

# Statistiek voor MBW / KW

## Uitwerkingen huiswerkopgaven

Dr. ir. D.A.M.P. Blom & Dr. J.B.M. Melissen

2025

---

# Week 13: correlatie en regressie

## Hoofdstuk 13

**Opdracht 13.m1:** Bij regressieanalyse wordt de grootte die aangeeft hoe de variabele  $Y$  (gemiddeld) reageert op veranderingen in de variabele  $X$  aangeduid met ...

- (a) de constante term.
- (b) de storingsterm.
- (c) de richtingscoëfficiënt van de regressielijn.
- (d) de correlatiecoëfficiënt.

### Uitwerking

Bij een regressieanalyse op de onafhankelijke (verklarende) variabele  $X$  en de afhankelijke (te verklaren) variabele  $Y$  proberen we een lineair verband te vinden. De regressielijn is van de vorm  $Y = \alpha + \beta X$ , waarbij  $\alpha$  de constante term is en  $\beta$  de richtingscoëfficiënt van de regressielijn. Deze richtingscoëfficiënt geeft aan hoe de variabele  $Y$  gemiddeld reageert op een verandering van  $X$ .

Het juiste antwoord is dus (c).

**Opdracht 13.m2:** Als bij een regressieanalyse een correlatiecoëfficiënt wordt gevonden ter grootte van  $-0,50$ , dan ...

- (a) moet er een rekenfout zijn gemaakt.
- (b) is de variabele  $Y$  ongeschikt als verklarende variabele.
- (c) zullen bij hogere  $X$ -waarden doorgaans lagere  $Y$ -waarden worden aangetroffen.
- (d) kan de storingsterm geen positieve variantie hebben.

### Uitwerking

Een negatieve correlatiecoëfficiënt duidt op het feit dat als de verklarende variabele  $X$  stijgt, dan is er een negatieve trend voor de te verklaren variabele  $Y$ . Dat houdt dus in dat bij hogere  $X$ -waarden doorgaans lagere  $Y$ -waarden worden aangetroffen.

Het juiste antwoord is dus (c).

**Opdracht 13.m3:** Als bij lineaire regressie wordt vastgesteld dat  $\beta = 0$ , dan ...

- (a) hebben de variabele  $X$  en  $Y$  een tegengestelde relatie.

- (b) is het onderscheidingsvermogen  $1 - \beta$  gelijk aan 1.
- (c) toont de variabele  $X$  geen relatie met de variabele  $Y$ .
- (d) passen de punten perfect op de lijn.

#### Uitwerking

Het juiste antwoord is (c).

**Opdracht 13.m4:** Als bij een enkelvoudige lineaire regressieanalyse blijkt dat geldt  $\alpha = 0$  voor de berekende regressielijn, dan ...

- (a) geldt altijd  $\alpha = 0$  voor het regressiemodel.
- (b) is de correlatie tussen  $X$  en  $Y$  gelijk aan nul.
- (c) gaat de berekende regressielijn door de oorsprong.
- (d) loopt de berekende lijn horizontaal.

#### Uitwerking

Een regressielijn is altijd van de vorm  $Y = \alpha + \beta X$ . Als dan blijkt dat  $\alpha = 0$ , houden we een vorm  $Y = \beta X$  over. Zodra we  $X = 0$  invullen, krijgen we ook  $Y = \beta \cdot 0 = 0$ . Dit betekent dus dat de regressielijn door de oorsprong, oftewel het punt  $(0, 0)$  gaat.

Het juiste antwoord is dus (c).

**Opdracht 13.m6:** Bij een regressieanalyse wordt het verband onderzocht tussen de omzet  $Y$  ( $\times 10000$  euro) van een bedrijf en de reclame-uitgaven  $X$  ( $\times 1000$  euro) in de voorafgaande periode. De leverde als regressievergelijking op:

$$Y = 220 + 32X.$$

Stel dat in een nieuwe periode 3000 euro wordt uitgegeven aan reclame, dan is de voorspelde omzet op basis van de regressievergelijking gelijk aan

- (a) 756000 euro.
- (b) 2296000 euro.
- (c) 96220 euro.
- (d) 3160000 euro.

#### Uitwerking

Stel dat in een nieuwe periode 3000 euro wordt uitgegeven, dat betekent dus dat  $X = 3$ . Invullen in de regressielijn geeft ons  $Y = 220 + 32 \cdot 3 = 316$ . Aangezien  $Y$  gemeten wordt in tienduizenden euro's, betekent dit dat de voorspelde omzet gelijk is aan  $316 \cdot 10000 = 3160000$  euro.

Het juiste antwoord is dus (d).

**Opdracht 13.1:** Een fabrikant van synthetische vezels onderzoekt of het krimpen van de vezels samenhangt met de temperatuur waarbij ze worden gewassen. Er wordt achtmaal

een proef verricht waarbij de vezels gedurende 30 minuten aan een bepaalde temperatuur worden blootgesteld. De geconstateerde krimp werd (in procenten van de oorspronkelijke lengte) als volgt vastgesteld:

Temperatuur (°C)	60	70	80	90	100	75	85	100
Krimp (%)	1,2	1,9	2,8	3,8	4,2	2,6	3,2	4,5

Het verband tussen temperatuur en krimp willen we beschrijven door een regressielijn te berekenen op basis van de waargenomen uitkomsten.

- (a) Welke sommaties moeten we berekenen om de regressiecoëfficiënten te kunnen bepalen?

#### Uitwerking

Bij enkelvoudige regressie is de regressielijn van de vorm  $Y = a + b \cdot X$ , waarbij we de regressiecoëfficiënten  $a$  en  $b$  bepalen aan de hand van

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

We hebben dus de sommaties (of gemiddeldes)  $\sum x_i$ ,  $\sum y_i$ ,  $\sum x_i \cdot y_i$  en  $\sum x_i^2$

- (b) Bereken de sommaties in een rekenschema.

#### Uitwerking

De sommaties worden geïllustreerd in onderstaand rekenschema:

$x$	$y$	$xy$	$x^2$
60	1.2	72	3600
70	1.9	133	4900
80	2.8	224	6400
90	3.8	342	8100
100	4.2	420	10000
75	2.6	195	5625
85	3.2	272	7225
100	4.5	450	10000
<hr/>			
$\bar{x} = 82.5$	$\bar{y} = 3.025$	$\overline{xy} = 263.5$	$\overline{x^2} = 6981.25$

- (c) Bereken de coëfficiënten  $a$  en  $b$  van de lineaire vergelijking.

#### Uitwerking

We kunnen de gemiddeldes uit het bovenstaande rekenschema invullen om de

coëfficiënten te krijgen van de regressielijn.

$$\begin{aligned}b &= \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \\&= \frac{263,5 - 82,5 \cdot 3,025}{6981,25 - (82,5)^2} \\&= \frac{13,938}{175} = 0,0796 \\a &= \bar{y} - b \cdot \bar{x} \\&= 3,025 - 0,0796 \cdot 82,5 \\&= -3,5455.\end{aligned}$$

De formule van de regressielijn behorende bij deze steekproef is dus gelijk aan  $Y = -3,5455 + 0,0796X$ .

(d) Voorspel de krimp als de temperatuur  $X$  de waarde  $65^\circ\text{C}$  heeft.

#### Uitwerking

Een voorspelling voor de krimp bij een temperatuur van  $65^\circ\text{C}$  vinden we simpelweg door  $X = 65$  in te vullen in de regressielijn. Dit geeft een voorspelde waarde van  $Y = -3,5455 + 0,0796 \cdot 65 = 1,63125$ .

Dat betekent dat bij een temperatuur van  $65^\circ\text{C}$  de vezels naar verwachting met ongeveer 1,63% zullen krimpen.

**Opdracht 13.4:** De Stichting voor Veilig Verkeer doet onderzoek naar de invloed van alcohol op het bloedalcoholgehalte (promillage). In een onderzoek onder tien mannen heeft men het promillage gemeten, twee uur na het begin van het drinken van vijf glazen bier. Tevens heeft men van deze tien personen het lichaamsgewicht gemeten. De gegevens staan in de tabel.

Promillage	Gewicht (in kg)
1,06	61
0,77	82
0,72	86
0,95	70
0,65	96
0,83	80
0,99	67
0,73	90
0,84	75
0,96	73

Een onderzoeker heeft het idee dat het gewicht van een persoon invloed heeft op het bloedalcoholgehalte en onderzoekt dit met regressie.

(a) Welke van de variabelen is de te verklaren variabele?

### Uitwerking

De onderzoeker wil toetsen of het gewicht van een persoon invloed heeft op het promillage alcohol in het bloed van testpersonen.

Dit houdt dus in dat de verklarende variabele  $X$  het gewicht (in kg) is en de te verklaren variabele  $Y$  het promillage alcohol in het bloed.

(b) Stel de vergelijking van de regressielijn op.

### Uitwerking

Bij enkelvoudige regressie is de regressielijn van de vorm  $Y = a + b \cdot X$ , waarbij we de regressiecoëfficiënten  $a$  en  $b$  bepalen aan de hand van

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

We moeten dus een tabel construeren met de gemiddelde  $\bar{x}$ ,  $\bar{y}$ ,  $\overline{xy}$  en  $\overline{x^2}$ .

$x$	$y$	$xy$	$x^2$
61	1,06	64,66	3721
82	0,77	63,14	6724
86	0,72	61,92	7396
70	0,95	66,5	4900
96	0,65	62,4	9216
80	0,83	66,4	6400
67	0,99	66,33	4489
90	0,73	65,7	8100
75	0,84	63	5625
73	0,96	70,08	5329
<hr/>			
$\bar{x} = 78 \quad \bar{y} = 0,85 \quad \overline{xy} = 65,013 \quad \overline{x^2} = 6190$			

We hebben nu alle grootheden die benodigd zijn om de coëfficiënten van de regressielijn  $Y = a + b \cdot X$  te kunnen bepalen:

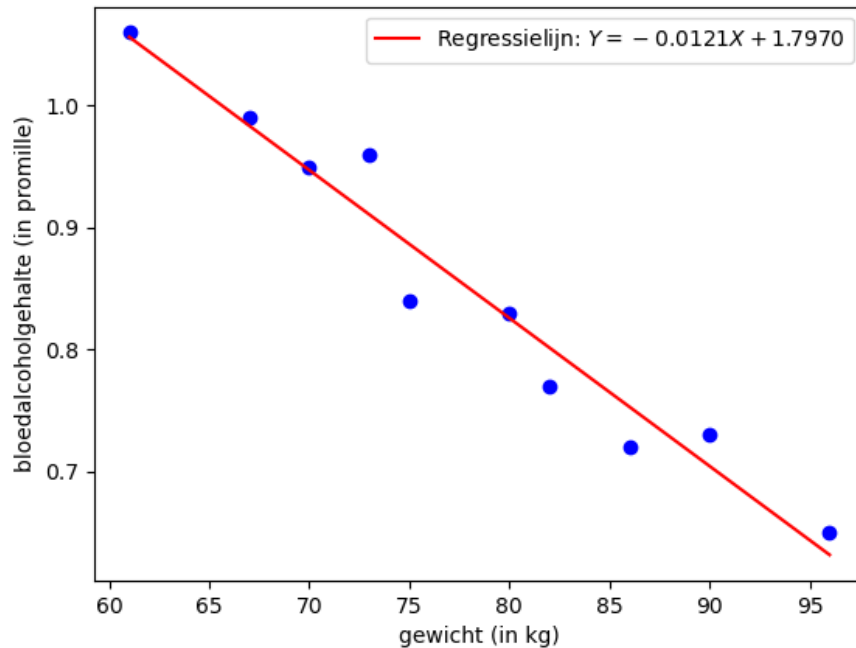
$$\begin{aligned} b &= \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \\ &= \frac{65,013 - 78 \cdot 0,85}{6190 - (78)^2} \end{aligned}$$

$$= \frac{-1,287}{106} \approx -0,0121$$

$$\begin{aligned} a &= \bar{y} - b \cdot \bar{x} \\ &= 0,85 - (-0,0121 \cdot 78) \\ &\approx 1,7970. \end{aligned}$$

De formule van de regressielijn behorende bij deze steekproef is dus gelijk aan  $Y = 1,7970 - 0,0121 \cdot X$ .

Spreadingsdiagram: gewicht (in kg) vs. bloedalcoholgehalte (in promille)



(c) Bereken de correlatiecoëfficiënt.

#### Uitwerking

De correlatiecoëfficiënt  $r(x, y)$  berekenen we op basis van de formule

$$r(x, y) = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}.$$

Dat betekent dat we naast de bovenstaande tabel moeten uitbreiden met een kolom voor  $y^2$ :

$x$	$y$	$xy$	$x^2$	$y^2$
61	1,06	64,66	3721	1,1236
82	0,77	63,14	6724	0,5929
86	0,72	61,92	7396	0,5184
70	0,95	66,5	4900	0,9025
96	0,65	62,4	9216	0,4225
80	0,83	66,4	6400	0,6889
67	0,99	66,33	4489	0,9801
90	0,73	65,7	8100	0,5329
75	0,84	63	5625	0,7056
73	0,96	70,08	5329	0,9216
<hr/>				
$\bar{x} = 78$	$\bar{y} = 0,85$	$\overline{xy} = 65,013$	$\overline{x^2} = 6190$	



De correlatiecoëfficiënt  $r(x, y)$  is dus gelijk aan

$$\begin{aligned} r(x, y) &= \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2) \cdot (\overline{y^2} - \bar{y}^2)}} \\ &= \frac{65,013 - 78 \cdot 0,85}{\sqrt{(78^2 - 6190) \cdot (0,85^2 - 0,739)}} \\ &= \frac{-1,287}{1,318} \\ &\approx -0,9761. \end{aligned}$$

*Sidenote: de correlatiecoëfficiënt  $r(x, y)$  ligt erg dicht in de buurt van  $-1$ , wat inhoudt dat er een sterke negatieve correlatie is tussen de variabelen bloedalcoholgehalte en gewicht. Naarmate iemand zwaarder is, zal het promillage in het bloed ook kleiner zijn bij consumptie van vijf glazen bier.*

- (d) Geef het voorspelde promillage (twee uur na het begin van het drinken van vijf glazen bier) voor een man van 85 kg.

#### Uitwerking

Een voorspelling voor het promillage twee uur na het begin van het drinken van vijf glazen bier bij een man die 85 kg weegt vinden we simpelweg door  $X = 85$  in te vullen in de regressielijn. Dit geeft een voorspelde waarde van  $Y = 1,7970 - 0,0121 \cdot 85 \approx 0,7650$ .

Dat betekent dat bij een gewicht van 85 kg, het bloedalcoholgehalte naar verwachting gelijk is aan 0,7650 promille.

- (e) Uit onderzoek is gebleken: het promillage ligt, twee uur na het begin van het drinken van zes glazen bier (bij mannen) 20 % hoger dan na het drinken van vijf glazen. Stel de vergelijking van de regressielijn voor dit geval op.

#### Uitwerking

Merk op dat we nu alle waarden voor  $y$  moeten vermenigvuldigen met 1,2 (20 % meer dan bij vijf glazen bier). Dat betekent ook dat  $\bar{y}$  en  $\overline{xy}$  met een factor 1,2 groter worden.

De formule van de (nieuwe) regressielijn  $Y = a_n + b_n \cdot X$  kunnen we opstellen met behulp van:

$$\begin{aligned} b_n &= \frac{\overline{xy_n} - \bar{x}_n \cdot \bar{y}_n}{\overline{x_n^2} - (\bar{x}_n)^2} = \frac{(1,2 \cdot \overline{xy}) - \bar{x} \cdot (1,2 \cdot \bar{y})}{\overline{x^2} - (\bar{x})^2} = \frac{1,2 \cdot (\overline{xy} - \bar{x} \cdot \bar{y})}{\overline{x^2} - (\bar{x})^2} = 1,2 \cdot b \\ a_n &= \bar{y}_n - b_n \cdot \bar{x}_n = 1,2 \cdot \bar{y} - 1,2 \cdot b \cdot \bar{x} = 1,2 \cdot (\bar{y} - b \cdot \bar{x}) = 1,2 \cdot a \end{aligned}$$

Dit betekent dus dat zowel het snijpunt  $a$  met de  $y$ -as als de richtingscoëfficiënt  $b$  van de regressielijn met een factor 1,2 groter worden.

De nieuwe regressielijn – uitgaande van consumptie van zes glazen bier – kan

dus worden beschreven met

$$Y = 1,2 \cdot a + 1,2 \cdot b \cdot X = 2,1564 - 0,0146 \cdot X.$$

**Opdracht 13.5:** De weerstand van een blokje van een bepaalde metaalsoort hangt af van de temperatuur. Bij verschillende temperaturen ( $X$ ) werd de weerstand ( $Y$ ) gemeten. De resultaten zijn weergegeven in de volgende tabel:

Temperatuur ( $^{\circ}\text{C}$ )	-100	-50	0	50	100	150	200
Weerstand ( $\Omega$ )	13,2	17,5	21,3	25,4	28,9	32,8	36,8

- (a) Ga ervan uit dat het verband tussen temperatuur en weerstand lineair is. Bepaal de regressielijn met de temperatuur als verklarende variabele.

#### Uitwerking

De regressielijn is van de vorm  $Y = a + b \cdot X$ , waarbij we de regressiecoëfficiënten  $a$  en  $b$  bepalen aan de hand van

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$
$$a = \bar{y} - b \cdot \bar{x}$$

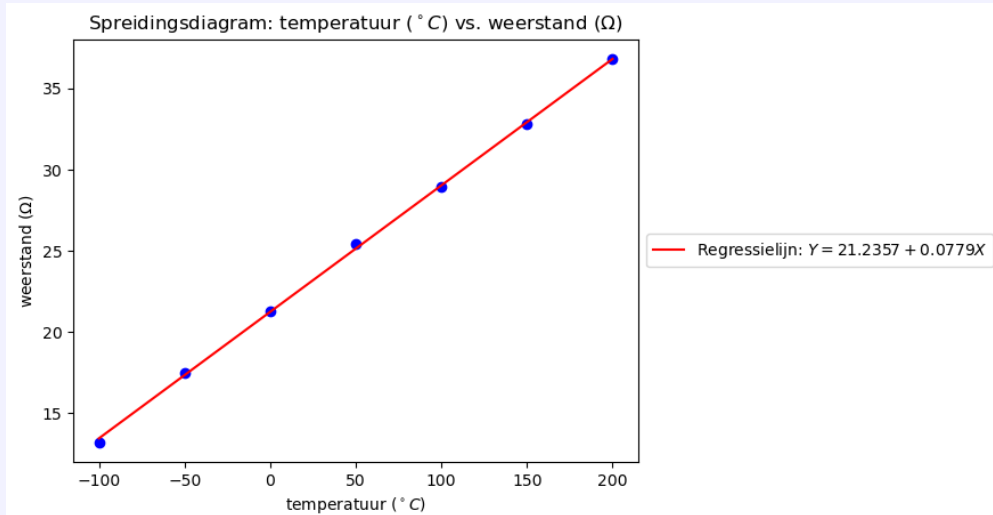
We moeten dus een tabel construeren met de gemiddelde  $\bar{x}$ ,  $\bar{y}$ ,  $\overline{xy}$  en  $\overline{x^2}$ .

$x$	$y$	$xy$	$x^2$	$y^2$
-100	13,2	-1320	10000	174,24
-50	17,5	-875	2500	306,25
0	21,3	0	0	453,69
50	25,4	1270	2500	645,16
100	28,9	2890	10000	835,21
150	32,8	4920	22500	1075,84
200	36,8	7360	40000	1354,24
<hr/>				
$\bar{x} = 50$	$\bar{y} = 25,13$	$\overline{xy} = 2035$	$\overline{x^2} = 12500$	$\overline{y^2} = 692,09$

We hebben nu alle grootheden die benodigd zijn om de coëfficiënten van de regressielijn  $Y = a + b \cdot X$  te kunnen bepalen:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$
$$= \frac{2035 - 50 \cdot 25,1286}{12500 - 50^2}$$
$$= \frac{778,5714}{10000} \approx 0,0779$$
$$a = \bar{y} - b \cdot \bar{x}$$
$$= 25,1286 - 0,0779 \cdot 50$$
$$\approx 21,2357.$$

De formule van de regressielijn behorende bij deze steekproef is dus gelijk aan  $Y = 21,2357 + 0,0779 \cdot X$ .



- (b) Welke weerstand kan men verwachten bij 400 °C?

#### Uitwerking

De weerstand  $Y$  die we verwachten bij een temperatuur  $X$  vinden we door in te vullen in de regressielijn. Hieruit volgt dat bij een temperatuur van  $X = 400$  °C de verwachte weerstand gelijk is aan ongeveer  $Y = 21,2357 + 0,0779 \cdot 400 \approx 52.38$  Ω.

*Sidenote: we zijn in dit geval ver buiten de dataset aan het extrapoleren, want we hebben de regressielijn bepaald op basis van metingen voor temperaturen tussen  $-100$  °C en  $200$  °C. We moeten dus voorzichtig zijn met de aanname dat deze voorspelde waarde van de weerstand bij 400 °C nauwkeurig zou zijn.*

- (c) Bij experimenten als het hier genoemde wordt de temperatuur vaak uitgedrukt in Kelvin (K). Hoe luidt de vergelijking van de regressielijn als de temperatuur in K wordt uitgedrukt? (Het nulpunt van de Kelvin-temperatuurschaal ligt bij  $-273$  °C. Een stijging met 1 °C is gelijk aan een stijging met 1 K.)

#### Uitwerking

Zodra we de temperatuurschaal veranderen van Celsius (°C) naar Kelvin (K), moet bij iedere waarde van  $X$  precies 273 worden opgeteld. Laat nu  $Z$  de temperatuur zijn in K. Dan geldt dus  $Z = X + 273$ , oftewel  $X = Z - 273$ . Invullen in de berekende regressielijn geeft:

$$\begin{aligned} Y &= 21,2357 + 0,0779 \cdot (Z - 273) \\ &= (21,2357 + 0,0779 \cdot -273) + 0,0779 \cdot Z \\ &= -0,0310 + 0,0779 \cdot Z. \end{aligned}$$

Deze lijn heeft dus exact dezelfde richtingscoëfficiënt (en is dus evenwijdig aan de regressielijn met de verklarende variabele  $X$ ), maar heeft een lager gelegen snijpunt met de  $Y$ -as.

**Opdracht 13.9:** Een consumentenorganisatie wil onderzoeken hoe de samenhang is tussen de ouderdom van automobielen en de jaarlijkse kosten voor onderhoud. Voor auto's van

het merk Inceatti zijn tien exemplaren in het onderzoek betrokken geweest. Dit leverde de volgende gegevens:

Auto nummer	Leeftijd (X) in jaren	Onderhoudskosten (Y) in de laatste 12 maanden
A	1,6	450
B	6,8	930
C	5,1	670
D	7,2	920
E	3,5	640
F	6,2	900
G	9,4	1440
H	5,5	760
I	8,0	1260
J	2,7	530
Totaal	56,0	8500

Verder zijn de volgende sommaties gegeven:

$$\sum x_i = 56, \sum y_i = 8500, \sum x_i^2 = 367,24, \sum y_i^2 = 8102000 \text{ en } \sum x_i y_i = 54132.$$

- (a) Bereken de regressielijn waardoor de onderhoudsuitgaven worden verklaard uit de leeftijd.

#### Uitwerking

De regressielijn is van de vorm  $Y = a + b \cdot X$ , waarbij we de regressiecoëfficiënten  $a$  en  $b$  bepalen aan de hand van

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Veel van het rekenwerk is al gegeven in de opdracht. Het is dus niet nodig de hele rekentabel uit te werken. We hoeven alleen maar de sommaties om te zetten naar gemiddeldes door te delen door de steekproefgrootte  $n = 10$ . Dit geeft als gemiddeldes

$$\bar{x} = 5,6, \quad \bar{y} = 850, \quad \overline{x^2} = 36,724, \quad \overline{y^2} = 810200, \quad \overline{xy} = 5413,2.$$

We hebben nu alle grootheden die benodigd zijn om de coëfficiënten van de regressielijn  $Y = a + b \cdot X$  te kunnen bepalen:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$= \frac{5413,2 - 5,6 \cdot 850}{36,724 - (5,6)^2}$$

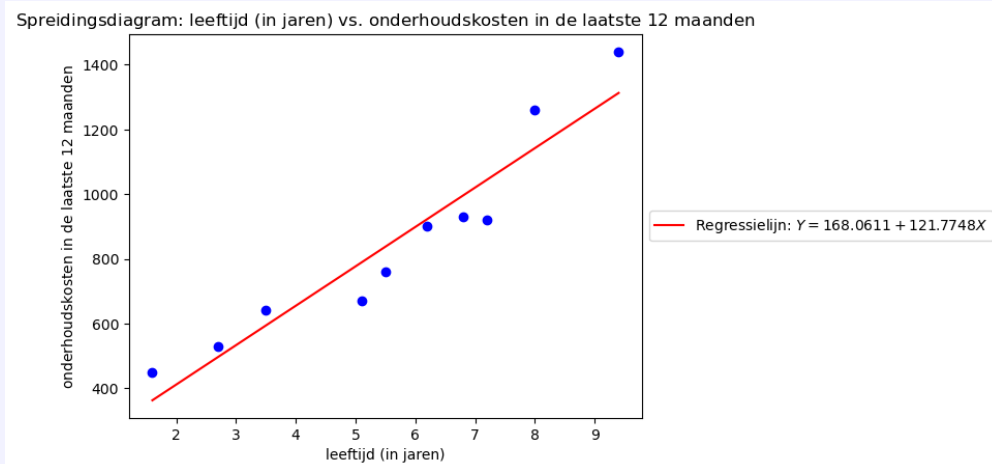
$$= \frac{653,2}{5,364} \approx 121,7748$$

$$a = \bar{y} - b \cdot \bar{x}$$

$$= 850 - 121,7748 \cdot 5,6$$

$$\approx 168,0611.$$

De formule van de regressielijn behorende bij deze steekproef is dus gelijk aan  $Y = 168,0611 + 121,7748 \cdot X$ .



- (b) Bereken op basis van de gegevens van (a) een 95%-voorspellingsinterval voor de onderhoudskosten van een willekeurige auto van tien jaar oud.

### Uitwerking

De eerste stap is om de puntschatting voor  $Y$  te bepalen aan de hand van de regressielijn  $Y = 168,0611 + 121,7748X$  door  $X = 10$  in te vullen. Dit geeft ons een puntschatting van  $y_0 = 168,0611 + 121,7748 \cdot 10 \approx 1385,8091$ . Daarnaast kunnen we de standaardafwijking  $\sigma$  van de storingsterm  $\varepsilon$  schatten:

$$\begin{aligned}
 s_\varepsilon &= \sqrt{\frac{n}{n-2} \cdot (\overline{y^2} - a \cdot \bar{y} - b \cdot \overline{xy})} \\
 &= \sqrt{\frac{10}{8} \cdot (810200 - 168,0611 \cdot 850 - 121,7748 \cdot 5413,2)} \\
 &\approx 100,9746
 \end{aligned}$$

Vervolgens kunnen we een puntschatting berekenen van de standaardafwijking van  $Y$  gegeven  $X = x_0$ :

$$\begin{aligned}
 s_f &= s_\varepsilon \cdot \sqrt{1 + \frac{1}{n} \cdot \left(1 + \frac{(x_0 - \bar{x})^2}{(\overline{x^2} - \bar{x}^2)}\right)} \\
 &= 100,9746 \cdot \sqrt{1 + \frac{1}{10} \cdot \left(1 + \frac{(10 - 5,6)^2}{(36,724 - 5,6^2)}\right)} \\
 &\approx 122,0468
 \end{aligned}$$

Omdat we de standaardafwijkingen geschat hebben en de storingstermen normaal verdeeld zijn, moeten we werken met de  $t$ -verdeling met  $df = n - 2 = 8$  vrijheidsgraden. De grenswaarde die hoort bij een betrouwbaarheidsniveau  $\alpha = 0,05$  is gelijk aan

$$t = \text{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 2) = \text{InvT}(\text{opp} = 0,975; \text{df} = 8) \approx 2,3060$$

Het 95%-betrouwbaarheidsinterval voor de gemiddelde  $Y$  voor gegeven  $X = x_0$  kan dus worden beschreven door

$$\begin{aligned} & [y_0 - t \cdot s_f; y_0 + t \cdot s_f] \\ & = [1385,8091 - 2,3060 \cdot 122,0468; 1385,8091 + 2,3060 \cdot 122,0468] \\ & \approx [1104,3688; 1667,2494] \end{aligned}$$

Met 95% betrouwbaarheid liggen de onderhoudskosten in de laatste 12 maanden van een willekeurige Inceatti auto van 10 jaar oud tussen ongeveer 1104 en 1667 euro.

- (c) Veronderstel eens dat de hier bedoelde regressielijn exact bekend zou zijn, bijvoorbeeld omdat een alomvattend onderzoek naar onderhoudskosten heeft plaatsgevonden waarbij alle 50000 Inceatti-auto's die in Nederland rondrijden, betrokken waren. Dit bleek een regressielijn op te leveren met  $\alpha = 320$  en  $\beta = 130$ . Verder geldt voor de storingsterm  $\sigma_\varepsilon = 65$ . Geef hiermee een 99%-voorspellingsinterval voor het bedrag aan jaarlijkse onderhoudskosten voor een willekeurige Inceatti-auto van tien jaar oud.

#### Uitwerking

Het daadwerkelijke verband tussen de variabelen “leeftijd (in jaren)” ( $X$ ) en “onderhoudskosten in de laatste 12 maanden” ( $Y$ ) is gelijk aan

$$Y = \alpha + \beta \cdot X + \varepsilon = 320 + 130 \cdot X + \varepsilon,$$

waarbij de storingsterm  $\varepsilon \sim N(0, 65)$ .

Voor een willekeurige Inceatti-auto van tien jaar oud geldt dus  $X = 10$ . Dat betekent dat we  $Y$  volledig kunnen beschrijven met alleen de storingsterm  $\varepsilon$  als kansvariabele:

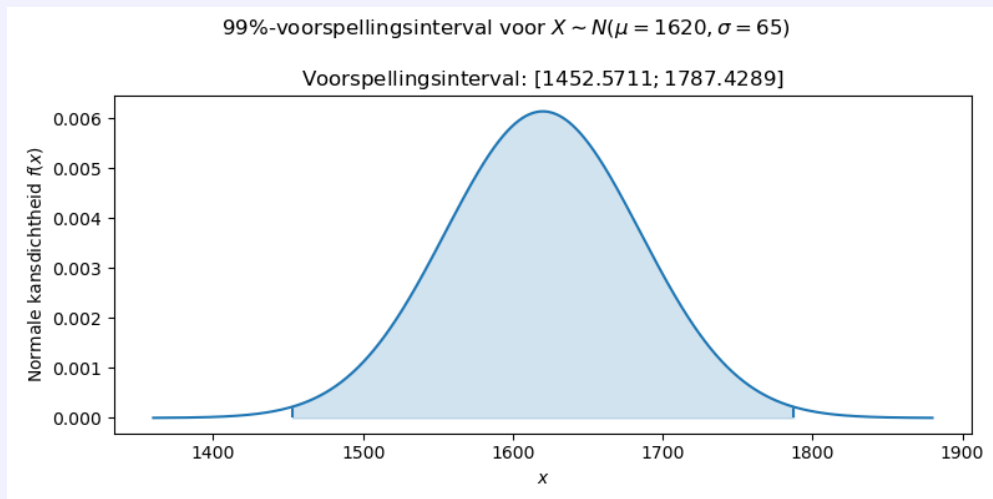
$$Y = 320 + 130 \cdot 10 + \varepsilon = 1620 + \varepsilon.$$

We kunnen  $Y$  dus als een kansvariabele zien die normaal verdeeld is met verwachtingswaarde  $\mu = 1620$  en standaardafwijking  $\sigma = 65$ . Omdat we een 99%-voorspellingsinterval willen bepalen (dus het betrouwbaarheidsniveau  $\alpha = 0,01$ ), geldt dat  $z_{\alpha/2} = \text{InvNorm}(\text{opp} = 1 - \alpha/2) = \text{InvNorm}(\text{opp} = 0,995) \approx 2,5758$ .

Het 99%-voorspellingsinterval voor een volgende uitkomst van  $Y$  wordt dan gegeven door

$$\begin{aligned} & [\mu - z_{\alpha/2} \cdot \sigma; \mu + z_{\alpha/2} \cdot \sigma] \\ & = [1620 - 2,5758 \cdot 65; 1620 + 2,5758 \cdot 65] \\ & \approx [1452,5711; 1787,4289] \end{aligned}$$

Voor een willekeurige Inceatti-auto van tien jaar oud zullen met 99% zekerheid de onderhoudskosten in de laatste 12 maanden tussen de 1452 en 1788 euro bedragen.



**Opdracht 13.11:** Voor een groep studenten wordt een tweetal test scores vastgesteld. Een voor de  $A$ -vakken en een voor de  $B$ -vakken. De resultaten waren als volgt:

Student	$A$	$B$	$C$	$D$	$E$	$F$	$G$	$H$
$A$ -score	40	60	50	45	74	53	45	42
$B$ -score	88	35	45	80	35	60	70	85

Bereken hiervoor de rangcorrelatiecoëfficiënt van Spearman.

#### Uitwerking

De eerste stap bij het berekenen van Spearman's correlatiecoëfficiënt is het bepalen van de rankings van de uitkomsten voor de  $A$ - en  $B$ -scores:

<b>Rangnummers <math>A</math>-scores</b>	1	7	5	3,5	8	6	3,5	2
<b>Rangnummers <math>B</math>-scores</b>	8	1,5	3	6	1,5	4	5	7
<b>Verschillen <math>d_i</math></b>	-7	5,5	2	-2,5	6,5	2	-1,5	-5
<b>Kwadratische verschillen <math>d_i^2</math></b>	49	30,25	4	6,25	42,25	4	2,25	25

De som van de kwadratische rangnummerverschillen is gelijk aan  $\sum_i d_i^2 = 163$ . Aangezien de steekproefgrootte gelijk is aan  $n = 8$ , is de rangcorrelatiecoëfficiënt van Spearman gelijk aan

$$\begin{aligned}
 r_s &= 1 - \frac{6 \cdot \sum_i d_i^2}{n^3 - n} \\
 &= 1 - \frac{6 \cdot 163}{8^3 - 8} \\
 &\approx -0,9405.
 \end{aligned}$$

Omdat de rangcorrelatiecoëfficiënt  $r_s$  dichtbij  $-1$  ligt, geldt dat de rangnummers een nagenoeg tegengestelde samenhang vertonen.

*Side note: merk op dat we in deze opgave meerdere keren dezelfde  $A$ -scores hebben waargenomen (in dit bijvoorbeeld 45). In dat geval zijn deze respectievelijk op de derde en vierde plaats in de ranking (van laag naar hoog). Omdat je niet eenduidig kunt zeggen*

welke van de twee derde is en welke vierde, wordt uitgegaan van het gemiddelde van de rankings, in dit geval dus 3,5.

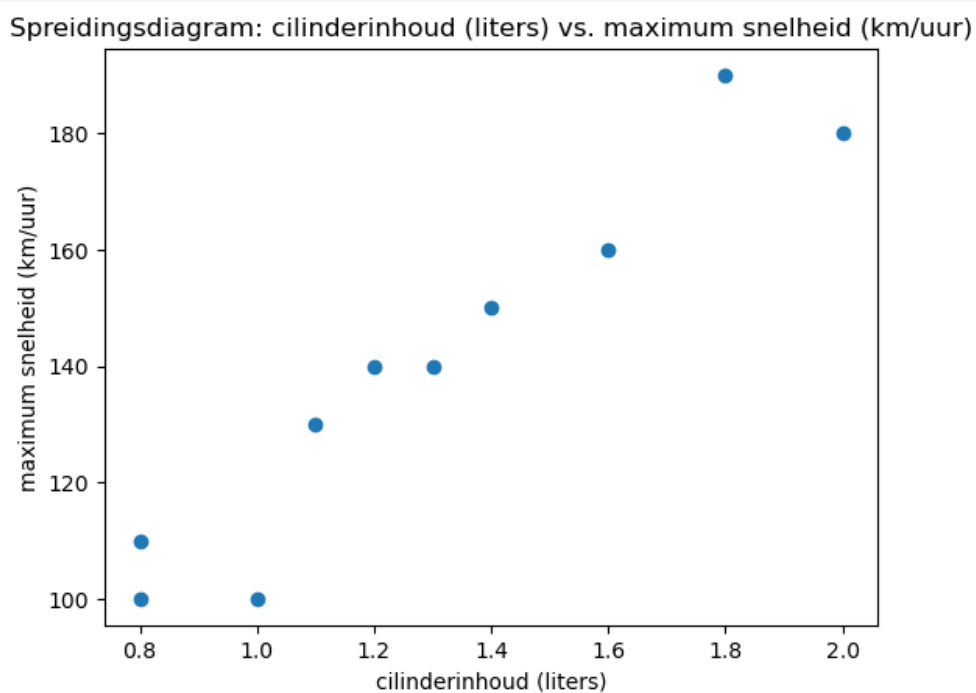
**Opdracht 13.15:** Er is een onderzoek gehouden naar het verband tussen motorinhoud en maximumsnelheid van personenauto's. Er werden tien auto's getest. De resultaten waren als volgt:

Auto nr.	Cilinderinhoud	Max. snelheid (km/uur)
<i>I</i>	1,2	140
<i>II</i>	0,8	110
<i>III</i>	0,8	100
<i>IV</i>	2,0	180
<i>V</i>	1,4	150
<i>VI</i>	1,0	100
<i>VII</i>	1,6	160
<i>VIII</i>	1,8	190
<i>IX</i>	1,3	140
<i>X</i>	1,1	130

(a) Teken de gegevens in een spreidingsdiagram.

#### Uitwerking

Bij het tekenen van een spreidingsdiagram is het belangrijk om van te voren te bepalen welke variabele de verklarende (onafhankelijke) variabele  $X$  is en welke variabele de te verklaren (afhankelijke) variabele  $Y$  is. In dit geval is het het meest logisch om aan de hand van de cilinderinhoud te bepalen wat de te bereiken maximum snelheid is van een auto.



(b) Bepaal het verband tussen de variabelen met behulp van lineaire regressie.



### Uitwerking

Uit het spreidingsdiagram is op te merken dat het verband tussen de variabelen cilinderinhoud en maximum snelheid redelijk lineair is. Er valt dus zeker iets te zeggen voor de keuze om lineaire regressie toe te passen.

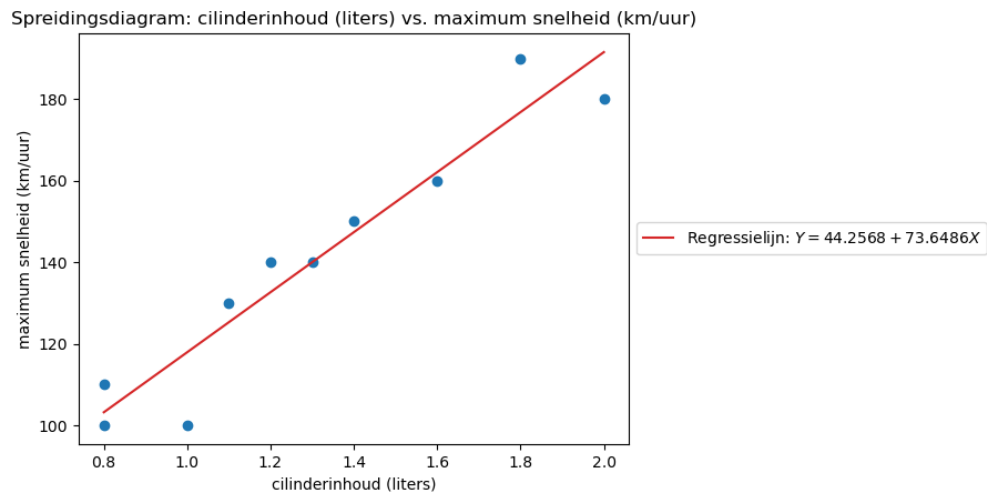
Om de bijbehorende regressielijn te bepalen, construeren we allereerst de reken-tabel met de gemiddeldes  $\bar{x}$ ,  $\bar{y}$ ,  $\overline{xy}$ ,  $\overline{x^2}$  en  $\overline{y^2}$ :

$x$	$y$	$xy$	$x^2$	$y^2$
1,2	140	168	1,44	19600
0,8	110	88	0,64	12100
0,8	100	80	0,64	10000
2,0	180	360	4,0	32400
1,4	150	210	1,96	22500
1,0	100	100	1,0	10000
1,6	160	256	2,56	25600
1,8	190	342	3,24	36100
1,3	140	182	1,69	19600
1,1	130	143	1,21	16900
<hr/>				
$\bar{x} = 1,3 \quad \bar{y} = 140 \quad \overline{xy} = 192,9 \quad \overline{x^2} = 1,838 \quad \overline{y^2} = 20480$				

Op basis van de bovenstaande tabel kunnen we de coëfficiënten  $a$  en  $b$  van de regressielijn bepalen:

$$\begin{aligned} b &= \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \\ &= \frac{192,9 - 1,3 \cdot 140}{1,838 - (1,3)^2} \\ &= \frac{10,9}{0,148} \approx 73,6486 \\ a &= \bar{y} - b \cdot \bar{x} \\ &= 140 - 73,6486 \cdot 1,3 \\ &\approx 44,2568. \end{aligned}$$

De formule van de regressielijn behorende bij deze steekproef is dus gelijk aan  $Y = 44,2568 + 73,6486X$ .



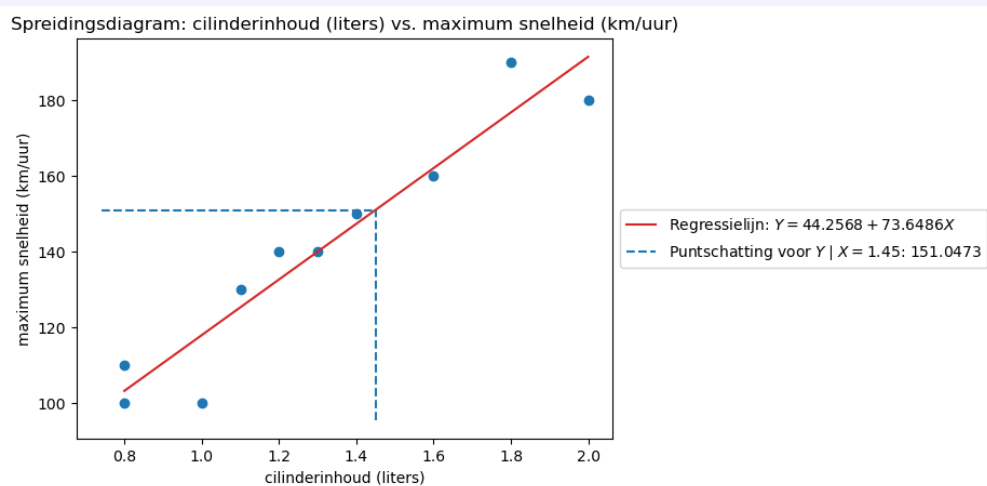
- (c) Voorspel aan de hand van en de gevonden lijn de maximumsnelheid van een auto met een cilinderinhoud van 1,5 liter.

#### Uitwerking

Aan de hand van de regressielijn kunnen we een voorspelling doen voor de maximumsnelheid van een auto met een cilinderinhoud van 1,5 liter. Dit doen we door simpelweg  $X = 1,5$  in te vullen en de bijbehorende  $Y$ -waarde te berekenen:

$$Y = a + b \cdot X = 44,2568 + 73,6486 \cdot 1,5 = 154,7297.$$

De verwachting is dus dat een willekeurige auto met een cilinderinhoud van 1,5 liter een maximumsnelheid heeft van 155 kilometer per uur.



- (d) Geef een schatting van de variantie van de storingsterm.

#### Uitwerking

Bij lineaire regressie nemen we aan dat de storingsterm  $\varepsilon$  normaal verdeeld is met verwachtingswaarde  $\mu_\varepsilon = 0$  en (onbekende) standaardafwijking  $\sigma_\varepsilon$ . We kunnen een schatting van de variantie  $\sigma^2$  berekenen aan de hand van de volgende

formule:

$$\begin{aligned}s_{\varepsilon}^2 &= \frac{n}{n-2} \cdot (\overline{y^2} - a \cdot \bar{y} - b \cdot \overline{xy}) \\&= \frac{10}{8} \cdot (20480 - 44,2568 \cdot 140 - 73,6486 \cdot 192,9) \\&\approx 96,5372.\end{aligned}$$

- (e) Geef een 95 %-voorspellingsinterval voor de maximumsnelheid van een auto waarvan is gegeven dat deze een cilinderinhoud van 1,45 liter heeft.

#### Uitwerking

We willen een voorspellingsinterval bepalen voor de maximumsnelheid van een willekeurige auto met een cilinderinhoud van 1,45 liter.

De eerste stap is om de puntschatting voor  $Y$  te bepalen aan de hand van de regressielijn  $Y = 44,2568 + 73,6486 \cdot X$  door  $X = 1,45$  in te vullen. Dit geeft ons een puntschatting van  $y_0 = 44,2568 + 73,6486 \cdot 1,45 \approx 151,0473$ . Daarnaast kunnen we de standaardafwijking  $\sigma$  van de storingsterm  $\varepsilon$  schatten:

$$\begin{aligned}s_{\varepsilon} &= \sqrt{\frac{n}{n-2} \cdot (\overline{y^2} - a \cdot \bar{y} - b \cdot \overline{xy})} \\&= \sqrt{\frac{10}{8} \cdot (20480 - 44,2568 \cdot 140 - 73,6486 \cdot 192,9)} \\&\approx 9,8253.\end{aligned}$$

Vervolgens kunnen we een puntschatting berekenen van de standaardafwijking van  $Y$  voor gegeven  $X = x_0$ :

$$\begin{aligned}s_f &= s_{\varepsilon} \cdot \sqrt{1 + \frac{1}{n} \cdot \left(1 + \frac{(x_0 - \bar{x})^2}{\overline{x^2} - \bar{x}^2}\right)} \\&= 9,8253 \cdot \sqrt{1 + \frac{1}{10} \cdot \left(1 + \frac{(1,45 - 1,3)^2}{1,838 - 1,3^2}\right)} \\&\approx 10,3759.\end{aligned}$$

Omdat we de standaardafwijkingen geschat hebben en de storingstermen normaal verdeeld zijn, moeten we werken met de  $t$ -verdeling met  $df = n - 2 = 8$  vrijheidsgraden. De  $t$ -waarde die hoort bij een betrouwbaarheidsniveau  $\alpha = 0,05$  is gelijk aan

$$t = \text{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 2) = \text{InvT}(\text{opp} = 0,975; \text{df} = 8) \approx 2,3060.$$

Het 95 %-betrouwbaarheidsinterval voor de gemiddelde  $Y$  voor gegeven  $X = 1,45$  kan dus worden beschreven door

$$\begin{aligned}&[y_0 - t \cdot s_f; y_0 + t \cdot s_f] \\&= [151,0473 - 2,3060 \cdot 10,3759; 151,0473 + 2,3060 \cdot 10,3759] \\&\approx [127,1205; 174,9741].\end{aligned}$$

Met 95 % betrouwbaarheid ligt de maximumsnelheid van een willekeurige auto met een cilinderinhoud van 1,45 liter tussen ongeveer 127 en 175 kilometer per uur.

