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 grootheid 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 α de constante term is en β 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

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

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 geillustreerd 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
$\overline{x} = 82.5$	$\overline{y} = 3.025$	$\overline{xy} = 263.5$	$\overline{x^2} = 6981.25$

(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.

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

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 $65\,^{\circ}\mathrm{C}$ heeft.

Uitwerking

Een voorspelling voor de krimp bij een temperatuur van $65\,^{\circ}\mathrm{C}$ 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 $65\,^{\circ}\mathrm{C}$ 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

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

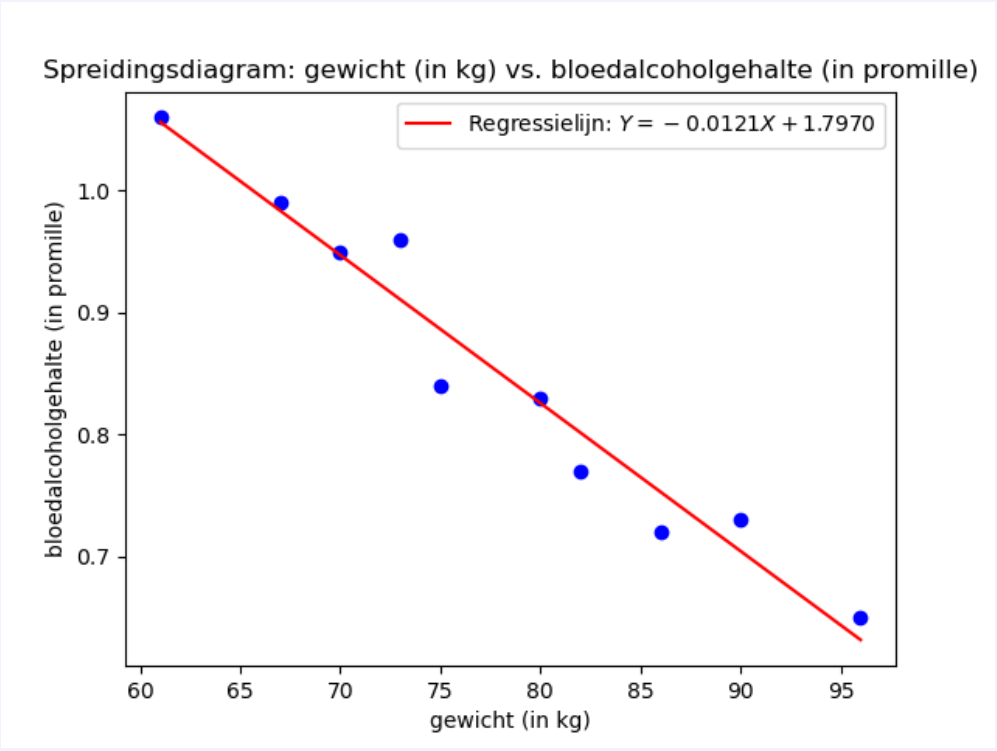
We moeten dus een tabel construeren met de gemiddelde $\overline{x}$, $\overline{y}$, $\overline{xy}$ en $\overline{x^2}$.

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
$\overline{x} = 78$	$\overline{y} = 0,85$	$\overline{xy} = 65,013$	$\overline{x^2} = 6190$

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

$$\begin{aligned} b &= \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \\ &= \frac{65,013 - 78 \cdot 0,85}{6190 - (78)^2} \\ &= \frac{-1,287}{106} \approx -0,0121 \\ a &= \overline{y} - b \cdot \overline{x} \\ &= 0,85 - (-0,0121 \cdot 78) \\ &\approx 1,7970. \end{aligned}$$

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



(c) Bereken de correlatiecoëfficiënt.

Uitwerking

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

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

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	$0,9216$
$\overline{x} = 78$	$\overline{y} = 0,85$	$\overline{xy} = 65,013$	$\overline{x^2} = 6190$	

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

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

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 $85\,\text{kg}$.

Uitwerking

Een voorspelling voor het promillage twee uur na het begin van het drinken van vijf glazen bier bij een man die $85\,\text{kg}$ 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 $85\,\text{kg}$, 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 waardes voor y moeten vermenigvuldigen met $1,2$ ($20\,\%$ meer dan bij vijf glazen bier). Dat betekent ook dat $\overline{y}$ en $\overline{xy}$ met een factor $1,2$ groter worden.

De formule van de (nieuwe) regressielijn $Y = a_\text{n} + b_\text{n} \cdot X$ kunnen we opstellen met behulp van:

$$b_\text{n} = \frac{\overline{xy}_\text{n} - \overline{x}_\text{n} \cdot \overline{y}_\text{n}}{\overline{x^2}_\text{n} - (\overline{x}_\text{n})^2} = \frac{(1,2 \cdot \overline{xy}) - \overline{x} \cdot (1,2 \cdot \overline{y})}{\overline{x^2} - (\overline{x})^2} = \frac{1,2 \cdot (\overline{xy} - \overline{x} \cdot \overline{y})}{\overline{x^2} - (\overline{x})^2} = 1,2 \cdot b$$

$$a_\text{n} = \overline{y}_\text{n} - b_\text{n} \cdot \overline{x}_\text{n} = 1,2 \cdot \overline{y} - 1,2 \cdot b \cdot \overline{x} = 1,2 \cdot (\overline{y} - b \cdot \overline{x}) = 1,2 \cdot a$$

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 (°C)	-100	-50	0	50	100	150	200
Weerstand (Ω)	$13,2$	$17,5$	$21,3$	$25,4$	$28,9$	$32,8$	$36,8$

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

Uitwerking

(b) Welke weerstand kan men verwachten bij $400\,°\mathrm{C}$?

Uitwerking

(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 $-273\,°\mathrm{C}$. Een stijging met $1\,°\mathrm{C}$ is gelijk aan een stijging met $1\,\mathrm{K}$.)

Uitwerking

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

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

Uitwerking

(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 $\sigma_e = 65$. Geef hiermee een 99%-voorspellingsinterval voor het bedrag aan jaarlijkse onderhoudskosten voor een willekeurige Inceatti-auto van tien jaar oud.

Uitwerking

Opdracht 13.11: Voor een groep studenten wordt een tweetal testscores 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

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	Onderhoudskosten (Y) in de laatste 12 maanden
I	$1,2$	140
II	$0,8$	110
III	$0,8$	100
IV	$2,0$	180
V	$1,4$	150
VI	$1,0$	100
VII	$1,6$	160
$VIII$	$1,8$	190
IX	$1,3$	140
X	$1,1$	130

(a) Teken de gegevens in een spreidingsdiagram.

Uitwerking

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

Uitwerking

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

Uitwerking

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

Uitwerking

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

Uitwerking