer waaraan de volgende massa's y_i (in gram) werden gehangen en de uitrekkingen x_i (in cm) werden gemeten.

x_i	0	1	1.9	2.8	3.7	4.7	5.6	6.6	7.5	8.5	9.3
y_i	0	51.8	101.3	148.4	201.5	251.1	302.3	350.9	397.1	452.5	496.3

b) Vergelijk het resultaat van a) met het regressiemodel $y = a\,x + b$.

c) Zijn de correlatiecoëfficiënten voor a) en b) dezelfde ?
En de determinatiecoëfficiënten ?

9. Het verband tussen spanning U en stroom I bij een niet-Ohmse weerstand wordt gegeven door $U = C \cdot I^{\beta}$ (U gemeten in volt en I in ampère), met C en β materiaalconstanten. Concrete metingen van U (in V) en I (in mA) levert de volgende resultaten :

I_i	5	6	7	8	9	10	11	12	13	14	15	16
U_i	6.90	7.14	7.35	7.53	7.70	7.82	7.97	8.08	8.17	8.27	8.38	8.45

I_i	17	18	19	20	21	22	23	24	25	26	27	28
U_i	8.52	8.58	8.66	8.72	8.77	8.80	8.88	8.93	8.97	9.00	9.04	9.12

Bepaal de beste waarden voor C en β aan de hand van deze meetwaarden. Ga na dat de **TI-83** deze waarden berekent met een logaritmische transformatie van de data en lineaire regressie.

10. Gegeven de lengte x in cm en de massa y in kg van 10 lukraak gekozen studenten.

x_i	163	185	180	175	168	175	191	180	160	183
y_i	60	90	78	81	71	79	104	84	64	83

Vind de exacte determinatiecoëfficiënt R^2 voor de onderstaande niet lineaire modellen en vergelijk deze met de lineaire determinatiecoëfficiënt r^2 van de getransformeerde data.

Model	R^2	r^2
LnReg ($a + b.\ln(x)$)		0.892
ExpReg ($a.b^x$)		0.915
PwrReg ($a.x^b$)		0.911