Inhoudstafel

Hoofdstuk 7
Correlatie- en regressieanalyse

1. Symmetrische associatiematen voor kenmerken op metrisch niveau

Dit hoofdstuk is een van de belangrijkste in een inleidend handboek voor toegepaste statistiek voor sociale wetenschappers. Dit hoofdstuk handelt over de lineaire samenhang tussen metrische kenmerken. Je kan je afvragen waarom deze methoden zo populair zijn in de criminologie. Je zou kunnen stellen: veel variabelen zijn toch niet echt van het metrische meetniveau? Veel onderzoek in de criminologie is gebaseerd op indicatoren die worden samengesteld op basis van vragenlijsten. Veel vragenlijsten zijn gebaseerd op uitspraken en die uitspraken hebben een beperkt aantal antwoordcategorieën. De antwoorden uit onderzoek met vragenlijsten kunnen worden verwerkt tot indicatoren die heel gedetailleerde informatie bevatten, alvast informatie die gedetailleerd genoeg is om lineaire correlatie-analyse en lineaire regressie-analyse op toe te passen. Daarnaast bestaan er natuurlijk wel een heleboel kenmerken die van nature uit metrisch zijn van aard. Een voorbeeld is het aantal delicten dat jongeren plegen binnen de tijdspanne van een jaar, het aantal maal dat men slachtoffer wordt van een misdrijf, het aantal keren dat men in de gevangenis werd opgesloten voor een druggerelateerd feit, het aantal (buitenechtelijke) kinderen dat men heeft, hoeveel maanden men leefloon geniet, enz.

Al deze kenmerken kunnen gemeten worden met een grote precisie, en kunnen met criminaliteit of de maatschappelijke reactie op criminaliteit in verband gebracht worden. Criminologen hebben zich sinds de negentiende eeuw heel intensief bezig gehouden met dergelijke vraagstellingen. De biologische school zocht naar biologische oorzaken van delinquent gedrag, de psychologische school zocht naar verbanden tussen de scores op psychologische testen en regelovertredend gedrag. Zulke studies leveren een antwoord op de vraag of er bijvoorbeeld een verband bestaat tussen de scores van iemand op een psychopathieschaal en de frequentie waarmee een persoon gewelddadige handelingen stelt. Dat zijn heel belangrijke vragen want ze leren ons iets over kenmerken zoals psychopathie en de kenmerken waarmee dit samenhangt. Vandaag de dag neemt men in de reclassering en hulpverlening vaak het geïntegreerd biospychosociaal model, waarbij gekeken wordt naar de samenhang tussen kenmerken gemeten op individueel niveau en kenmerken gemeten op groepsniveau als uitgangspunt.

Eén van de meest volledige overzichtswerken van alle correlaties is het grote correlatiehandboek "The Handbook of Crime Correlates" van de criminologen Lee Ellis, Kevin Beaver en John Wright. Dit handboek bevat een overzicht van kenmerken, gerangschikt volgens type en sterkte.

Om al die interessante studies te kunnen begrijpen, is het belangrijk dat je basisinzichten verkrijgt in de regressie- en correlatieanalyse. Stel je voor dat je niks kent van deze associatiematen en je gebruikt resultaten uit het eerste en beste artikeltje uit een tijdschrift in het kader van een paper die je moet schrijven of erger nog: in het kader van je scriptie. De kans is reëel dat je het artikel foutief interpreteert en ook dat je niet in staat bent het kaf van het koren te onderscheiden. Het duurt wel eventjes voor je dat kan, maar Rome werd ook niet op één dag gebouwd. Zo is het ook gesteld in de criminologie-opleiding. De criminoloog wordt ook niet op een dag gevormd. Gelukkig maar. Criminologen die gedetineerden loslaten op basis van foutief geïnterpreteerde tests, zijn geen goede zaak voor de samenleving. De tijd dat je dacht dat enkel ingenieurs een goede statistische opleiding moeten hebben, is wel degelijk voorbij. Je ontsnapt dus niet aan de lineaire regressie en correlatieanalyse. Maar geloof het of niet, de correlatie- en regressieanalyse kunnen best interessant zijn eenmaal je de logica ervan doorhebt. In het tijdperk van "Big Data" waar alle informatie voor het rapen ligt, en waar zoveel digitale informatie wordt bijgehouden, kan je als (criminologische) data-analist goed aan de slag en beleef je gouden tijden. Genoeg reclamepraat echter. We komen terug tot de orde van de dag.

De lineaire samenhang tussen twee kenmerken van het metrisch niveau kan worden bestudeerd aan de hand van de **covariatie, de covariantie en de correlatiecoëfficiënt**. Deze symmetrische associatiematen zijn verwant aan elkaar. Zij vormen de basis om de bivariate regressieanalyse te begrijpen. Aan regressieanalyse wordt in een volgende paragraaf aandacht besteed. Stel dat we geïnteresseerd zijn in de samenhang tussen de criminaliteitsgraad in Gentse buurten en het werkloosheidspercentage in diezelfde Gentse buurten. We verzamelen deze metrische gegevens voor alle Gentse buurten. Als blijkt dat hoge criminaliteitsgraden in buurten samenhangen met hoge werkloosheidspercentages, dan is er sprake van een positieve samenhang. Omgekeerd, als blijkt dat hoge criminaliteitsgraden samenhangen met lage werkloosheidsgraden, dan is er sprake van negatieve samenhang. Als er geen samenhang bestaat tussen beide metrische kenmerken dan zal de covariatie, covariantie maar ook de correlatie nul bedragen.

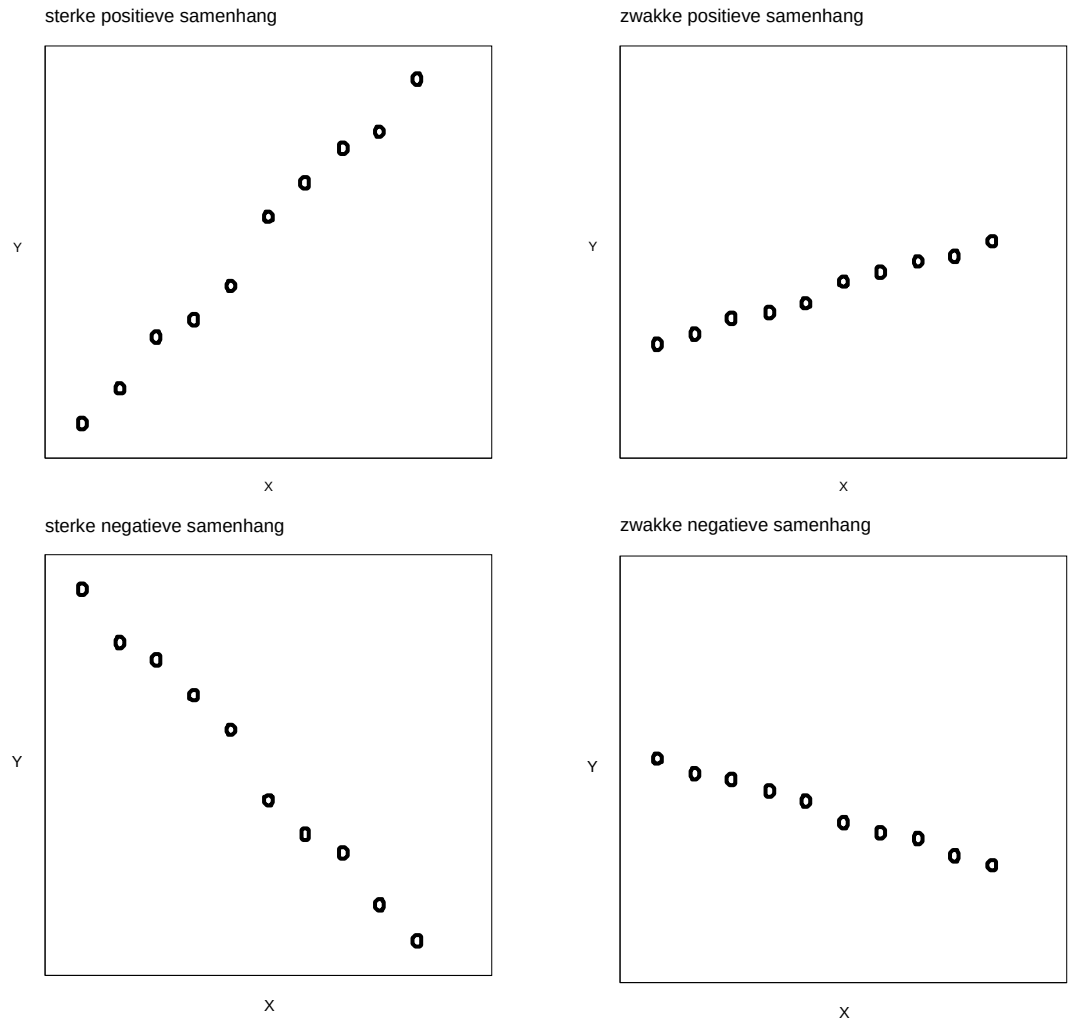
Om de bivariate lineaire samenhang tussen twee metrische kenmerken beter te begrijpen, kunnen we best beroep doen op een puntenwolk of **scatterplot**. Dit principe hebben we eerder al eens kort uitgelegd. Elk punt representeert een statistische eenheid (bijvoorbeeld een individu, een buurt) en elk punt bevat informatie over een X-variabele en een Y-variabele. Elk punt heeft dus wat we noemen xy-coördinaten. Deze coördinaten worden gegeven voor de kenmerken waarin men geïnteresseerd is, zoals de criminaliteitsgraad en de werkloosheidsgraad. Alle punten vormen samen een puntenwolk.

De puntenwolk is dus de verzameling van alle elementen uit onze steekproef waarbij geldt dat we voor elk element de waarde op de X-variabele en de waarde op de Y-variabele kunnen aflezen.

In de tekening hieronder zien we vier verschillende situaties. We bespreken ze even.

- Links bovenaan kan een puntenwolk gezien worden, waaruit een heel sterk positief verband blijkt te bestaan. Hoge waarden op de X-variabele gaan samen met hoge waarden op de Y-variabele. We spreken van een sterke positieve samenhang.
- Rechts bovenaan zien we eveneens een puntenwolk, maar de relatie tussen X en Y is toch minder uitgesproken. Je moet al veel beter je best doen om daar een sterk verband in te willen zien. Toegegeven, er zit een patroon is, maar de stijging is maar matig. We spreken hier duidelijk van een matig positief verband.
- Links onderaan kan een puntenwolk gezien worden, waaruit een heel sterk negatief verband blijkt te bestaan. Hoge waarden op de X-variabele gaan samen met lage waarden op de Y-variabele. We spreken van een sterke negatieve samenhang.
- Rechts onderaan zien we eveneens een puntenwolk, maar de relatie tussen X en Y is toch minder uitgesproken. Je moet al veel beter je best doen om daar een sterk verband in te willen zien. Toegegeven, er zit een patroon in, maar de daling is maar matig. We spreken hier duidelijk van een matig negatief verband.

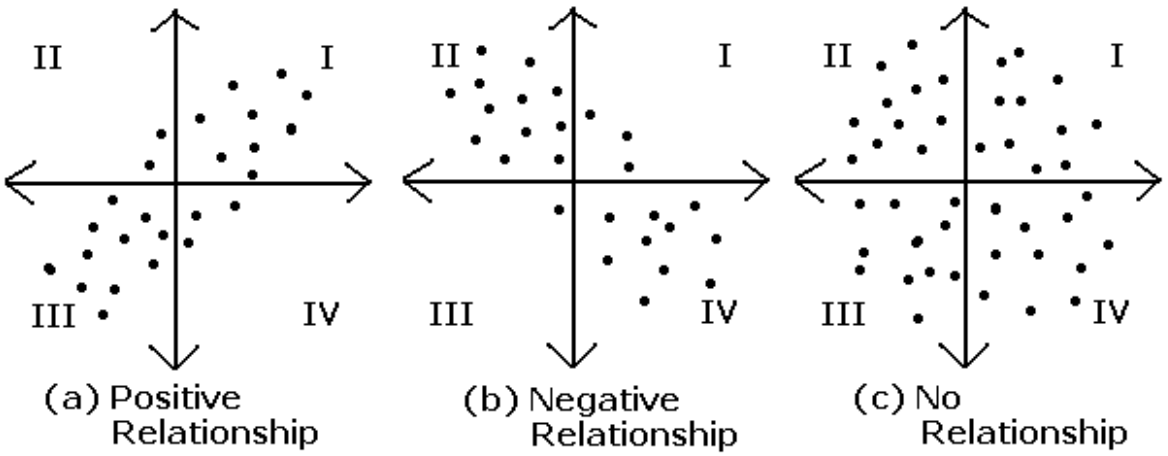
Vormen van lineaire samenhang tussen metrische kenmerken



Omdat elk van de beide kenmerken van het metrische niveau is, kunnen we van elk punt gemakkelijk het rekenkundig gemiddelde berekenen. We kunnen een denkbeeldig punt tekenen in de puntenwolk: het punt gegeven door de gemiddelde score op de X-variabele en de gemiddelde score op de Y-variabele, noemen we **het centrale punt van de puntenwolk.** Een nog andere benaming hiervoor is het **zwaartepunt van de tweedimensionele verdeling**. Deze tweedimensionele tegenhanger van het rekenkundige gemiddelde (gedragen door de

gemiddelde waarde op variabele X en de gemiddelde waarde op variabele Y) is letterlijk het "zwaartepunt" van de verdeling: het geeft een tweedimensionele centrale waarde in het vlak (x,y). Laat ons beginnen met een illustratie.

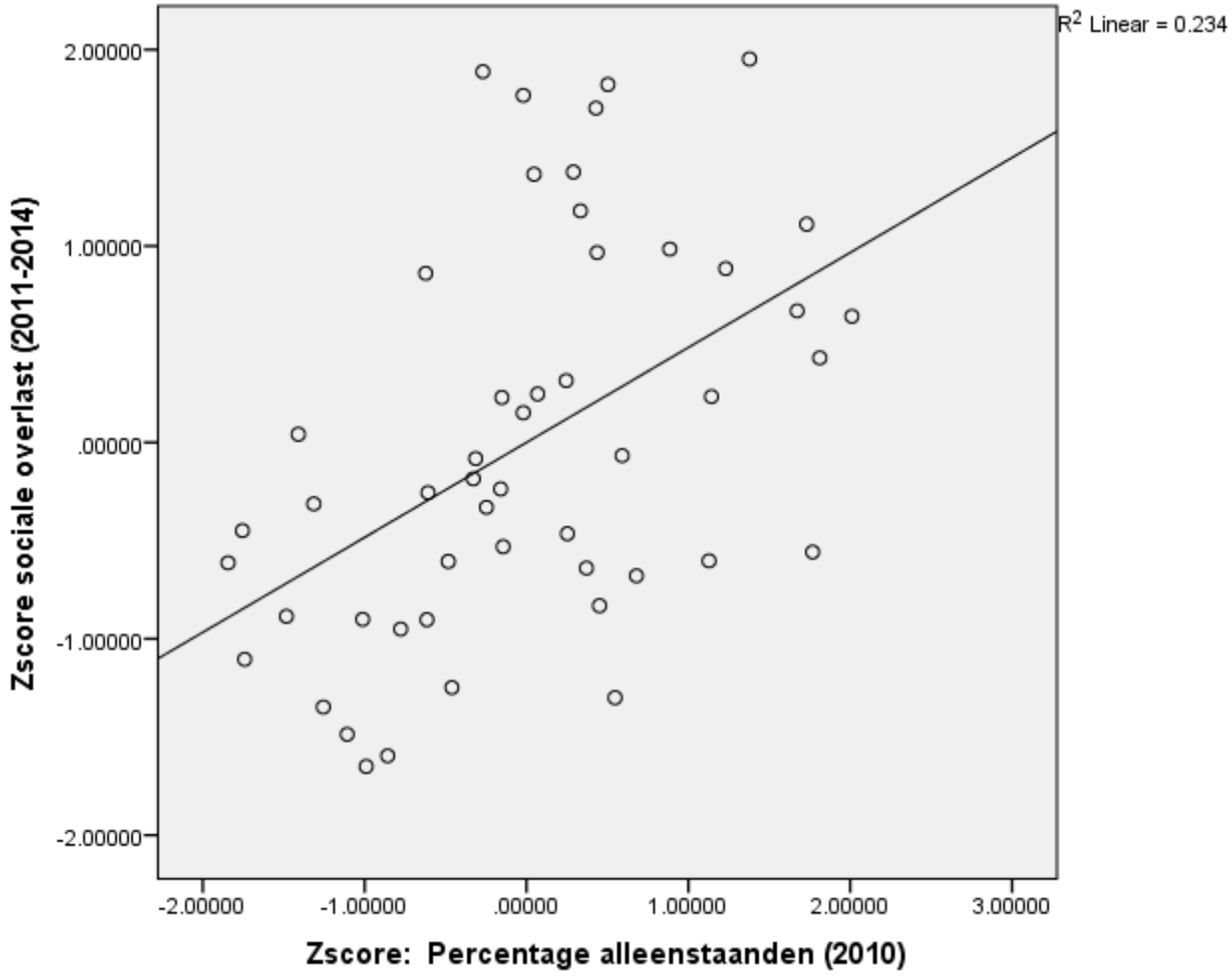
Het bepalen van de statistische samenhang tussen twee metrische kenmerken is een indicatie van het samen optreden van afwijkingen ten opzichte van het gemiddelde bij één van de variabelen met afwijkingen ten opzichte van het gemiddelde bij de andere variabele. De richting geeft aan of het over een *positief* of een *negatief* verband gaat. Bij een positief verband hangen hoge (lage) waarden van X (Y) samen met hoge (lage) waarden van Y (X). Bij een negatief verband zullen hoge (lage) waarden van X (Y) samengaan met lage (hoge) waarden op Y (X). De *sterkte* duidt op de mate waarin beide variabelen al dan niet samenhangen. Laten we de waarnemingen in het (x,y)-vlak beschouwen en een assenkruis door het bivariate zwaartepunt tekenen.



Kwadrant I: $(X - \overline{X})(Y - \overline{Y}) > 0$
Kwadrant II: $(X - \overline{X})(Y - \overline{Y}) < 0$
Kwadrant III: $(X - \overline{X})(Y - \overline{Y}) < 0$
Kwadrant IV: $(X - \overline{X})(Y - \overline{Y}) > 0$

Als het merendeel der (x,y)-waarnemingen in kwadrant II en IV liggen, dan vertonen deze punten een dalende tendens. Als het merendeel der (x,y)-waarnemingen in kwadrant I en III liggen, dan vertonen deze punten een stijgende tendens.

Voorbeeld: Puntenwolk van de relatie sociale overlast en het percentage alleenstaanden in Gentse buurten



Laten we een voorbeeld geven uit de praktijk van het geografisch criminologisch onderzoek. Als in een puntenwolk de punten een *lineair patroon* vormen dan is er sprake van een zekere *lineaire* samenhang tussen de variabelen. In het voorbeeld hierboven is er sprake van een positieve *samenhang*. De twee variabelen worden hier gepresenteerd onder de vorm van z-scores. Dit wil zeggen dat de variabelen gestandaardiseerd zijn en dat de gemiddelde score nul bedraagt en dat de standaardafwijking één bedraagt. Als je nu naar de Y-as kijkt, dan zie je de waarde van "nul". Idem voor de X-as. Als je deze beide punten met elkaar verbindt, dan heb je het denkbeeldige bivariate zwaartepunt. Uit deze puntenwolk leiden we een zeker patroon af: hoe hoger de score op de variabele "percentage alleenstaanden", hoe hoger de score voor de variabele "sociale overlast". Uit verschillende studies blijkt dat het percentage

alleenstaanden één van de sterkste predictoren voor criminaliteit en overlast lijkt te zijn. Vanuit de routine-activiteitentheorie van Marcus Felson werd gewezen op het feit dat alleenstaanden er een andere leefstijl op nahouden dan niet-alleenstaanden (althans gemiddeld genomen). Concreet komt de redenering hier op neer: in buurten waar veel alleenstaanden wonen, er minder informeel toezicht of "*guardianshi*p" is. Alleenstaanden gaan vaker uit dan samenwonenden en gezinnen met kinderen, waardoor de huizen en de straten aan minder toezicht worden blootgesteld. Hierdoor ontstaat een situatie waarin minder burgers bereid zijn om in te grijpen als er iets gebeurt. Dit is een interessante voedingsbodem voor criminaliteit, althans vanuit het opportuniteitsperspectief bekeken.

De hierboven geïllustreerde associatie kunnen we statistisch beschrijven aan de hand van drie belangrijke en onderling verworven associatiematen: de *covariatie, de covariantie* en *de correlatie*. Alvorens dieper in te gaan op de berekening van deze associatiematen, is het voor een goed begrip noodzakelijk te duiden op de mogelijkheden maar ook op de beperkingen ervan. Covariaties, covarianties en correlaties worden gebruikt om de samenhang tussen twee variabelen te schatten en het gaat hierbij om **symmetrische associatiematen**. Er is dus geen veronderstelling over causaliteit.

De covariatie

De covariatie wordt ook de **kruisproductensom** genoemd. De Engelstalige benaming hiervoor is de **Sum of Squares** (afgekort: **SSxy**) en ze stelt de mate voor waarin beide variabelen samen variëren (synoniem: co-variëren). Het is de **som van de *kruisproducten*.** Voor elke onderzoekseenheid kan je een kruisproduct berekenen. Eerder hebben we laten zien dat je voor elke onderzoekseenheid de afwijking tegenover het rekenkundig gemiddelde kan berekenen en dat vermenigvuldigen met zichzelf. Dat was eigenlijk een bijzonder geval van een kruisproduct, met name het kruisproduct met zichzelf. **Met kruisproduct bedoelen we het product** van een afwijking van een onderzoekseenheid tegenover de gemiddelde x-waarde, en de afwijking van een onderzoekseenheid tegenover de gemiddelde y-waarde. Als we deze oefening uitvoeren voor elk element in onze steekproef, dan krijgen we voor elke eenheid een nieuwe waarde. Als we die nieuwe waarden met elkaar optellen, dan krijgen we een nieuwe som. Dat is de kwadratensom en deze is heel belangrijk. Immers, op basis van de kruisproductensom worden de parameters van de bivariate associatie op metrisch niveau berekend.

$SS(x,y) = \{ (x_1 - \bar{x})\cdot(y_1 - \bar{y}) + ... + (x_n - \bar{x})\cdot(y_n - \bar{y}) \}$

$$SS_{xy} = \sum (X - \bar{X})(Y - \bar{Y})$$

Wanneer we de berekeningswijze van de covariatie bekijken, dan zien we dat de eerder geziene variatie eigenlijk een bijzonder geval is van de covariatie. De variatie is eigenlijk *de covariatie van een kenmerk met zichzelf*. In tegenstelling tot de variatie, waar de gesommeerde deviatiescores (van één variabele) gekwadrateerd worden, vermenigvuldigt men bij de covariatie de gesommeerde deviatiescores van de twee variabelen. Deze producten noemt men kruisproducten.

De covariantie

De covariatie is een maat die dezelfde nadelen heeft als de variatie. Omdat ze enkel gebaseerd is op de kruisproducten, krijgen we grote waarden. We moeten iets doen om deze maat te normeren. Een eerste belangrijke tussenstap is het berekenen van de covariantie. ***De covariantie (Sxy)*** van *x* en *y* is de ***kruisproductensom* van ($x_i - \bar{x}$) en ($y_i - \bar{y}$), gedeeld door n-1**.[10] Dus wordt de covariantie tussen x en y als volgt berekend:

$Cov(x,y) = \{ (x_1 - \bar{x})\cdot(y_1 - \bar{y}) + ... + (x_n - \bar{x})\cdot(y_n - \bar{y}) \} / n\text{-}1$

$$s_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1}$$

Gezien de covariantie afhankelijk is van de meeteenheid waarin de variabelen zijn opgenomen kan de absolute waarde van de covariantie weinig informatie bieden over de *sterkte* van de samenhang. Door te covariatie te delen hebben we het probleem van de normering nog niet opgelost. Vandaar dat ook hier nog steeds een vaste boven- en benedengrens ontbreekt, en dus ook het ontbreken van de mogelijkheid om waarden onderling te vergelijken. Wanneer bij wijze van voorbeeld inkomen wordt uitgedrukt *in jaarlijks inkomen* en leeftijd in jaren, bekomt men een grotere waarde van de covariantie dan wanneer

[10] Merk op dat we ook hier delen door n-1 en niet door n. De redenering is dezelfde als voorheen werd uiteengezet. **Wanneer onze gegevens berusten op steekproefresultaten wordt n-1 in de noemer gebruikt en wanneer we beschikken over gegevens afkomstig uit een voltallige populatie dan delen we door n.**

men werkt met *wekelijks inkomen*. Dat zou ook weer niet mogen. Een associatie tussen twee kenmerken mag niet afhankelijk zijn van de meting (uitgedrukt in weken dan wel in jaren). Een grotere waarde van de covariantie duidt dus niet op een sterkere samenhang maar is een rechtstreeks gevolg van het feit dat de numerieke waarde en de spreiding van inkomen groter is bij jaarlijks dan bij wekelijks inkomen. Aan de relatie tussen inkomen en leeftijd is echter niets veranderd. De oplossing is het standaardiseren van de covariantie. Deze gestandaardiseerde covariantie is gekend als de product-moment correlatiecoëfficiënt van Pearson.

De product-moment correlatiecoëfficiënt van Pearson

De **correlatiecoëfficiënt**, ook wel product-moment correlatiecoëfficiënt van Pearson genoemd, is gelijk aan de covariantie tussen X en Y in gestandaardiseerde vorm. De formule ziet er als volgt uit:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Om de correlatie te berekenen volstaat het dus de covariantie Cov(x,y) te standaardiseren door de covariantie te delen door Sx*Sy. Standaardisatie maakt de interpretatie van de associatie tussen twee metrische variabelen eenvoudiger. De correlatie varieert dankzij het proces van standaardisering van min één tot plus één; waarbij nul wijst op gebrek aan correlatie.

Hier volgen enkele vuistregels voor de interpretatie van de associaties:

- **0 – 0,10 zeer zwak/geen verband;**
- **0,11 – 0,30 zwak verband;**
- **0,31 – 0,50 redelijk verband;**
- **0,51 – 0,80 sterk verband;**
- **0,81 – 0,99 zeer sterk verband;**
- **1 perfect verband.**

Het gaat hier om absolute waarden, ongeacht het teken. Er dient verder op gewezen te worden dat men veelvuldig de fout maakt om op basis van een lage waarde van de correlatie tussen twee variabelen te besluiten dat er geen verband zou bestaan tussen beide kenmerken. Men dient zich steeds te realiseren dat een correlatie toetst naar een *lineair* verband tussen twee variabelen. Het verdient aanbeveling om een puntenwolk (of scatterplot) van beide variabelen

te bestuderen aangezien het mogelijk is dat de variabelen wel een samenhang vertonen maar dat dit verband niet lineair is.

2. Covariatie, covariantie en correlatie: een uitgewerkt rekenvoorbeeld

Een heleboel complexe multivariate analysetechnieken die criminoloog-onderzoekers gebruiken zijn gebaseerd op deze bivariate associatiematen. Precies omdat deze associatiematen zo belangrijk zijn in de statistiek, moeten studenten deze zeer goed kunnen interpreteren, maar ook beseffen hoe deze maten uitgerekend worden. We geven een uitgewerkt rekenvoorbeeld. We berekenen de associatie tussen twee testscores die studenten hebben op twee psychosociale proeven ter voorbereiding van een sollicitatiegesprek voor een job als strategisch analist bij de federale politie. We zetten even de stappen op een rijtje die je moet maken om de covariantie zelf uit te rekenen. In de praktijk maken criminologen echter gebruik van statistische verwerkingspakketten om deze berekening uit te voeren. Toch is het noodzakelijk dat je snapt wat er in werkelijkheid achter de schermen gebeurt wanneer je de samenhang tussen kenmerken berekent aan de hand van software pakketten. Het begrijpen van wat een statistische analyse doet, zal ervoor zorgen dat de kans dat je de resultaten van een analyse verkeerd interpreteert, aanzienlijk verkleint.

Tabel: tussenstappen bij het berekenen van een correlatie

Student	ScoreT1	ScoreT2	$x_1 - \bar{x}$	$y_1 - \bar{y}$	$(x_1 - \bar{x})*(x_1 - \bar{x})$	$(y_1 - \bar{y})*(y_1 - \bar{y})$	$(x_1 - \bar{x})*(y_1 - \bar{y})$
An	30,00	65,00	0	2	0	4	0
Arno	45,00	75,00	15	12	225	144	180
Bart	35,00	60,00	5	-3	25	9	-15
Björn	20,00	50,00	-10	-13	100	169	130
Delphine	40,00	80,00	10	17	100	289	170
Hanne	35,00	75,00	5	12	25	144	60
Henk	30,00	70,00	0	7	0	49	0
Ines	30,00	75,00	0	12	0	144	0
Jeroen	25,00	55,00	-5	-8	25	64	40
Jurgen	20,00	40,00	-10	-23	100	529	230
Kim	40,00	75,00	10	12	100	144	120
Robert	25,00	60,00	-5	-3	25	9	15
Nele	20,00	60,00	-10	-3	100	9	30
Sara	25,00	50,00	-5	-13	25	169	65
Sofie	30,00	55,00	0	-8	0	64	0
N= 15 $\bar{x}$ = 30 $\bar{y}$ = 63					Sum of squares (Variatie) 850	Sum of squares (Variatie) 1940	Covariatie of Kruisproducten som 1025

Stappen te volgen in het uitrekenen van een correlatie:

Hieronder hebben we de stappen uitgeschreven. We brengen in herinnering dat we gebruik maken van dezelfde gegevens, met name de testscores van studenten op twee examens. Echter, in plaats van te kijken naar de individuele variatie in de testscores, gaan we nu kijken naar de mate waarin deze twee testscores samenhangen.

1. Bereken het ***rekenkundig gemiddelde*** van de twee variabelen, zoals eerder werd uiteengezet.
2. Bereken de ***afwijkingen*** van elke onderzoekseenheid ten opzichte van het rekenkundig gemiddelde voor de beide variabelen X en Y. Anders gesteld: bereken de deviatiescores voor elke onderzoekseenheid op basis van de variabelen X en Y.
3. ***Kwadrateer de deviatiescore op basis van X en op basis van Y***. Op die manier leg je de basis voor de berekening van de variatie in X en Y en de covariatie tussen X en Y.
4. Neem *de* ***kwadratensom*** *van de deviatiescores op basis van X en op basis van Y.*
5. Bereken de ***variantie*** **van X** en de ***variantie*** **van Y**. Dit gebeurt door zowel variatie van X als de variatie van Y te delen door N-1.
6. Bereken de vierkantswortel van de variantie in X en de vierkantswortel van de variantie in Y. Je hebt nu ook de ***standaardafwijking*** van X en van Y.
7. Bereken de **kruisproductensom** en je hebt de ***covariatie tussen X en Y***.
8. **Deel de kruisproductensom door n-**1 en je hebt de ***covariantie*** tussen X en Y.
9. ***Vermenigvuldig*** de standaardafwijking van X met de standaardafwijking van Y.
10. Deel de covariantie tussen X en Y door de vermenigvuldiging van de standaardafwijking van X met de standaardafwijking van Y. Dit resultaat is de ***correlatiecoëfficiënt van Pearson***.

Als je berekening juist is stel je vast dat de correlatie 0.79 bedraagt. Een heranalyse met het statistisch verwerkingspakket SPSS leert ons hetzelfde. De uitkomst van deze heranalyse ziet er zo uit:

Descriptive Statistics

	Mean	Std. Deviation	N
Score T1	30,0000	7,79194	15
Score T2	63,0000	11,77164	15

Correlations

		ScoreT1	ScoreT2
Score T1	Pearson Correlation	1	,798(**)
	Sum of Squares and Cross-products	850,000	1025,000
	Covariance	60,714	73,214
	N	15	15
ScoreT2	Pearson Correlation	,798(**)	1
	Sum of Squares and Cross-products	1025,000	1940,000
	Covariance	73,214	138,571
	N	15	15

** Correlation is significant at the 0.01 level (2-tailed).

3. De bivariate lineaire regressieanalyse als asymmetrische analysetechniek

Ook wetenschappers en beleidsmakers houden van voorspellingen dat iemand een misdrijf gaat plegen. Zou het mogelijk zijn om te interveniëren vlak voor het misdrijf plaatsvindt? Neen, we leven nog steeds niet in zo een "brave new world". Denk bijvoorbeeld aan de film "Minority Report". Maar toch is voorspellen nuttig. Rechters willen graag rekening houden met recidivestudies en criminele carrièrestudies wanneer zij de moeilijke beslissingen moeten nemen wat te doen met iemand die voor de rechter verschijnt. Een jonge student verschijnt voor de eerste keer voor een rechter. Deze jongeman heeft te snel gereden. Deze jongeman heeft geen criminele antecedenten en goede schoolresultaten en bovendien alles om het te maken. Het kan zijn dat deze jongeman eventjes is uitgegleden en een fout gemaakt heeft. Dat gebeurt. Een andere jongeman heeft een ernstig coke-probleem en heeft onder invloed gereden. De kans dat die laatste nog eens onder invloed achter het stuur kruipt is mogelijks groter dan de kans dat de andere jongeman met het onberispelijke verleden onder invloed achter het stuur kruipt. Zou het mogelijk zijn te voorspellen op basis van een reeks achtergrondkenmerken wat er gebeurt? Stel je nu eens voor dat je in het kader van je masterproef geïnteresseerd bent in criminele recidive. Je wil wel eens weten of je kan voorspellen hoeveel keer iemand in de gevangenis zal belanden in de volwassenheid op basis van de scores op een IQ-test die werd afgenomen in de kindertijd. Een eerste vraag die bij je opkomt is wellicht: kan dat wel? Immers, je verzamelt informatie over individuen in een steekproef als ze pakweg allemaal tien jaar zijn en na dertig jaar ga je deze

steekproefpersonen terug opzoeken en je stelt hen bijkomende vragen en vraagt hun criminele gegevens op. Er bestaan studies die dergelijke gegevens hebben bijgehouden en er zijn criminologen die dat onderzocht hebben. Eén van die historische figuren was professor David Farrington. Als je voorspellingen wil maken heb je meestal wel wat criminologische theorie als gods. Denk aan wat in het inleidende hoofdstuk werd gezegd. We gaan gemakshalve uit van de idee dat er een verband bestaat tussen IQ en criminele recidive. Waarom zou zo een verband er moeten zijn? Wel, het is mogelijk dat individuen met een laag IQ minder goed de gevolgen van hun gedrag inschatten, of minder goed kunnen inschatten of ze wel eens door de politie zouden opgepakt kunnen worden. Dus: onze hypothese is zeker plausibel. Ze is zelfs onderzocht door de criminologen James Q. Wilson en Richard Herrnstein in de jaren 1980. We verzamelen gegevens over de criminele recidive, die we onze afhankelijke variabele noemen. Deze afhankelijke variabele noemen we ook wel het explanandum. We willen weten of we de scores op de afhankelijke variabele kunnen verklaren op basis van een onafhankelijke variabele. De onafhankelijke variabele is het explanans. We gaan een voorspelling maken en we doen dat door een lineaire regressieanalyse uit te voeren.

De volledige naam van deze analysetechniek is de *OLS-lineaire regressieanalyse*. OLS staat voor *Ordinary Least Squares* en is het principe dat gebruikt wordt om de regressieanalyse mathematisch uit te voeren. Het begrip OLS leggen we verderop uit. Hou dat begrip nog eventjes in het achterhoofd. De regressieanalyse is eerst en vooral een **asymmetrische** associatiemaat. We begeven ons nu op het terrein van de verklarende statistiek, die gebruikt wordt ter toetsing van criminologische theorieën en die gebruikt wordt in het predictie-onderzoek. Het doel van een bivariate lineaire regressieanalyse bestaat uit het statistisch verklaren van de variatie in een **afhankelijke variabele** (de responsvariabele of het explanandum genoemd) op basis van een **onafhankelijke variabele** (predictor-variabele of explanans). Er wordt in het theorietoetsend onderzoek een causaal verband tussen beide variabelen verondersteld waarmee men de afhankelijke variabele tracht te verklaren. In het predictie-onderzoek wordt dat causaal verband niet altijd verondersteld, men wil vooral weten of een variabele een voorspellend karakter heeft, want men wil bij de berechtiging rekening houden met een reeks van kenmerken van de persoon in kwestie.

Het uitvoeren van een enkelvoudige lineaire regressieanalyse levert een statistische vergelijking op waarmee de afhankelijke variabele voorspeld kan worden op basis van de

onafhankelijke variabele. Bij enkelvoudige lineaire regressie kan deze vergelijking op twee manieren genoteerd worden:

$Y = \beta_0 + \beta_1 X + e$ *of* *Aantal veroordelingen* $= \beta_0 + \beta_{1\,IQ} + e$

of nog

$Y = a + b_1 X + e$

In deze vergelijking is **Y de geobserveerde afhankelijke variabele**. X is de onafhankelijke variabele. De β's zijn de (populatie)parameters die met de regressieanalyse worden geschat en worden daarom ook wel de *regressiecoëfficiënten* genoemd.

(1) β_0 is de *(regressie)constante of het intercept*. Deze wordt ook met het symbool a aangeduid. Dat zagen we in de tweede notatie. **Deze constante drukt de verwachte of voorspelde waarde van Y uit wanneer X nul bedraagt.**[11] Het intercept is niet steeds betekenisvol. Onderzoekers zijn veel meer geïnteresseerd in de effecten van de onafhankelijke variabele.

(2) β_1 (in de tweede notatie $b_{1)}$ is het ongestandaardiseerde *regressiegewicht* dat de helling van de regressielijn aanduidt. **β_1 geeft aan met hoeveel eenheden Y toeneemt als X met één eenheid toeneemt.** Als het IQ van een persoon gemeten is aan de hand van een gekende IQ test (zoals bijvoorbeeld de Stanford binet test) dan is de richtingscoëfficient gelijk aan de toename in Y als het IQ met een eenheid toeneemt. Dat zegt wellicht ook niet veel. Daarom kunnen we een onafhankelijke variabele ook standaardiseren vooraleer we de analyse uitvoeren. Dan krijgt de richtingscoëfficiënt wel betekenis. Een standaardafwijking heeft betekenis. We weten dat IQ normaal verdeeld is en we kunnen ons dus wel iets voorstellen bij een toename van een standaardafwijking. De waarde van het regressiegewicht kan zowel positief als negatief zijn. Bij een positieve waarde treedt er een stijging op van Y per eenheidstoename van X. Bij een negatieve waarde treedt er een

[11] Wil men dat deze waarde wel informatief wordt, kan men de onafhankelijke variabele centreren rond het gemiddelde. Op die manier staat nul voor een gemiddelde waarde, en krijgt het intercept de betekenis van de score op Y voor iemand met een gemiddelde score op X.

daling op van Y per eenheidstoename van X. Hoe groter de waarde van β_1 des te groter de verandering van Y bij een verandering van X. Door te werken met een gestandaardiseerde onafhankelijke variabele krijgt ook het intercept betekenis, want dit is de waarde voor Y als X nul is. We weten dat nul wijst op de gemiddelde score, als X gestandaardiseerd is.

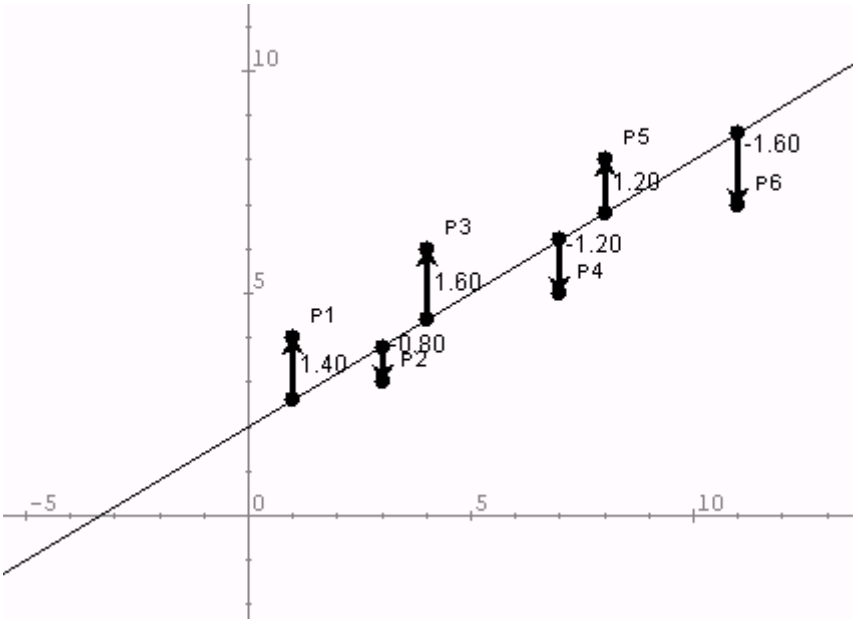
(3) De *e* in de vergelijking staat voor de **foutenterm** ('error'). Dit is het **verschil tussen de werkelijke (geobserveerde) waarde van Y en de door het statistische regressiemodel voorspelde waarde van Y**. Geen enkele voorspelling is immers perfect in de sociale wetenschappen. Er is in de praktijk dus altijd een verschil tussen een observatie en een predictie. Voor sommige individuen zullen we heel goede voorspellingen kunnen maken, voor anderen heel slechte voorspellingen. Dit verschil tussen de predictie en de geobserveerde waarde wordt ook wel het *residu of de residuele term* genoemd. De som van alle residuen bedraagt nul. De variantie van *e* is gelijk ongeacht de waarde van X. Deze foutenterm is niet gecorreleerd met X.

Op onderstaande grafiek zijn de punten $P_1(x_1,y_1)$ tot en met $P_6(x_6,y_6)$ van een puntenwolk getekend. De rechte $y = a + bx$ werd ook getekend en de punten $(x_i, \hat{y}_i)$ op de rechte die telkens corresponderen met de punten P_i.

Opmerking: de voorspelde waarde op basis van X wordt aangeduid met de notatie $\hat{y}_i$

We spreken $\hat{y}_i$ ***uit (Engels:" Y-hat") als de verwachte / voorspelde waarde van Y op basis van X.***

Figuur: De regressierechte van y gegeven x



Bekijk de relatie tussen x en y in de figuur hierboven. Indien men veronderstelt dat y afhangt van x, bepaalt men de regressierechte van y gegeven x. Deze lijn vat de lineaire relatie tussen beide variabelen zo goed mogelijk samen. Een algemene voorstelling van een lineair verband tussen twee variabelen wordt gegeven door de formule. Men beschouwt een theoretisch model in hetwelk de veranderlijke y kan beschouwd worden als een lineaire functie van x

$Y = \beta_0 + \beta_1 X + e$

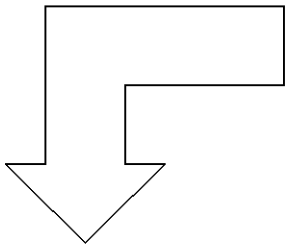
Voor elke waarneming x_i heeft men twee overeenkomende waarden van y: deze zijn: (1) de geobserveerde waarneming y_i en (2) de voorspelde waarde op basis van de onafhankelijke variabele, waarbij men uitgaat van lineariteit.

Het is ontzettend belangrijk de redenering en de betekenis van de parameters te kennen, zodat men geen foutieve interpretatie maakt van de resultaten van een regressieanalyse. Men dient de geschatte regressieparameters goed te kunnen interpreteren en verwoorden in termen van inhoudelijke betekenis.

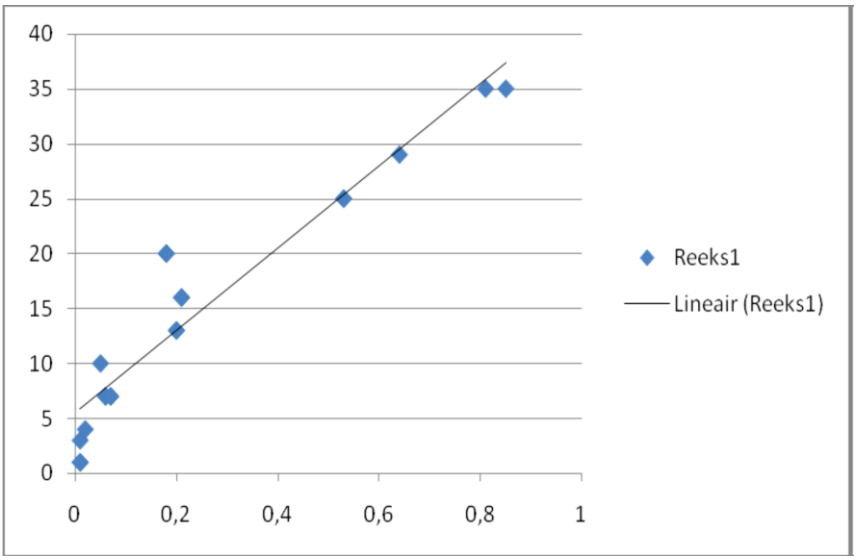
Als de geobserveerde waarden van de afhankelijke variabele in een puntenwolk geplaatst worden tegenover de geobserveerde waarden van de onafhankelijke variabele verkrijgt men een visueel beeld van (x,y)-coördinaten. Nu tracht men via een regressievergelijking het

verband tussen de onafhankelijke en afhankelijke variabele voor te stellen door middel van een rechte zodanig dat de globale voorspellingsfout die men maakt zo klein mogelijk is. We trachten de analysetechniek zo eenvoudig mogelijk voor te stellen. Hieronder zien we voor tien statistische eenheden de observaties voor twee criminologisch relevante kenmerken, met name X en Y. Uit de ruwe observaties is het niet duidelijk of deze kenmerken samenhangen. We maken een puntenwolk en we zien een sterke samenhang. Het lijkt er sterk op dat we een voorspelling kunnen maken van Y op basis van X. Dit zien we alleen al uit de puntenwolk. Maar hoe goed zijn deze voorspellingen? En hoe drukken we deze uit? De bivariate regressieanalyse is de techniek bij uitstek die je deze onderzoeksvraag leert te beantwoorden

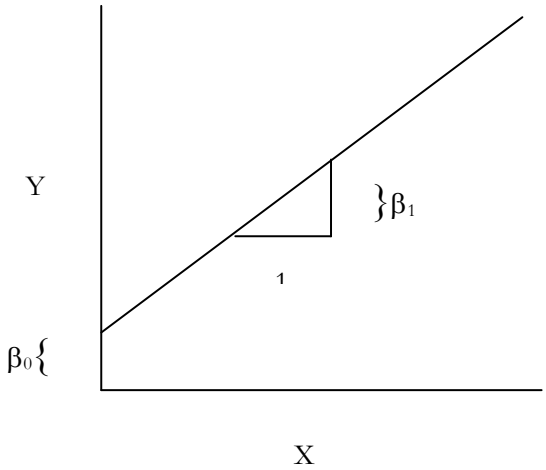
Figuur: X-Waarden en Y-waarden voor tien statistische eenheden



Eenheid	X	Y
1	0,64	29
2	0,21	16
3	0,85	35
4	0,53	25
5	0,02	4
6	0,01	1
7	0,21	16
8	0,18	20
9	0,06	7
10	0,20	13



Hoe bepalen we het intercept en de hellingshoek (richtingscoëfficiënt)?



Het bepalen van de regressiecoëfficiënten gebeurt via hetgeen men noemt "de gewone kleinste kwadratenoplossing". Om een zo goed mogelijk model te bekomen, worden de parameters a en b zodanig bepaald *dat de gekwadrateerde afwijking van de regressierechte minimaal wordt.* Anders gezegd: de gewone kleinste kwadraten methode (Engels: ***Ordinary***

Least Squares of OLS) levert een zodanige formule voor de regressielijn dat de gekwadrateerde afstand van alle datapunten tot die lijn (Engels: **residual sum of squares**) minimaal is. Bijgevolg is de kwadratische afwijking van de punten t.o.v. de regressierechte een kwaliteitsmaat voor de gevonden rechte. Men noemt deze de *residuele variatie*. De regressielijn gaat per definitie door het zwaartepunt van de puntenwolk. **Het minimaliseren van de residuen (verschillen tussen observaties en voorspellingen) is wat er gebeurt in deze asymmetrische analysetechniek.**

We stellen nu de regressievergelijking op en hanteren het principe van de kleinste kwadratenoplossing om de onderzoeksvraag te beantwoorden:

$\hat{y}_i = a + b_1 X_i$

In deze vergelijking zijn a en b_1 de steekproefparameters; a is *de regressieconstante* en b_1 is het *regressiegewicht*. $\hat{y}_i$ is dus de y-waarde die voorspeld wordt voor subject i op basis van de waarde van x voor subject i. Nu is het dus de bedoeling a en b_1 zo te kiezen dat de som van de gekwadrateerde afwijkingen tussen de geobserveerde waarden en de verwachte waarden zo klein mogelijk is. Er kan worden aangetoond dat de optimale waarden van b_0 en b_1 de volgende zijn:

$a = \overline{Y} - b_1 \overline{X}$

$b_1 = r(X,Y)\, S_Y/S_x$

waarbij:

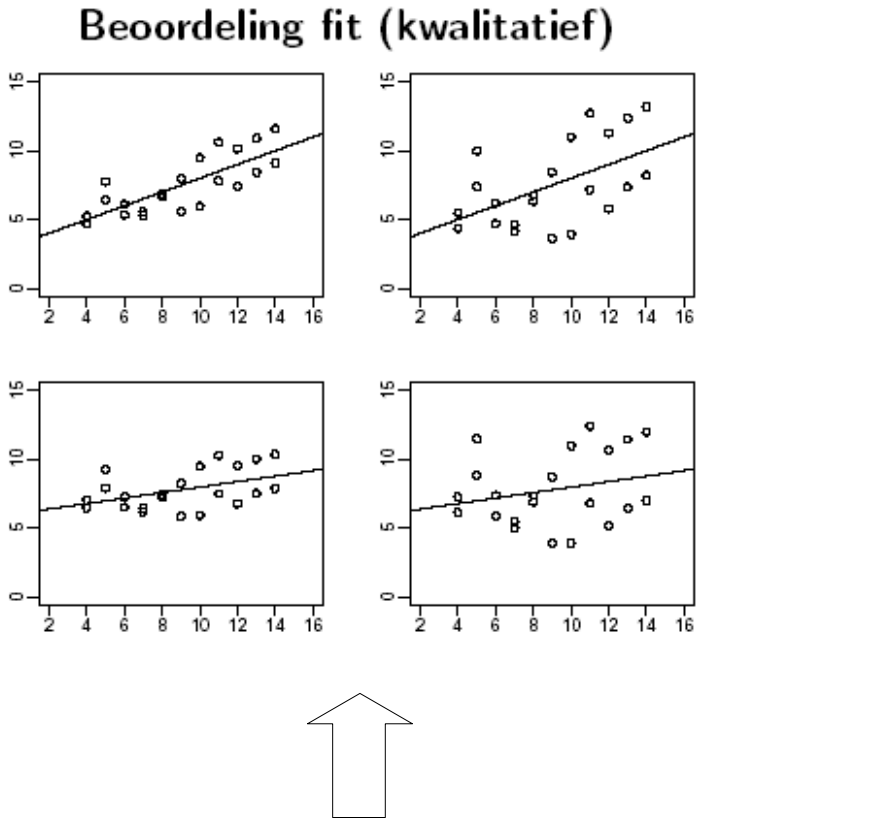
- **$\overline{Y}$ en $\overline{X}$ de (steekproef)gemiddelden zijn van Y en X,**
- **S_y en S_x de steekproefstandaarddeviaties**
- **en r(X,Y) de steekproefcorrelatie tussen X en Y.**

De interpretatie van deze parameters kan gebeuren zoals hierboven werd beschreven bij de modelformulering. Bovenstaande vergelijking wordt ook wel een **statistisch model** genoemd. Een belangrijke vraag die de criminoloog dient te beantwoorden is de volgende: **hoe goed past het statistische model bij de geobserveerde data? We noemen het beantwoorden van deze vraag ook wel de evaluatie van de "model fit (Engels: to fit= passen)". De**

model fit zegt dus iets over hoe goed onze predicties het doen tegenover onze observaties. Of anders: past het statistische model goed bij de data? Als het er goed bij past, vat het de relatie tussen X en Y heel goed samen.

We kunnen dit "tentatief" doen en een eerste blik werpen op regressiemodellen. Hieronder presenteren we een aantal situaties die zich in de praktijk van de statistische analyse kunnen voordoen. Let op het volgende: de voorbeelden werden zo geselecteerd dat je het verschil goed merkt. Er zijn situaties waarin de observaties of punten heel dicht bij de best passende lijn liggen en er zijn situaties waar de punten iets verder van de lijn liggen. Deze twee situaties zijn denkbeeldig voor elke rechte (ongeacht hoe steil deze is). Door een puntenwolk te tekenen en de best passende rechte op de puntenwolk te tekenen, zie je hoe dicht de observaties bij de lijn liggen en daardoor zie je al op het eerste zicht of de voorspelling goed dan wel zwak is. Hoe dichter bij de lijn, hoe beter.

Figuur: model fit beoordeling op kwalitatieve wijze



- Punten dicht bij de lijn: betere fit
- Lijn heeft grotere helling (t.o.v. range van Y): betere fit
- In welke plaatjes dus een goede fit?

Het mag duidelijk zijn dat we met het blote oog niet in staat zijn *precieze uitspraken* te doen over de model fit. Daarom doen we beroep op een reeks coëfficiënten die de kwaliteit van de regressielijn helder uitdrukken. Eenmaal de regressiecoëfficiënten gekend zijn, dient nog bepaald te worden of en hoe goed men de waarden van de afhankelijke variabele (Y) kan voorspellen op basis van de waarden van de onafhankelijke variabele (X). We willen weten hoeveel van de variatie[12] in Y men kan "verklaren" op basis van de variatie in X.

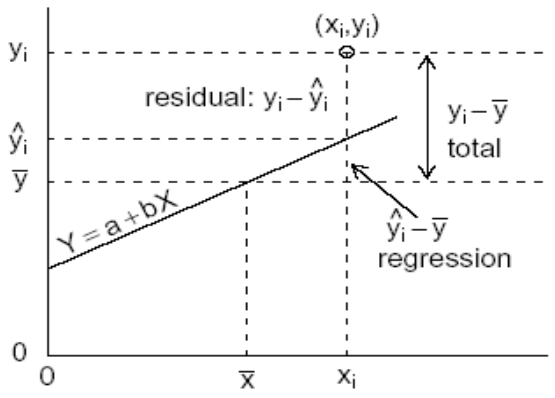
[12] Inhoudelijk: de verschillen in Y-waarden, maar ook statistisch: de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde in Y vormen de statistische uitdrukking voor variatie.

De afwijkingen van het gemiddelde van Y bestaan uit twee componenten: ***toevallige afwijkingen (residual)*** **en** ***verklaarde afwijkingen***: afwijkingen van het rekenkundig gemiddelde verklaard door de regressielijn.

De totale variatie van Y kan opgesplitst worden in twee delen.

- **Een eerste gedeelte** vertegenwoordigt de variantie in Y die 'verklaard' wordt door X. We noemen dit gedeelte van de variatie in Y de **"regression sum of squares"**. Dit gedeelte wordt steeds groter naarmate men er beter in slaagt Y-waarden te voorspellen die verder afwijken van de voorspelling die men zou maken als men geen informatie heeft. Als we geen informatie hebben over een onafhankelijke variabele, dan is onze beste voorspelling nog steeds het rekenkundig gemiddelde $(\overline{Y})$[13].
- **Een tweede gedeelte** vertegenwoordigt de *foutenvariatie*, dit wil zeggen de mate waarin de voorspelde waarden van Y afwijken van de geobserveerde waarden van Y. We noemen deze ook de **"residual sum of squares"**.
- De **totale variatie in Y** noemen we de *Total sum of squares* (zie univariate statistiek) en is dus mathematisch niets anders dan **de som van de *Regression sum of squares* en de *residual sum of square*s**. In formule en getekend krijgen we het volgende resultaat:

Figuur: de regressierechte in technische termen ontleed



Naam	formule
Regression	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
Residual	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$

[13] Indien men over geen enkele informatie beschikt om de voorspelling op te baseren, maakt men de kleinste globale voorspellingsfout door het gemiddelde van de variabele als voorspelde waarde te gebruiken.

Een goede maat voor de kwaliteit van de regressievoorspelling is de verhouding van de verklaarde variatie t.o.v. de totale variatie. Deze wordt de *determinatiecoëfficiënt* (R^2) genoemd. De coëfficiënt $\mathbf{R}^2$ ("r-kwadraat") is de proportionele reductie in de voorspellingsfout die ontstaat door op de regressielijn te steunen bij de voorspelling van de afhankelijke variabele en is dus **de proportie van de totale variatie in Y die door X wordt verklaard**. Deze coëfficiënt kan waarden aannemen gaande van 0 tot 1. **In een bivariate regressieanalyse is de determinatiecoëfficiënt gelijk aan het kwadraat van de correlatiecoëfficiënt tussen X en Y.**

Laten we de teller en noemer in de formule van de determinatiecoëfficiënt ontleden:

$$\mathbf{R}^2 = \frac{\sum_{i=1}^{n}\left[\hat{y}_i - \overline{Y}\right]^2}{\sum_{i=1}^{n}\left[y_i - \overline{Y}\right]^2}$$

Teller: Regression sum of squares

Noemer: Total sum of squares

De **teller** bevat variatie die bestaat uit de som van het gekwadrateerde verschil tussen de voorspelde waarde van Y en de gemiddelde waarde van Y. De gemiddelde waarde is immers onze beste voorspeller als we geen onafhankelijke variabele hebben.
De **noemer** bestaat uit de som van het gekwadrateerde verschil tussen de geobserveerde waarde van Y en het gemiddelde van Y.

Omdat er wordt gesteund op de regressielijn voor de voorspelling van Y, is $\mathbf{R}^2$ in zuiver technisch opzicht een symmetrische maat. Het is wel mogelijk om een **asymmetrische interpretatie** aan $\mathbf{R}^2$ geven.

4. Zelf uitrekenen van de parameters van de regressierechte

We werken een rekenvoorbeeld uit in de hieronder gepresenteerde tabel. We baseren ons op het voorbeeld dat we eerder gaven bij de berekening van de metrische associatiematen de covariatie, covariantie en correlatie. Het is duidelijk dat je al deze maten dient te kennen, wil je een regressieanalyse verstaan.

Student	**ScoreT1**	**ScoreT2**	$x_1 - \bar{x}$	$y_1 - \bar{y}$	$(x_1 - \bar{x})^* (x_1 - \bar{x})$	$(y_1 - \bar{y})^* (y_1 - \bar{y})$	$(x_1 - \bar{x})^* (y_1 - \bar{y})$	Predictie Y op basis van X
An	30,00	65,00	**0**	**2**	0	4	0	63
Arno	45,00	75,00	**15**	**12**	225	144	180	81
Bart	35,00	60,00	**5**	**-3**	25	9	-15	69
Björn	20,00	50,00	**-10**	**-13**	100	169	130	51
Delphine	40,00	80,00	**10**	**17**	100	289	170	75
Hanne	35,00	75,00	**5**	**12**	25	144	60	69
Henk	30,00	70,00	**0**	**7**	0	49	0	63
Ines	30,00	75,00	**0**	**12**	0	144	0	63
Jeroen	25,00	55,00	**-5**	**-8**	25	64	40	57
Jurgen	20,00	40,00	**-10**	**-23**	100	529	230	51
Kim	40,00	75,00	**10**	**12**	100	144	120	75
Robert	25,00	60,00	**-5**	**-3**	25	9	15	57
Nele	20,00	60,00	**-10**	**-3**	100	9	30	51
Sara	25,00	50,00	**-5**	**-13**	25	169	65	57
Sofie	30,00	55,00	**0**	**-8**	0	64	0	63
N= 15 $\bar{x}$ **= 30** $\bar{y}$ **= 63**					**Sum of squares 850**	**Sum of squares 1940**	**Covariatie 1025**	

Stappen te volgen in het uitrekenen van een bivariate regressie:

1. Bereken het *rekenkundig gemiddelde* van de twee variabelen, zoals eerder werd uiteengezet.
2. Bereken de *afwijkingen* van elke eenheid ten opzichte van het rekenkundig gemiddelde voor de beide variabelen X en Y.
3. *Kwadrateer de afwijkingen* van elke eenheid met het rekenkundig gemiddelde. Op die manier leg je de basis voor de berekening van de variatie in X en Y en covariatie tussen X en Y.
4. Neem de *som van de gekwadrateerde afwijkingen* tegenover de gemiddelde waarde van X en neem de som van de gekwadrateerde afwijkingen tegenover de gemiddelde waarde van Y.
5. Bereken de *variantie* van X en de variantie van Y. Dit gebeurt door zowel variatie van X als de variatie van Y te delen door n-1.
6. Bereken de vierkantswortel van de variantie in X en de vierkantswortel van de variantie in Y. Je hebt nu ook de *standaardafwijking* van X en van Y.

7. Bereken de kruisproductensom en je hebt de *covariatie*.
8. Deel de kruisproductensom door n-1 en je hebt de *covariantie* tussen X en Y.
9. *Vermenigvuldig* de standaardafwijking van X met de standaardafwijking van Y.
10. Deel de covariantie tussen X en Y door de vermenigvuldiging van de standaardafwijking van X met de standaardafwijking van Y. Nu heb je de *correlatiecoëfficiënt.* Deze is *gelijk aan de gestandaardiseerde richtingscoëfficiënt.*
11. Bereken de *ongestandaardiseerde richtingscoëfficiënt*. Deze is gelijk aan de covariantie tussen X en Y gedeeld door de variantie in X.
12. Het *intercept* (a) kan worden berekend op basis van voorgaande info:

    $a = \bar{y} - B\ \bar{x} = 26.824$

13. Op basis van de regressievergelijking Y(hat) = a + b.X, kan nu de predictie van Y berekend op worden op basis van X (bijvoorbeeld als X= 30: 26.824 + 1.206. 30 = 63).
14. Bereken de *determinatiecoëfficiënt*. Deze is gelijk aan het kwadraat van de bivariate correlatiecoëfficiënt. Let wel: dit geldt enkel in het geval van bivariate regressieanalyse.
15. Bereken de *aliënatiecoëfficiënt.* Deze kan gemakkelijk berekend worden: 1 – determinatiecoëfficiënt. Dit is **de proportie van de totale variatie in Y die NIET door X kan verklaard worden.**

De uitkomsten van deze regressieanalyse werden met SPSS uitgevoerd en kunnen hieronder geraadpleegd worden. Als deze uitkomsten ietwat verschillen met de door jullie uitgerekende uitkomsten, bedenk dan dat SPSS niet afrondt. Jullie worden wel verwacht af te ronden.

Beschrijvende stats voor de afhankelijke en onafhankelijke variabele

	N	Minimum	Maximum	Mean	Std. Deviation
Test_1 (X)	15	20.00	45.00	30.0000	7.79194
Test_2 (Y)	15	40.00	80.00	63.0000	11.77164
Valid N (listwise)	15				

Correlaties en kruisproductensom

		Test_1	Test_2
Test_1	**Pearson Correlation**	1	.798**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	850.000	1025.000
	Covariance	60.714	73.214
	N	15	15
Test_2	**Pearson Correlation**	.798**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	1025.000	1940.000
	Covariance	73.214	138.571
	N	15	15

**. Correlation is significant at the 0.01 level (2-tailed).

Coefficients[a]

Model		**Unstandardized Coefficients**		**Standardized Coefficients**	t	Sig.
		B	**Std. Error**	**Beta**		
1	(Constant)	**26.824**	7.807		3.436	.004
	Test_1	**1.206**	.252	**.798**	4.778	.000

a. Dependent Variable: Test_2 de rico bedraagt 1.206 en deze is gelijk aan 0.798 (11.77164/7.79194) oftewel de correlatiecoëfficiënt maal de standaardafwijking van Y gedeeld door de standaardafwijking van X.

Model Summary

Model	R	**R Square**	Adjusted R Square	Std. Error of the Estimate
1	.798[a]	**.637**	.609	7.35878

a. Predictors: (Constant), Test_1

De determinatiecoëfficiënt is het kwadraat van de correlatiecoëfficiënt en bedraagt zoals je uit de output kan afleiden 0.637. Dit wil zeggen dat 63.7% van de variabiliteit in T2 kan verklaard worden op basis van T1.

5. De rapportage van de belangrijkste parameters van de regressierechte in een rapport

Een belangrijke vraag waar studenten mee worstelen, is de vraag naar de rapportage van parameters uit een regressierechte in een rapport, bachelor- of masterproef. We geven een voorbeeld uit eigen verzamelde onderzoeksgegevens over criminaliteit en werkloosheid in 210 Antwerpse buurten.

Afhankelijke variabele: criminaliteitsgraad opzettelijke slagen	**Ongestandaardiseerde Coëfficiënt B**	**Gestandaardiseerde coëfficiënt β**
Intercept (B0 of a)	-0.045	**--**
Onafhankelijke variabele (B1): werkloosheidsgraad	**1.068**	**0.682**

Determinatiecoëfficiënt: 0.464 (Noot: vetgedrukte coëfficiënten zijn statistisch significant- zie verder)

De criminaliteitsgraad is de uitkomstvariabele en de werkloosheidsgraad is de onafhankelijke variabele (ook wel predictor-variabele of voorspeller genoemd in de literatuur). De voorspelde waarde voor de criminaliteitsgraad van geweld = -0.045 + 1.068 (werkloosheidsgraad).

Op basis van de werkloosheidsgraad zijn we in staat om 46.4 procent van de variabiliteit in de criminaliteitsgraad voor opzettelijke slagen te voorspellen. 46.4 procent van de geobserveerde verschillen in de criminaliteitsgraad van Vlaamse gemeenten kan worden verklaard vanuit de werkloosheidsgraad van de gemeente. Dit is een relatief goede voorspelling, want ongeveer de helft van de variatie in de criminaliteitsgraad kan gebeuren op basis van slechts één kenmerk, de werkloosheid. Deze bevinding mag niet op zichzelf staan, maar moet de onderzoeker-criminoloog aanzetten tot het bedenken van een verklaring. Waarom is dit zo? Welk mechanisme gaat hierachter schuil? Gemeenten zijn geen causale actoren, met andere woorden, we moeten op zoek gaan naar het causale mechanisme op een lager niveau. We moeten “de black box” opendoen en kijken welke factoren mensen er toe aanzetten om meer delicten te plegen in gemeenten met een hoge werkloosheidsgraad. De vaststelling is eeuwen oud, maar als criminoloog mogen we ons niet blindstaren op de cijfertjes op zich. Dit zou resulteren in cijferfetisjisme en dat is wel het laatste dat wij met deze cursus beogen.

6. En wat als de meetniveaus van twee variabelen verschillend zijn?

Hierboven werden de meest klassieke bivariate beschrijvende associatiematen besproken. Van een criminoloog wordt verwacht dat hij of zij verantwoord kan kiezen tussen een reeks associatiematen. Soms zijn de meetniveaus van twee criminologisch relevante variabelen verschillend. Strikt genomen dient men zich te houden aan het meetniveau van de variabele met het laagste meetniveau. Is de afhankelijke variabele gemeten op het metrisch niveau (bijvoorbeeld het aantal criminele veroordelingen) en de onafhankelijke variabele gemeten op het nominaal niveau (bijvoorbeeld het volgen van een bepaalde behandeling of interventie), dan dient men strikt genomen een analysetechniek te kiezen op het nominaal niveau. Hiertoe dienen we de afhankelijke variabele te hercoderen naar een lager meetniveau. Hierdoor ontstaat er eigenlijk informatieverlies. In deze situatie bestaat er een veel gebruikte analysetechniek: **het vergelijken van gemiddelden**. In deze situatie worden dan de gemiddelde scores beschreven voor elke subgroep. De vergelijking van gemiddelde scores tussen groepen komt heel vaak voor en we behandelen deze techniek later wanneer we de variantieanalyse introduceren als onderdeel van de inferentiële statistiek.

Er zijn echter enkele andere vaak voorkomende situaties waar we het nog niet over gehad hebben. Wat doe je als een variabele ordinaal is en een andere variabele nominaal? In zo een situatie kiest men strikt genomen een analysetechniek op nominaal niveau. Is een afhankelijke variabele van het nominaal niveau en een onafhankelijke variabele van het metrisch niveau, dan kiest men ook strikt genomen voor een analysetechniek gemeten van het nominaal niveau. Je merkt dat er heel wat situaties bestaan die niet voldoen aan hetgeen we tot hiertoe behandeld hebben. Dat wil zeggen dat we hier eigenlijk nog maar het topje van de ijsberg hebben behandeld. Desalniettemin is het zo dat wie zich doorheen deze cursus sleept, een goede basis heeft voor vervolgcursussen. Een introductie tot alle niet-metrische analysetechnieken die criminologen gebruiken wanneer zij afhankelijke variabelen hebben gemeten op ordinaal niveau of nominaal niveau en onafhankelijke variabelen op metrisch niveau valt buiten het bestek van deze inleidende cursus. Echter, onthoud dat er ordinale en nominale varianten van de regressieanalyse bestaan.[14]

[14] Een zeer vaak voorkomende analysetechniek wanneer de afhankelijke variabele van het nominale niveau is, is de binomiale logistische regressie. Van alle niet-metrische (multivariate en bivariate) analysetechnieken komt deze wellicht het meest voor in de criminologische wetenschappen. Binomiale logistische regressie analyse is geschikt voor een afhankelijke variabele die categorisch van aard is: er zijn maar twee categorieën. We behandelen deze techniek echter niet in het licht van deze syllabus.

7. Leerdoelen

Dit hoofdstuk beoogt diverse leerdoelen. Je wordt verwacht te weten onder welke omstandigheden je kiest voor een correlatie of voor een regressieanalyse. We verwachten dat je zelf (weliswaar met behulp van een rekenmachine) een correlatie- en regressieanalyse kan uitvoeren. Je dient alle coëfficiënten die betrekking hebben op zulke analyses goed te begrijpen: de ongestandaardiseerde en gestandaardiseerde richtingscoëfficiënt, het intercept, R-kwadraat, het begrip "model fit". Maak daarom een inventaris van begrippen en tracht deze uit te leggen in je eigen woorden. Houd daarbij de statistische achtergrond in het oog.

Samenvattend schema voor bivariate beschrijvende analysetechnieken

VERBANDEN TUSSEN 2 VARIABELEN: symmetrische associatiematen			
	Nominaal	***Ordinaal***	***Metrisch***
verbanden tussen 2 variabelen	***Percentageverschil, odds ratio, Chi², Phi, Cramer's V***	***Spearman's rang-correlatie; Kendall's Tau-b, gamma***	***Correlatie-analyse***

Dependente bivariate analysetechnieken

Onafhankelijke variabele	**Afhankelijke variabele**	**ANALYSETECHNIEK**	**Parameters**
Interval/Ratio	***Interval/Ratio***	**Lineaire Regressieanalyse**	**Intercept** **Ongestandaardiseerde en gestandaardiseerde richtingscoëfficiënt** **Determinatiecoëfficiënt** **Regression Sum of Squares** **Residual Sum of Squares**