Hoofdstuk 10
De partiële correlatie als introductie tot de multivariate statistiek

1. Inleiding

Tot nu begaven we ons op het domein van de bivariate statistiek. Bivariate analyses zijn zinvol. Ze leren ons of twee kenmerken samen geobserveerd worden bij statistische onderzoekseenheden dan louter op basis van toeval kan verwacht worden. Bivariate analyses hebben echter belangrijke beperkingen. In de sociale verklarende wetenschappen zijn we vaak geïnteresseerd in de effecten van meerdere onafhankelijke variabelen op een afhankelijke variabele. Dat is in de criminologie niet anders. Het grote probleem in de bivariate analyse zit hem in het feit dat we uit een bivariate analyse weinig kunnen besluiten. Een bivariaat verband kan potentieel veroorzaakt worden door een derde, storende variabele waar we geen rekening mee hielden. Vaak omdat andere onafhankelijke variabelen, die aan eenzelfde afhankelijke variabele gerelateerd zijn, onderling ook samenhangen, is het bivariate effect vaak een overschatting van de realiteit. Dit is een zeer belangrijke reden om meer dan één onafhankelijke variabele in een analyse in te voeren. Van zodra we meerdere onafhankelijke variabelen in een analyse inbrengen, begeven we ons op het domein van de **multivariate statistiek**. Het meest éénvoudige model is het meervoudige *regressiemodel* met twee onafhankelijke variabelen. Dit is een eenvoudige extensie van de bivariate regressieanalyse. In de bivariate lineaire regressie onderzoeken we het effect van één onafhankelijke variabele op één afhankelijke variabele (onder de veronderstelling dat het verband tussen beide variabelen lineair is). Bij de meervoudige lineaire regressie onderzoeken we de effecten van meer dan één onafhankelijke variabele, maar nog steeds op slechts één afhankelijke variabele.[16]

Uiteraard is de lineaire wereld van de multivariate analyse niet beperkt tot het uitvoeren van een meervoudige regressieanalyse. De meervoudige lineaire regressieanalyse leent zich echter perfect om uiteen te zetten wat de beperkingen zijn van bivariate analyses. Waarom is de meervoudige analyse van statistische gegevens zo belangrijk in de sociale wetenschappen in het algemeen en de criminologie in het bijzonder? Hiervoor zijn verschillende redenen te bedenken.

[16] Zijn er meerdere afhankelijke variabelen, dan spreken we van *multivariate lineaire regressie*.

- **In de eerste plaats** is het zo dat de sociale werkelijkheid multivariaat is. Er zijn nu eenmaal **meerdere determinanten** verbonden aan regelovertredend gedrag. De criminoloog die ervoor kiest deze werkelijkheid aan de hand van statistische analysetechnieken te onderzoeken, dient zich hiervoor te wenden tot de multivariate analyse.
- **Ten tweede** is het zo dat het in de criminologie moeilijk is om fenomenen **geïsoleerd** te bestuderen. In gecontroleerde wetenschappen is het wel mogelijk om fenomenen aan de hand van experimenten in geïsoleerde positie te bestuderen. Dit gaat slechts in beperkte mate in de sociale wetenschappen. Zowel de causale als niet causale analyse van criminaliteitsfenomenen maakt daarom gebruik van het niet experimentele onderzoeksdesign als alternatief. Dit is het principe van **de statistische controle**. Elke variabele die gecorreleerd is met de onafhankelijke variabele en die meebepalend kan zijn voor de score op de afhankelijke variabele, is een storende variabele. We geven een voorbeeld uit onderzoek naar individuele verschillen in delinquent gedrag: we zijn geïnteresseerd in de impact van morele normen op individuele betrokkenheid bij criminaliteit. Delinquente normen kunnen samenhang met de opvoeding, de sociale controle in het gezin, de bindingen die jongeren hebben met de conventionele samenleving, de buurt waarin men opgroeit, de mate waarin jongeren in contact komen met criminele rolpatronen,… We kunnen ons bijvoorbeeld afvragen of het verband tussen delinquente waarden en individuele verschillen in delinquentie zou veranderen als we ook rekening zouden houden met ouderlijke controle.

  Controleren voor een storende variabele, bijvoorbeeld ouderlijke controle, betekent dat we de samenhang tussen delinquente waarden en delinquent gedrag bestuderen voor jongeren uit gezinnen met eenzelfde niveau van ouderlijke controle. We houden de variabele ‘ouderlijke controle’ constant, zodat we alleen kijken naar variatie in de afhankelijke variabele ‘delinquent gedrag’ en in de onafhankelijke variabele ‘delinquente waarden’ die niet samenhangt met een verschil in ouderlijke controle.
- **Ten derde** is het zo dat achter een bivariaat verband meer kan schuil gaan dan op het eerste zicht lijkt. We leggen dit verderop gedetailleerd uit met een criminologisch voorbeeld aan de hand van het *“schijneffect”* of de *spurieuze verbanden, maar ook de conditionele causaliteit of de statistische interactie*. Een spurieus verband verwijst naar een correlatie tussen X en Y omdat beide variabelen afhangen van een derde variabele Z. We behandelen het ‘schijneffect’ grondig in het vervolg van dit hoofdstuk. Interactie betekent dat de samenhang tussen twee variabelen afhankelijk is van de waarden van

een andere variabele. Er is sprake van interactie wanneer het effect van een variabele X op een variabele Y verschilt naargelang de waarden van een variabele Z. Bijvoorbeeld: omgevingskenmerken kunnen een effect hebben op criminaliteit maar de sterkte van het effect kan afhankelijk zijn van bepaalde individuele kenmerken. We behandelen statistische interactie grondig in hoofdstuk 11.

In de volgende paragrafen leggen we de theoretische achtergrond van de partiële correlatie uit. We tonen hoe je zelf handmatig een partiële correlatiecoëfficiënt berekent.

2. De partiële correlatiecoëfficiënt

De partiële correlatie is de correlatie tussen twee variabelen (X1 en Y), onder statistische controle van één of meerdere storende variabelen. De partiële correlatieanalyse wordt uitgevoerd wanneer we willen nagaan of :

- de samenhang tussen X1 en Y **spurieus** is**.**
- de samenhang tussen X1 en Y **indirect** is.

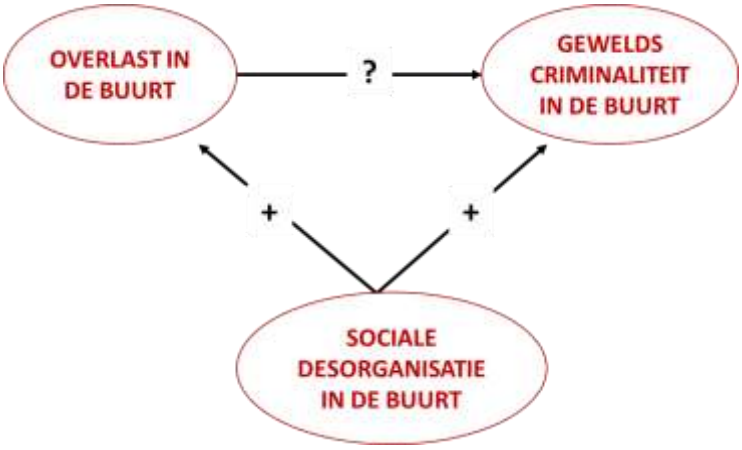
Hierna geven we twee criminologische voorbeelden.

Spurieuze samenhang of een schijnverband: een voorbeeld uit de sociale desorganisatietheorie

Het uitgangspunt betreft de samenhang tussen '*overlast in de buurt'* en '*geweldscriminaliteit in de buurt'*. De vraag is of deze samenhang werkelijk kan beschouwd worden als een effectrelatie: als er meer overlast in de buurt is, leidt dit dan naar meer geweldscriminaliteit in de buurt ? Een alternatieve verklaring is immers eveneens mogelijk: namelijk een derde variabele kan verantwoordelijk zijn voor zowel 'overlast in de buurt' als voor 'geweldscriminaliteit in de buurt'. In dat geval zou de samenhang tussen deze twee variabelen 'wegverklaard' worden door die derde variabele. De oorspronkelijk geobserveerde samenhang zou dan een schijnverband zijn.

Een voorbeeld van zo'n derde variabele is *'sociale desorganisatie in de buurt'.* Veronderstel een positieve relatie tussen sociale desorganisatie in de buurt en overlast in de buurt (hoe hoger de niveaus van sociale desorganisatie, hoe meer overlast) en veronderstel een positieve relatie tussen sociale desorganisatie in de buurt en geweldscriminaliteit in de buurt (hoe hoger de niveaus van sociale desorganisatie, hoe hoger de niveaus geweldscriminaliteit).
Figuur 1 geeft een visuele voorstelling van een spurieuze relatie of een schijnverband.

Figuur 1: Een spurieuze relatie: voorbeeld van een schijnverband uit de sociale desorganisatietheorie



De plustekens verwijzen naar een positieve samenhang, het vraagteken verwijst naar een potentieel spurieuze relatie. De bestaande bivariate samenhang is het gevolg van *een gemeenschappelijke oorzaak*. In het voorbeeld wordt sociale desorganisatie gezien als de gemeenschappelijke oorzakelijke factor die verklaart waarom buurten zowel te maken hebben met overlast als met geweldscriminaliteit.

In deze analyse zijn drie variabelen betrokken[17], gemeten op intervalmeetniveau. De gemeenschappelijke oorzaak 'sociale desorganisatie' is de **controlevariabele**. . De toegepaste analysetechniek is **de partiële correlatieanalyse.** De achterliggende veronderstelling is dat de oorspronkelijke samenhang tussen overlast en criminaliteit verdwijnt wanneer men **controleert** voor sociale desorganisatie. Concreet betekent dit dat buurten die verschillen in mate van overlast ook verschillen in mate van geweldscriminaliteit maar dat deze co-variatie volledig te wijten is aan verschillen in sociale desorganisatie. Voor buurten met gelijke niveaus van sociale desorganisatie, verdwijnt de samenhang tussen overlast en geweldscriminaliteit. De samenhang tussen overlast en geweldscriminaliteit is **spurieus of een schijnverband.**

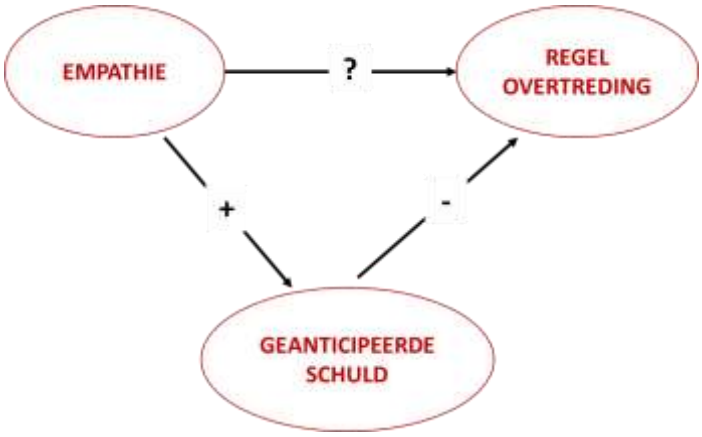
[17] Een partiële correlatieanalyse hoeft zich niet te beperken tot 3 variabelen (X1 en Y onder controle van X2). De analyse kan uitgebreid worden naar meerdere controlevariabelen. De partiële correlatie kan berekend worden tussen X1 en Y onder controle van X2, X3, X4. In dit geval is de eerste stap in de analyse geen bivariate maar een multiple regressie analyse van X1 op X2, X3 en X4, hetgeen resulteert in de residuele term X1 – X1hat. De tweede stap in de analyse is een multiple regressie analyse van Y op X2, X3 en X4, hetgeen resulteert in de residuele term Y – Yhat. De zero-order correlatie tussen beide residuele termen is dan de partiële correlatiecoëfficiënt onder controle van X2, X3 en X4.

Indirecte relatie : toepassing van de empathie-ontwikkelingstheorie van Hoffman (2001)[18]

Het uitgangspunt betreft de samenhang tussen '*empathie'* en *'regelovertreding'*. De vraag is of deze samenhang werkelijk kan beschouwd worden als een effectrelatie: leidt meer/minder empathie tot minder/meer regelovertreding ? Het is immers mogelijk dat de samenhang te wijten is aan een vorm van indirecte samenhang via een intermediaire variabele. Een voorbeeld van zo'n intermediaire variabele is '*geanticipeerde schuld'.*

Veronderstel: we observeren een positieve relatie tussen empathie en geanticipeerde schuld (hoe meer empathie, hoe meer geanticipeerde schuld) en een negatieve relatie tussen geanticipeerde schuld en regelovertreding (hoe meer geanticipeerde schuld, hoe minder regelovertreding). Dit mechanisme van **indirecte samenhang** is visueel voorgesteld in Figuur 2.

Figuur 2: Voorbeeld van een indirecte relatie gebaseerd op het empathie-ontwikkelingsmodel van Hoffman (2001)



Het plus- en minteken verwijzen respectievelijk naar een positieve en negatieve samenhang. Het vraagteken refereert ernaar dat de oorspronkelijk veronderstelde samenhang tussen empathie en regelovertreding niet direct maar indirect van aard is. De toegepaste analysetechniek is ook in dit geval een **partiële correlatieanalyse.** Net zoals in het geval van spurieuze samenhang (zie hierboven) is de variantie die empathie (X1) en regelovertreding (Y) delen met elkaar, in het geval van indirecte samenhang, (bijna) volledig te wijten aan de variantie die beide variabelen delen met de controlevariabele 'geanticipeerde schuld'. Dit betekent concreet dat voor respondenten met gelijke niveaus van geanticipeerde schuld, de

[18] Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

relatie tussen empathie en regelovertreding verdwijnt. Dit is ook een toepassing van het mechanisme van de partiële correlatie analyse: de samenhang tussen empathie en regelovertreding is **indirect.**

Dus, het mechanisme van de partiële correlatieanalyse is hetzelfde voor zowel een spurieuze als indirecte samenhang. De richting van de pijlen in het conceptueel diagram is echter verschillend. In geval van:

- **spurieuze relatie** is de controle variabele een gemeenschappelijke oorzaak (zie Figuur 1). (X1 ⟵ Z ⟶ Y)
- **Indirecte relatie** is de controle variabele een intermediaire variabele (zie Figuur 2).
  (X1 ⟶ Z ⟶ Y)

Waarom is de berekening van een partiële correlatie nodig ?

Waarom de berekening van een partiële correlatie nodig is, kunnen we best uitleggen aan de hand van een venndiagram.

Veronderstel dat we gegevens hebben over drie variabelen:

X = **Inspanningen** gedaan voor het instuderen van het vak statistiek

Y = **De behaalde score** van de student op het examen statistiek

Z = De mate waarin de student **angst heeft voor het vak statistiek**

We vinden de volgende samenhangen :

X versus Y: $\mathbf{r}_{XY}$ = +0.20 $\mathbf{r}^2_{XY}$ =0 .04

X versus Z: $\mathbf{r}_{XZ}$ = +0.80 $\mathbf{r}^2_{XZ}$ = 0.64

Y versus Z: $\mathbf{r}_{YZ}$ = —0.40 $\mathbf{r}^2_{YZ}$ = 0.16

We zien een zwakke-matige positieve samenhang tussen inspanningen (X) en behaalde scores (Y) : hoe meer inspanningen, hoe hoger de scores (ryz = 0.20).

We zien een zeer sterke positieve samenhang tussen inspanningen (X) en angst voor het vak (Z) : hoe meer inspanningen, hoe meer angst voor het vak (rxz = 0.80).

We zien een matig-sterke negatieve samenhang tussen behaalde scores (Y) en angst voor het vak (Z) : hoe hoger behaalde scores, hoe lager angst voor het vak (ryz = -0.40).

We constateren dus een samenhang tussen inspanningen en behaalde scores (tussen X en Y) maar ook tussen inspanningen en angst (tussen X en Z) en tussen behaalde scores en angst (tussen Y en Z).

We zien dat inspanningen (X) 4% van de variabiliteit deelt met behaalde scores (Y) , 64% van de variabiliteit deelt met angst voor het vak (Z) en dat behaalde scores (Y) 16% van de variabiliteit deelt met angst voor het vak (Z).

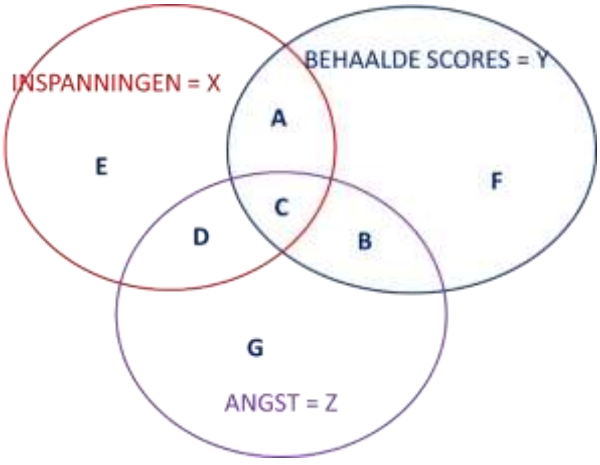
Dus: inspanningen en behaalde scores (X en Y) delen variabiliteit of gemeenschappelijke variantie met angst voor het vak (Z).

Om een maat te krijgen van <u>de unieke samenhang</u> tussen inspanningen en behaalde scores moeten we rekening houden met angst voor het vak.

Of anders geformuleerd: om een zuivere schatting te krijgen van de relatie tussen X en Y, moet de gemeenschappelijke variantie met Z, een "potentiële confounder" verwijderd worden. Die gemeenschappelijke variantie kan er immers toe leiden dat een bivariate analyse misleidende resultaten geeft.

Figuur 3 geeft een visuele voorstelling van de variantie van elke variabele en de gedeelde variantie.

Figuur 3: Drie variabelen en de gedeelde variabiliteit



We stellen X, Y en Z voor als cirkels. De volledige oppervlakte van elke cirkel is gelijk aan 100% van de variatie in elke variabele.

Figuur 3 toont tevens de overlap of variabiliteit die de variabelen met elkaar delen[19]:

[19] Om helemaal correct te zijn, zou de variatie die de variabelen met elkaar delen *proportioneel* moeten voorgesteld worden. Voorbeeld: als twee variabelen 10% van de variabiliteit gemeenschappelijk hebben, dan zou 10% van hun cirkeloppervlakte moeten ingekleurd worden als gemeenschappelijk. We hebben er in dit voorbeeld voor geopteerd om dit niet te doen teneinde de interpreteerbaarheid te vergroten.

- A = variabiliteit in behaalde scores (Y) **uniek** gedeeld met inspanningen (X)
- B = variabiliteit in behaalde scores (Y) **uniek** gedeeld met angst voor het vak (Z)
- C= variabiliteit in behaalde scores (Y) **gedeeld met inspanningen (X) en angst voor het vak (Z)**
- D= variabiliteit in inspanningen (X) **uniek** gedeeld met angst voor het vak (Z)
- E= variabiliteit in inspanningen (X) **niet** gedeeld met behaalde scores (Y) noch met angst voor het vak (Z)
- F = variabiliteit in behaalde scores (Y) **niet** gedeeld met inspanningen (X) noch met angst voor het vak (Z)
- G = variabiliteit in angst voor het vak (Z) **niet** gedeeld met inspanningen (X) noch met behaalde scores (Y)
- A + C = gedeelde variabiliteit tussen inspanningen en behaalde scores (X en Y) = 4%
  Waarvan A = unieke variabiliteit tussen inspanningen en behaalde scores
  Waarvan C = variabiliteit die inspanningen en behaalde scores delen met angst voor het vak

- C + B = gedeelde variabiliteit tussen behaalde scores en angst voor het vak (Y en Z) = 16%
  Waarvan B = unieke variabiliteit tussen behaalde scores en angst voor het vak
  Waarvan C = variabiliteit die behaalde scores en angst voor het vak delen met inspanningen

- C + D = gedeelde variabiliteit tussen inspanningen en angst voor het vak (X en Z) = 64%
  Waarvan D = unieke variabiliteit tussen inspanningen en angst voor het vak
  Waarvan C = variabiliteit die inspanningen en angst voor het vak delen met behaalde scores

Bedenk nu het volgende :

Gebied A = de variabiliteit in behaalde scores (Y) die **uniek** gedeeld wordt met inspanningen (X) maar in combinatie met C is dit de totale variatie in behaalde scores (Y) die gedeeld wordt met inspanningen (X) = 4%. De correlatie tussen behaalde scores (Y) en inspanningen (X) leert ons dat beide variabelen 4% van de variabiliteit delen. In werkelijkheid is maar een gedeelte uniek (gebied A) want er is een gedeelde overlap met angst voor het vak (gebied C).

Daarnaast hebben zowel inspanningen (X) als behaalde scores (Y) een unieke overlap met angst voor het vak (Z): respectievelijk gebied D en gebied B.

De partiële correlatie is de unieke samenhang tussen twee variabelen X en Y onder controle van Z. Die unieke samenhang is de variatie in X en Y (gebied A) die overblijft als de volledige samenhang met Z is 'weggenomen' (dus gebieden B, C en D).

De partiële correlatie analyse is een procedure die de gemeenschappelijke variantie verwijdert en de correlatie weergeeft tussen X en Y, onder de statistische controle van Z.

Het komt er dus op aan om de gemeenschappelijke variantie die X en Z, en Y en Z hebben op voorhand uit te schakelen, zodat wat overblijft de waargenomen variantie is in X en de waargenomen variantie in Y.

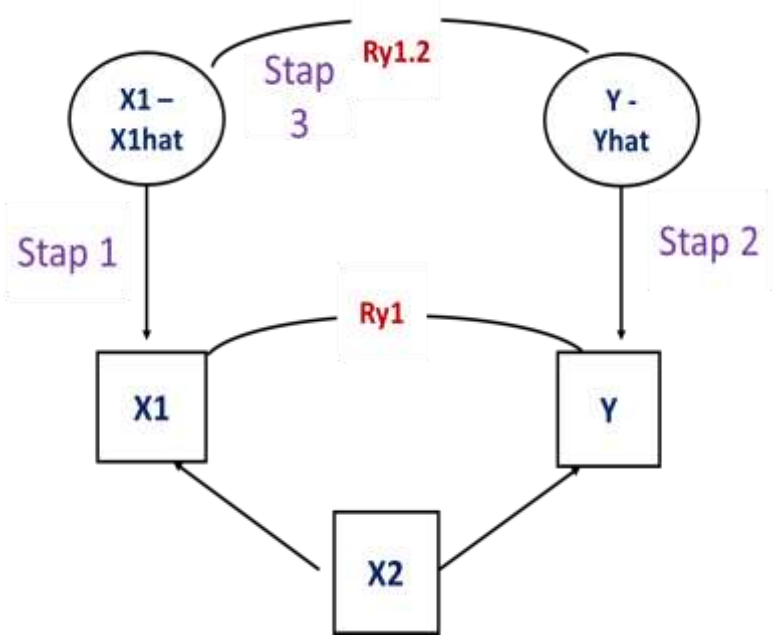
De partiële correlatie kan berekend worden aan de hand van een reeks van regressievergelijkingen.

3. De berekening van de partiële correlatiecoëfficiënt ahv regressievergelijkingen

Hierna volgt een uitgewerkt voorbeeld van de berekening van de partiële correlatiecoëfficiënt op basis van 3 variabelen, X1 , X2 en Y.

Figuur 3 geeft een schematische voorstelling van de werkwijze.

Figuur 3: Schematische voorstelling van een partiële correlatie analyse op basis van 3 variabelen X1, X2 en Y



We nemen als uitgangspunt de samenhang tussen X1 en Y (= **Ry1**) en onderzoeken de partiële correlatie tussen X1 en Y onder controle van X2 (= **Ry1.2**).
De beste manier om het principe van partiële correlatieanalyse te begrijpen is door de variantie die X2 deelt met X1 en die X2 deelt met Y te verwijderen zodat de partiële correlatie de zero-order correlatie (= zonder controlevariabelen) is tussen de residuele termen X1 – X1hat en Y – Yhat. De berekening verloopt in drie stappen.

STAP 1 : Verwijder de variantie die X1 en X2 met elkaar delen

- Voer een bivariate regressie analyse uit van X1 op X2
- Bereken de verwachte waarden voor X1 (=X1hat) op basis van de regressievergelijking

  $X1hat = a + b.X2$
- Bereken de residuele waarden: X1 - X1hat

STAP 2 : Verwijder de variantie die Y en X2 met elkaar delen

- Voer een bivariate regressie analyse uit van Y op X2
- Bereken de verwachte waarden voor Y (= Yhat) op basis van de regressievergelijking :

  $Yhat = a + b.X2$
- Bereken de residuele waarden: Y – Yhat

STAP 3 : bereken de partiële correlatiecoëfficiënt: ry1.2

- Bereken de correlatie tussen de residuele termen : X1 – X1hat en Y – Yhat.
- De zero-order correlatie tussen de residuele termen is de partiële correlatiecoëfficiënt **ry1.2** (= notatie voor de correlatie tussen Y en X1 onder controle van X2).

STAP 4 : voer een significantietoets uit van de partiële correlatiecoëfficiënt

- **bereken de t-ratio**

$$t = \frac{Rxy.z * \sqrt{N-3}}{\sqrt{1 - R^2xy.z}}$$

- **zoek de kritieke t-waarde op in de t-tabel, gegeven het aantal vrijheidsgraden en een gekozen α (=kans op een type I-fout)**
-

Data-matrix: een fictief voorbeeld

De statistische analyse-eenheden zijn individuen (N=10). Alle variabelen zijn gemeten op metrisch meetniveau met scores die gaan van 0 tot 9.

- Y = afhankelijke variabele
- X1 = onafhankelijke variabele
- X2 = controle variabele.

	Y	**X1**	**X2**
	3	2	1
	3	1	2
	2	5	3
	6	2	3
	4	3	4
	5	6	5
	6	5	6
	4	8	7
	9	7	8
	8	3	9
gemiddelde	**5,0**	**4,2**	**4,8**
standaardafwijking	**2,261**	**2,348**	**2,658**

Correlatiematrix	**Y**	**X1**	**X2**
Y	1		
1	0,251	1	
X2	0,777	0,612	1

STAP 1 : Verwijder de variantie die X1 en X2 met elkaar delen

(voer een regressie analyse uit van X1 op X2)

$$X1hat = a + b.X2$$

$$X1hat = 1{,}60 + 0{,}54.X2$$

X1	**X2**	**X1hat (=voorspelde waarden X1 op basis van X2)**	**X1-X1hat (= residu)**
2	1	2,14 (= 1.60 + 0.54.1)	**-0,14 (= 2 – 2.14)**
1	2	2,69 (= 1.60 + 0.54.2)	**-1,69 (= 1 – 2.69)**
5	3	3,23	**1,77**
2	3	3,23	**-1,23**
3	4	3,77	**-0,77**
6	5	4,31	**1,69**
5	6	4,85	**0,15**
8	7	5,39	**2,61**
7	8	5,93	**1,07**
3	9	6,47	**-3,47**

STAP 2 : Verwijder de variantie die Y en X2 met elkaar delen
(voer een regressie analyse uit van Y op X2)

Yhat = a + b. X2

Yhat = 1,83 + 0,66.X2

Y	**X2**	**Yhat (=voorspelde waarden Y op basis van X2)**	**Y- Yhat (=residu)**
3	1	2,49 (= 1.83 + 0.66.1)	**0,51 (= 3 – 2.49)**
3	2	3,15	**-0,15**
2	3	3,81	**-1,81**
6	3	3,81	**2,19**
4	4	4,47	**-0,47**
5	5	5,13	**-0,13**
6	6	5,79	**0,21**
4	7	6,45	**-2,45**
9	8	7,11	**1,89**
8	9	7,77	**0,23**

STAP 3 : Bereken de partiële correlatiecoëfficiënt: ry1.2

X1 – X1hat	**(X1 – X1hat)²**	**Y – Yhat**	**(Y – Yhat)²**	**(X1 – X1hat)(Y – Yhat)**
-0.14	0.02	0.51	0.26	-0.07
-1.69	2.86	-0.15	0.02	0.25
1.77	3.13	-1.81	3.28	-3.20
-1.23	1.51	2.19	4.80	-2.69
-0.77	0.59	-0.47	0.22	0.36
1.69	2.86	-0.13	0.02	-0.22
0.15	0.02	0.21	0.04	0.03

2.61	6.81	-2.45	6.00	-6.39
1.07	1.14	1.89	3.57	2.02
-3.47	12.04	0.23	0.05	-0.80
	30.98		18.26	-10.71

$$Ry1.2 = \frac{\sum(X1 - X1hat)(Y - Yhat)}{\sqrt{\sum(X1 - X1hat)^2 \sum(Y-Yhat)^2}} = \frac{-10.71}{\sqrt{30,98*18,26}} = \frac{-10.71}{23.78} = \mathbf{-0.45}$$

Een partiële correlatiecoëfficiënt kan net zoals de correlatiecoëfficiënt een waarde aannemen die varieert van -1 tot +1. De interpretatie ervan is identiek aan de correlatiecoëfficiënt.
Hier: de partiële correlatiecoëfficiënt tussen Y en X1 onder controle van X2 bedraagt -0.45 : dit is een matig-sterke negatieve samenhang (hoe meer…., hoe minder….).

STAP 4 : voer een significantietoets uit van de partiële correlatiecoëfficiënt

De statistische significantie van een partiële correlatie kan worden getest aan de hand van een t-ratio. We gebruiken volgende formule (toegepast op het voorbeeld) :

$$t = \frac{Ry1.2*\sqrt{N-3}}{\sqrt{1 - R^2y1.2}}$$

Waarbij t= t ratio als statistische significantietoets voor de partiële correlatiecoëfficiënt
N = aantal paren scores voor X1 en Y (hier = 10)
Ry1.2 = waarde van de partiële correlatiecoëfficiënt (hier = -.45)

$$t = \frac{-.45*\sqrt{10-3}}{\sqrt{1 - 0.20}} = \frac{-1.19}{0.89} = -1.34$$

Df = N – 3 = 7
α = .05
kritieke t-waarde in t-tabel : 2.365

Aangezien de t-ratio de kritieke t-waarde gevonden in de t-tabel, voor 7 vrijheidsgraden en een $\alpha = .05$, niet overschrijdt, besluiten we dat de partiële correlatiecoëfficiënt van -.45 voor 10 statistische eenheden niet statistisch significant is. We kunnen het resultaat niet veralgemenen naar de populatie omdat we niet kunnen uitsluiten dat het gevonden resultaat in deze steekproef 'toevallig' is (uiteraard hebben we hier te maken met een zeer kleine steekproef: N=10).

Besluit:

Het resultaat van de partiële correlatie analyse is in dit voorbeeld nogal drastisch: de samenhang tussen X1 en Y (= Ry1) bedraagt aanvankelijk r=0.25. Dit is een zwakke positieve relatie. Wanneer we echter controleren voor X2 wordt deze zwakke positieve relatie echter matig-sterk en negatief.

De partiële correlatiecoëfficiënt **Ry1.2 = -0.45**

In veel gevallen zal het zo zijn dat het bivariate verband tussen 2 variabelen zwakker wordt of zelfs volledig verdwijnt wanneer gecontroleerd wordt voor een derde variabele. Een partiële correlatie mag echter niet naïef begrepen worden als het verdwijnen van een samenhang tussen twee variabelen onder controle van een derde variabele. Ee partiële correlatie kan drastische gevolgen hebben, zoals we zagen in voorgaand voorbeeld. Waar de oorspronkelijke samenhang tussen X1 en Y zwak positief is (hoe meer X1, hoe meer Y), wordt deze niet alleen sterker onder controle van X2, maar verandert deze ook van teken (de samenhang wordt negatief).

Concreet betekent dit dat de relatie tussen X1 en Y voor personen met gelijke scores op X2 matig-sterk èn negatief wordt (hoe meer X1, hoe minder Y)!

De samenhang tussen X1 en Y is daarom een misleidende weergave van de werkelijkheid omdat de variantie die beide variabelen met elkaar delen hoofdzakelijk het resultaat is van de variantie die zij delen met X2.

Tabel 1 geeft een overzicht van mogelijke wijzigingen in de initiële samenhang tussen twee variabelen als gevolg van een partiële correlatie analyse en wat je hieruit als onderzoeker kan besluiten.

Tabel 1 : Mogelijke wijzigingen in bivariate samenhang als gevolg van partiële correlatie analyse

Initiële samenhang tussen X en Y	**Correlatie tussen X en Y onder controle van Z**	**Omschrijving**	**Besluit**
Rxy = 0.50	Rxy.z= 0.50	De initiële samenhang tussen X en Y onder controle van Z verandert niet.	***Een partiële correlatie maakt geen enkel verschil in de samenhang tussen X en Y.*** ***Z is noch een gemeenschappelijke oorzaak, noch een intermediaire variabele in de relatie tussen X en Y.***
Rxy = 0.50	Rxy.z= 0.00	De initiële samenhang tussen X en Y onder controle van Z verdwijnt volledig.	***Z is mogelijk een gemeenschappelijke oorzaak (= antecedent variabele) van de initiële samenhang tussen X en Y (spurieuze relatie).*** ***Z kan ook een intermediaire variabele zijn in de relatie tussen X en Y (indirecte relatie).***[20]
Rxy = 0.05	Rxy.z= 0.60	De initiële samenhang tussen X en Y onder controle van Z verandert van een zeer lage samenhang naar een sterke samenhang.	***Z is een suppressor-variabele in de relatie tussen X en Y. Z 'verbergt' de relatie tussen X en Y, die zichtbaar wordt wanneer de gedeelde variatie tussen X en Z en Y en Z 'verwijderd' wordt.***

[20] Het feit dat Z een gemeenschappelijke oorzaak (of antecedent variabele) is dan wel een intermediaire variabele die tussen de relatie van X en Y in staat, wordt bepaald op niet-statistische gronden. Op basis van een statistische partiële correlatie analyse kan enkel beslist worden of de samenhang tussen X en Y al dan niet wijzigt (behouden blijft, verzwakt, verdwijnt of versterkt).

4. Berekening van de partiële correlatiecoëfficiënt ahv rekenkundige formule

In de vorige paragraaf hebben we de partiële correlatiecoëfficiënt berekend aan de hand van een reeks regressievergelijkingen. Er bestaat echter een rekenkundige formule om uit de bivariate correlaties de partiële correlatie te berekenen.

$$\mathbf{r}_{X1Y\cdot X2} = \frac{\mathbf{r}_{X1Y} - (\mathbf{r}_{X1X2})(\mathbf{r}_{YX2})}{\mathbf{sqrt}[1 - \mathbf{r}^2_{X1X2}] \times \mathbf{sqrt}[1 - \mathbf{r}^2_{YX2}]}$$

We geven een voorbeeld:

X versus Y: $\mathbf{r}_{XY} = +.50$ $\mathbf{r}^2_{XY} = .25$
X versus Z: $\mathbf{r}_{XZ} = +.50$ $\mathbf{r}^2_{XZ} = .25$
Y versus Z: $\mathbf{r}_{YZ} = +.50$ $\mathbf{r}^2_{YZ} = .25$

$$\mathbf{r}_{XY\cdot Z} = \frac{0.50 - (0.50)(0.50)}{\mathbf{sqrt}[1 - 0.25] \times \mathbf{sqrt}[1 - 0.25]}$$

$$= +0.33$$

Besluit $\mathbf{r}^2_{XY\cdot Z} = 0.11$

Dezelfde redenering kan worden toegepast om de partiële correlatie te berekenen tussen X en Z, waarbij we de gemeenschappelijke variatie met Y moeten verwijderen.

$$\mathbf{r}_{XZ\cdot Y} = \frac{\mathbf{r}_{XZ} - (\mathbf{r}_{XY})(\mathbf{r}_{YZ})}{\mathbf{sqrt}[1 - \mathbf{r}^2_{XY}] \times \mathbf{sqrt}[1 - \mathbf{r}^2_{YZ}]}$$

en voor de berekening van de partiële correlatie tussen Y en Z, waarbij de effecten van X verwijderd zijn:

$$\mathbf{r}_{YZ\cdot X} = \frac{\mathbf{r}_{YZ} - (\mathbf{r}_{XY})(\mathbf{r}_{XZ})}{\mathbf{sqrt}[1 - \mathbf{r}^2_{XY}] \times \mathbf{sqrt}[1 - \mathbf{r}^2_{XZ}]}$$

Laten we een rekenvoorbeeld geven uit een criminologisch onderzoek. De variabelen C (crime / delinquency), **A** (Association delinquent peers) en **V** (low self-control). Er is duidelijk overlap tussen de drie theoretische variabelen uit de criminele etiologie. Er is een positief verband tussen lage zelfcontrole en delinquentie, tussen delinquente peers en lage zelfcontrole en tussen delinquentie en delinquente peers.

C versus A:	$\mathbf{r_{CA} = +0.49}$ $\mathbf{r^2_{CA} = 0.24}$
C versus V:	$\mathbf{r_{CV} = +0.73}$ $\mathbf{r^2_{CV} = 0.53}$
A versus V:	$\mathbf{r_{AV} = +0.59}$ $\mathbf{r^2_{AV} = .035}$

$$\mathbf{r}_{CA\cdot V} = \frac{\mathbf{r}_{CA} - (\mathbf{r}_{CV})(\mathbf{r}_{AV})}{\mathbf{sqrt}[1 - \mathbf{r}^2_{CV}] \times \mathbf{sqrt}[1 - \mathbf{r}^2_{AV}]}$$

$$\mathbf{r}_{CA\cdot V} = \frac{0.49 - (0.73)(0.59)}{\mathbf{sqrt}[1 - .53] \times \mathbf{sqrt}[1 - 0.35]}$$

$$\mathbf{r}_{CA\cdot V} = +.11$$

Besluit $\mathbf{r}^2_{CA\cdot V} = 0.01$

Hier zien we dat het bivariate verband een drastische overschatting van de realiteit laat zien. Wanneer we controleren voor V, dan is het verband tussen C en A quasi weg. Immers, een correlatie van 0.11 is niet echt noemenswaardig te noemen. Effecten onder .2 beschouwen we als uiterst zwak.

5. Suppressie-effect

Laten we tot slot het **suppressie-effect** nog even onder de loep nemen. Dit is belangrijk want het komt in de praktijk van het onderzoek wel eens voor en beginnende onderzoekers zijn zich er niet altijd van bewust. Dat heeft tot gevolg dat men resultaten verkeerd inschat en dat men een grotere kans heeft om nonsens-beleid te voeren. We hernemen het voorbeeld van hierboven. Veronderstel dat een professor metingen heeft uitgevoerd op drie kenmerken: inspanningen van studenten om een vak in te studeren, de scores van de studenten en de angst voor een vak bij studenten. Die drie kenmerken hebben elk een associatie met elkaar.

X = Inspanningen gedaan voor het instuderen van het vak

Y = De score van de student

Z = De mate waarin de student angst heeft voor het vak

Hier zijn de bivariate correlaties tussen de drie variabelen:

X versus Y: $\mathbf{r}_{XY}$ = +0.20 $\mathbf{r}^2_{XY}$ =0 .04

X versus Z: $\mathbf{r}_{XZ}$ = +0.80 $\mathbf{r}^2_{XZ}$ = 0.64

Y versus Z: $\mathbf{r}_{YZ}$ = —0.40 $\mathbf{r}^2_{YZ}$ = 0.16

Is het niet vreemd dat het verband tussen de scores van de student en de inspanningen gedaan door de student zo laag is? ($\mathbf{r}_{XY}$=+.20 en $\mathbf{r}^2_{XY}$=.04). Als je nader onderzoek doet, zal je zien wat er werkelijk aan de hand is. Als de angst voor het vak groter wordt, dan studeren studenten meer voor het vak, vandaar $\mathbf{r}_{XZ}$=+.80 en $\mathbf{r}^2_{XZ}$=.64. Aan de andere kant is het ook zo dat angst voor een vak en de score voor het vak negatief met elkaar correleren: angst kan de concentratie helemaal verstoren ($\mathbf{r}_{YZ}$=—.40 en $\mathbf{r}^2_{YZ}$=.16). Laten we het onderdrukkende effect van de angst voor het vak verwijderen en de bivariate correlatie tussen scores en inspanningen opnieuw bekijken. Wat zien we na wat rekenwerk?

$$\mathbf{r}_{XY\cdot Z} = \frac{0.20 - (0.80)(-.040)}{\mathbf{sqrt}[1 - 0.64] \times \mathbf{sqrt}[1 - 0.16]}$$

$\mathbf{r}_{XY\cdot Z}$ = +0.95

$\mathbf{r}_{XY}$=0.20 $\mathbf{r}^2_{XY}$=0.04
$\mathbf{r}_{XY\cdot Z}$=0.95 $\mathbf{r}^2_{XY\cdot Z}$=.090

De bivariate correlatie die eerder zwak was, heeft plaats gemaakt voor een heel sterke partiële correlatie: van een kleine 0.20 (bivariaat) naar een impressionante +0.95 (partieel).

6. Leerdoelen

Studenten dienen de grenzen van de bivariate statistiek te kennen en zelf in staat te zijn een partiële correlatie tussen twee variabelen onder de statistische controle van een derde variabele te berekenen. In werkelijkheid is het denkbaar dat de partiële correlatie wordt berekend tussen twee variabelen onder de controle voor meerdere variabelen, echter deze extensie maakt niet het voorwerp uit van dit verkennend handboek. Studenten dienen te begrijpen waarom multivariate analyses zo belangrijk zijn in de criminologie. Het principe van de statistische controle moet goed begrepen worden.