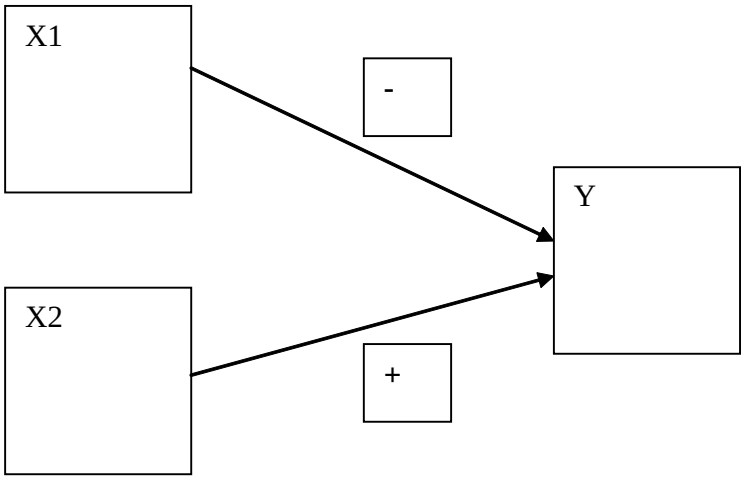
Hoofdstuk 11

Regressieanalyse met twee onafhankelijke variabelen

1. Inleiding

De **meervoudige (of multipele) regressieanalyse** wordt gebruikt wanneer men geïnteresseerd is in het verklaren van de spreiding in een afhankelijke variabele van het metrische niveau op basis van meerdere onafhankelijke variabelen die eveneens van het metrische niveau zijn. Bovendien: de werkelijkheid is complex en complexe fenomenen kunnen niet zomaar door één en slechts één kenmerk verklaard worden. De multipele regressie laat toe om de rechtstreekse statistische effecten van meerdere onafhankelijke variabelen tegelijkertijd te bestuderen. Dit basisformat (het pijltjesdiagram) heeft wat men noemt een ***convergente causale structuur***, dit wil zeggen: men is geïnteresseerd in de **rechtstreekse effecten van meerdere onafhankelijke variabelen (predictoren) op één afhankelijke variabele (het explanandum).**

Figuur: causaal diagram van een meervoudige regressie



We geven een voorbeeld aan de hand van twee onafhankelijke variabelen, X1 en X2.

- X1= de mate van sociale controle van de ouders op de adolescent. We noemen deze variabele verderop “sociale controle ouders”.
- X2= de mate waarin de adolescent het overtreden van regels als moreel aanvaardbaar beschouwt: We noemen deze variabele verderop “criminele waarden van adolescent”.

- Y= de frequentie van regelovertredend gedrag van de adolescent.

Vooraleer we als criminoloog-onderzoeker de analyse uitvoeren, is het belangrijk de veronderstellingen die de onderzoeker heeft, neer te schrijven. Waarom? Dit laat toe om achteraf te controleren of de veronderstellingen juist waren of fout. Veronderstellingen zijn vaak afgeleid uit theorie en het is belangrijk de terugkoppeling te maken.

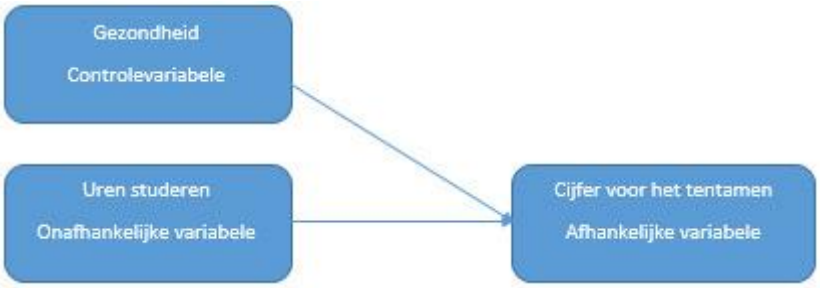
Hypothese 1: de mate van sociale controle van de adolescent heeft een **remmend** (= negatief) effect op de frequentie van regelovertreding, ***onafhankelijk van de mate waarin de adolescent het overtreden van regels als moreel aanvaardbaar beschouwt***.

Hypothese 2: De mate waarin de adolescent criminele waarden als moreel aanvaardbaar beschouwt, heeft een **duwend** (= positief) effect op de frequentie van regelovertreding van de adolescent, ***onafhankelijk van de sociale controle van de ouders***

2. De noodzaak voor het meten van controlevariabelen

Een controlevariabele **(confounder)** is een variabele die je meeneemt in je onderzoek, maar waar niet speciaal je aandacht naar uitgaat. Je neemt de variabele wel mee omdat deze invloed heeft op de afhankelijke variabele en omdat deze variabele ook samenhangt met de onafhankelijke variabele. De controlevariabelen weglaten uit het onderzoek zou betekenen dat de resultaten van je onderzoek minder accuraat zijn. Dit is vooral aan de orde wanneer je een statistische analyse doet en je een bepaalde oorzaak-gevolgrelatie statistisch wilt bewijzen. Sommige controlevariabelen kunnen geen oorzaken zijn, maar toch worden ze mee opgenomen. De reden waarom men er zoveel aandacht aan besteedt, is omdat men vreest dat een causale relatie zou verdwijnen onder de controle van de controlevariabele. Als men kan zeggen dat de resultaten van een studie, waarbij de relatie tussen enkele causale factoren en regelovertredend gedrag wordt bestudeerd, blijft bestaan onder controle van een aantal potentieel verstorende factoren, dan komt dat het geloof dat men zal hechten aan de resultaten ten goede. Wanneer we bijvoorbeeld het verband onderzoeken tussen de variabelen uit één theorie, en we kunnen zeggen dat het verband stand houdt onder controle van de variabelen uit een concurrerende theorie, dan is dit sterk.

Controlevariabelen in een conceptueel model



Voorbeeld controlevariabele in conceptueel model

In deze paragraaf geven we het voorbeeld van de student die veel studeert (onafhankelijke variabele) om zo een hoger cijfer te halen voor zijn tentamen (afhankelijke variabele). We voegen nu een controlevariabele aan het voorbeeld toe. Een voorbeeld van een controlevariabele in ons voorbeeld is 'Gezondheid'. Het valt te beargumenteren dat een slechte gezondheid ervoor zorgt dat de student een lager cijfer haalt voor het tentamen (invloed op afhankelijke variabele). Wanneer je kwantitatief onderzoek doet onder studenten om statistisch aan te tonen dat er een statistisch effect is van het aantal 'Uren studeren' en het behalen van 'Cijfer voor het tentamen' dan is het goed om additioneel te vragen naar de gezondheid van de student. Doe je dit niet, en neem je de variabele 'Gezondheid' dus ook niet mee als controlevariabele in je onderzoek, dan kan dit betekenen dat het aantonen van de oorzaak-gevolgrelatie een stuk moeilijker is. Het verband dat je vindt, kan namelijk voor een deel te wijten zijn aan de gezondheid van de student. Er zijn theoretisch verschillende scenario's mogelijk: gezondheid beïnvloedt hoe lang je kan studeren en heeft zo een indirect effect op je punten op je tentamen. Maar het kan zijn dat het rechtstreeks effect van het aantal uren studeren op het tentamencijfer verdwijnt wanneer je ook rekening houdt met gezondheid, en dat is een probleem, want dan vervalt het empirisch bewijs voor de hypothese dat er een rechtstreeks verband bestaat tussen het aantal uren studeren en het cijfer voor het tentamen. We willen de lezer er op wijzen dat controlevariabelen gebruiken belangrijk is, maar nooit ondoordacht mag gebeuren. Er moet een theoretische reden achter schuilgaan. Zo kan men zich de vraag stellen of we moeten controleren voor variabelen die geen oorzaken kunnen zijn in de betekenis die we aan oorzakelijkheid geven (productie, aanwezigheid van een mechanisme, manipuleerbaarheid). Echter, de traditie van het gebruiken van

controlevariabelen is zo sterk ingebakken in de sociale wetenschappen, dat men op de duur vergeet waarom men bepaalde controlevariabelen toevoegt aan de analyse.

3. De vergelijking tussen twee bivariate versus één meervoudige regressie

We spraken reeds over de mogelijke invloed van storende variabelen die samenhangen met de onafhankelijke variabele en die ook invloed uitoefenen op de afhankelijke variabelen. De onderlinge samenhang tussen onafhankelijke variabelen is op zich een realiteit die maakt dat we niet zomaar verschillende bivariate regressieanalyses kunnen uitvoeren als vervanging voor een meervoudige regressieanalyse. *De samenhang tussen onafhankelijke variabelen noemen we met een statistisch begrip "**multicollineariteit**".* In een lineaire regressieanalyse vertaalt zich deze samenhang in een correlatie tussen de onafhankelijke variabelen X1 en X2. Hoe sterker twee onafhankelijke variabelen samenhangen, hoe meer misleidend de uitkomsten van een bivariate analyse kunnen zijn. De samenhang tussen twee onafhankelijke variabelen mag ook niet te hoog zijn als we een meervoudige regressieanalyse willen uitvoeren. *Is deze samenhang hoger dan .50 dan dient men voorzichtig te zijn bij de interpretatie van de resultaten en is deze .80 of hoger, dan mag men absoluut geen meervoudige regressie met die onafhankelijke variabelen uitvoeren die dergelijke hoge samenhang hebben. De beide variabelen kunnen dan analytisch nauwelijks van elkaar onderscheiden worden en de resultaten van dergelijke analyse zijn hoogst onbetrouwbaar. De gestandaardiseerde effectparameters (richtingscoëfficiënten) kunnen dan onmogelijke waarden bekomen (groter dan 1 of kleiner dan –1).*

Daarom dient men bij een multivariate analyse steeds de voorbereidende controle op multicollineariteit door te voeren. Stel dat we geïnteresseerd zijn in het causale effect van criminele waarden en normen en de sociale controle in het gezin bij het verklaren van individuele verschillen in de frequentie van regelovertredend gedrag van jonge adolescenten. Waarom is het niet zonder gevaren om twee afzonderlijke bivariate regressieanalyses uit te voeren en de determinatiecoëfficiënten van beide afzonderlijke analyses bij elkaar op te tellen? Dit zou toch ogenschijnlijk het meest voor de hand liggen?

De partiële overlap of onderlinge samenhang tussen de beide onafhankelijke variabelen is de reden waarom we geen twee afzonderlijke bivariate analyses met elkaar mogen optellen. In het voorbeeld dat overigens kan worden overgedaan aan de hand van de oefendatabestanden, is de overlap tussen de twee onafhankelijke variabelen X1 en X2 reëel en bedraagt -0.455.

Correlations

		sociale controle ouders	criminele waarden
sociale controle ouders	Pearson Correlation	1	-,455**
	Sig. (2-tailed)		,000
	N	1521	1514
criminele waarden	Pearson Correlation	-,455**	1
	Sig. (2-tailed)	,000	
	N	1514	1542

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

Laten we nu eens de uitkomsten van de afzonderlijke bivariate regressieanalyses bekijken en vergelijken met de uitkomsten van één meervoudige regressieanalyse. We presenteren de gestandaardiseerde regressiecoëfficiënten omdat deze toelaten de vergelijking te maken tussen de sterkte van het effect van elke variabele op de afhankelijke variabele.

Tabel: Twee afzonderlijke bivariate regressieanalyses en één meervoudige regressieanalyse

Afhankelijke variabele: regelovertredend gedrag	**Model 1(bivariaat) β**	**Model 2 (bivariaat) β**	**Model 3 (multivariaat) β**
Sociale controle ouders (X1)	-0.382*	--	-0.198*
Criminele waarden (X2)	--	0.496*	0.404*
R-kwadraat	**14.6%**	**24.6%**	**27.5%**

* $p < 0.05$ of beter

Beide onafhankelijke variabelen verklaren een substantieel deel van de variantie in crimineel gedrag van jongeren: X1 verklaart 14.6% en X2 verklaart 24.6%. We kunnen hieruit echter niet besluiten dat beide variabelen samen 14.6% + 24.6% van de variatie in Y verklaren. Dit komt omdat de beide variabelen sterk samenhangen. Het kan niet anders of er is dus *gedeelde variantie* tussen X1 en X2, waardoor de uitvoering van twee bivariate analyses misleidend zou geweest zijn.

Bovendien leren we ook hier nog iets anders: de gestandaardiseerde richtingscoëfficiënten (de beta-waarden) zijn in de meervoudige regressieanalyse verschillend van de richtingscoëfficiënten afkomstig uit de bivariate regressieanalyses. Het effect van X1 op Y is kleiner onder controle van X2. Dit leiden we af uit de tabel. We zien dat het rechtstreeks effect van ouderlijke controle op delinquent gedrag groter is in een bivariate analyse dan in de meervoudige regressieanalyse waar we simultaan rekening houden met de invloed van criminele waarden. Het rechtstreeks effect van delinquente waarden is ook iets kleiner wanneer we rekening houden met het effect van ouderlijke controle.

4. De uitbreiding naar een meervoudige regressieanalyse

De bivariate regressievergelijking ziet er als volgt uit:

y= a+bx+e
$\hat{y} = a + bx$

Y is de geobserveerde score op het explanandum, a is het intercept en b is de richtingscoëfficiënt. Tenslotte is e de residuele term.

De multipele regressievergelijking is een uitbreiding van het bivariate model en ziet er als volgt uit:

$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$+e
$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$

Het vinden van de waarden voor de verschillende *b*-waarden vergt veel rekenwerk in de situaties waarin het aantal onafhankelijke variabelen meer dan twee bedraagt (k>2). Een snelle uitrekening van de netto-effecten van de b-waarden in de situatie van drie of meerdere onafhankelijke variabelen gebeurt daarom aan de hand van de matrix algebra. Studenten kunnen gerust zijn: de matrix algebra blijft buiten het bestek van dit handboek. Om inhoudelijk te begrijpen wat er gebeurt in de situatie van twee onafhankelijke variabelen, is echter minder rekenwerk nodig dan men op het eerste zicht zou denken.

5. Het relatieve belang van elke onafhankelijke variabele

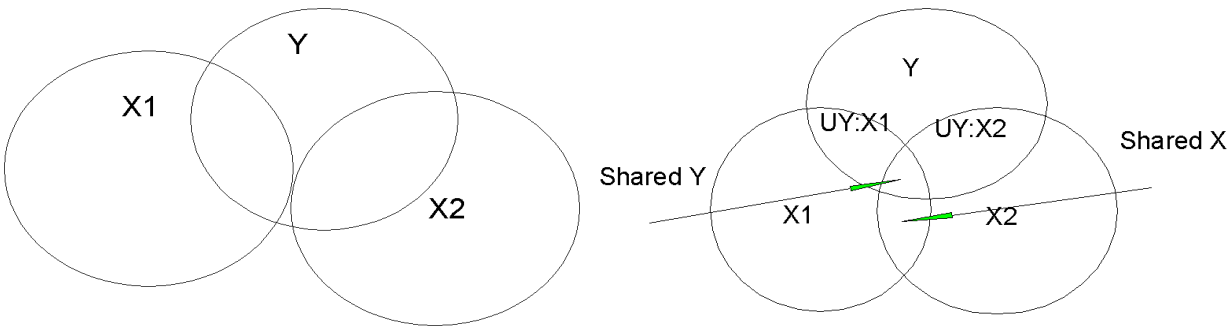
In een eenvoudige regressieanalyse met één afhankelijke variabele is er een afhankelijke variabele en een onafhankelijke variabele. De impact van de onafhankelijke variabele (hoe belangrijk is de onafhankelijke variabele bij de predictie van de afhankelijke variabele) wordt uitgedrukt aan de hand van de verklaarde variantie r^2. Als r^2 1.0 bedraagt, weten we dat de afhankelijke variabele perfect kan voorspeld worden op basis van de onafhankelijke variabele. Als r^2 nul bedraagt, weten we dat er alvast geen lineair verband bestaat tussen de onafhankelijke variabele en de afhankelijke variabele. In de situatie met twee of meerdere onafhankelijke variabelen verkrijgen we ook een verklaarde variantie, de totale R^2. Deze R^2 vertelt ons hoeveel van de variantie in Y kan verklaard worden op basis van de onafhankelijke variabelen samen. Deze totale verklaarde variantie vertelt ons iets over de relatieve belangrijkheid van de lineaire combinatie van de onafhankelijke variabelen ($b_1X_1+b_2X_2+...+b_kX_k$). Heel vaak willen we weten welke impact elke variabele op zichzelf heeft, onder controle van de overige onafhankelijke variabelen.

Opnieuw maken we gebruik van venndiagrammen om uit te leggen wat er gebeurt. **Venndiagrammen mogen misschien niet helemaal de meest correcte manier zijn om de idee achter een multivariate regressieanalyse uit te leggen, maar deze visuele voorstelling zorgt voor een beter begrip van het concept.**

We voorspellen Y op basis van twee onafhankelijke variabelen, X1 en X2. In de ideale situatie zijn X1 en X2 niet met elkaar gecorreleerd en dan krijgen we de situatie zoals deze is getekend in het linkse venndiagram.

Figuur: de relatie tussen de onafhankelijke variabelen

Geen multicollineariteit **Wel multicollineariteit**



De drie cirkels stellen onafhankelijke variabelen voor. Elke cirkel representeert de variabiliteit in de variabele. De grootte van de overlap kan worden afgelezen uit de grootte van de overlappende gebieden. In de situatie die links werd getekend, is er geen overlap tussen X1 en X2, beiden hebben wel elk een impact op Y. In de situatie rechts is er wel overlap tussen X1 en X2 en hier stelt zich hetzelfde probleem als we eerder beschreven bij de uiteenzetting over de partiële correlatie. We presenteren de correlatiematrix tussen de drie variabelen:

Tabel: correlatiematrix tussen de drie variabelen

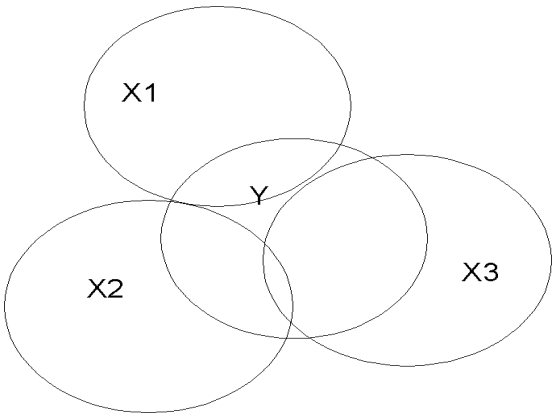
R	Y	X_1	X_2
Y	1		
X_1	0.50	1	
X_2	0.60	0.00	1

In het geval dat X1 en X2 niet gecorreleerd zijn kunnen we de totale verklaarde variantie op basis van X1 en X2 schatten door de afzonderlijke verklaarde varianties op te tellen: in ons voorbeeld is dat $0.50^2+0.60^2 = 0.25+0.36 = 0.61$.

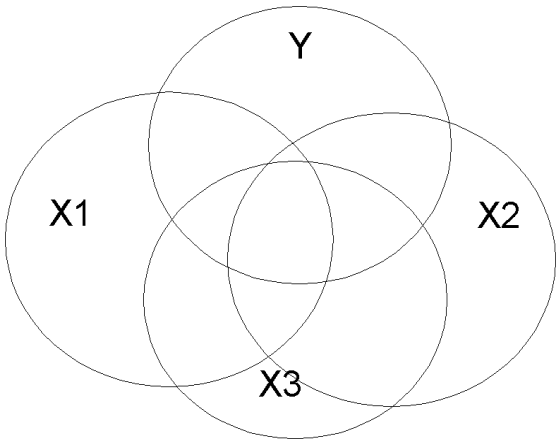
In de meeste gevallen is er echter wel een zekere mate van overlap te verwachten. In de sociale werkelijkheid hangen vele kenmerken nu eenmaal samen

Wanneer we dit idee doortrekken naar de multipele regressie met drie onafhankelijke variabelen dan is dit ook zo:

De quasi ideale situatie



De realiteit



Linksboven zien we een quasi ideale situatie: minimale overlap tussen de verklarende variabelen. Rechtsboven zien we de realiteit: een kluwen van correlerende variabelen. Laat dit de lezer niet afschrikken.

6. De berekening van de gestandaardiseerde gewichten (β1 en β2)

De parameters van een multivariate regressieanalyse kunnen gemakkelijk berekend worden op basis van de correlatiematrix. Dit is goed nieuws. De formule is eenvoudig en we verkrijgen onmiddellijk de gestandaardiseerde richtingscoëfficiënt: dit is tenslotte de coëfficiënt die we het meest nodig hebben om zinvolle uitspraken te kunnen doen over het relatieve belang van elk van de onafhankelijke variabelen. Het probleem met de ongestandaardiseerde regressiegewichten is dat zij elk in hun eigen metrische eenheid worden gepresenteerd en dus elke zinvolle vergelijking in de weg staan. De betekenis van de toename in Y op basis van de toename van X1 met één eenheid wordt uitgedrukt in de meeteenheid van X1. Hetzelfde geldt voor X2.

Wanneer we de gestandaardiseerde richtingscoëfficiënten willen berekenen op basis van de correlatiematrix tussen de onafhankelijke variabelen, moeten we volgende formule gebruiken

$$\beta_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$$ en

$$\beta_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}$$

waarbij r_{y1} de correlatie van y met X1 voorstelt, r_{y2} is dan de correlatie van y met X2, en r_{12} is de correlatie tussen X1 en X2. Let er op dat de twee formules bijna identiek zijn. Het verschil zit hem in de volgorde van de symbolen in de teller.

Als onze correlatiematrix er als volgt uitziet:

	Y	X_1	X_2
Y	1		

X_1	0.77	1	
X_2	0.72	0.68	1

dan bekomen we volgende uitkomsten voor de gestandardiseerde regressiegewichten van X1 en X2:

$$\beta_1 = \frac{.77 - (.72)(.68)}{1 - .68^2} = .521577$$

$$\beta_2 = \frac{.72 - (.77)(.68)}{1 - .68^2} = .365327$$

De berekening van de ongestandaardiseerde regressiecoëfficiënten (b1 en b2)

De berekening van de ongestandaardiseerde regressiegewichten kan eenvoudig door de gestandaardiseerde regressiegewichten te vermenigvuldigen met een coëfficiënt. Voor b1 is dat de standaardafwijking van Y gedeeld door de standaardafwijking van x1. Voor b2 is dat de standaardafwijking van y gedeeld door de standaardafwijking van x2.

$$b_1 = \left(\frac{r_{y,x1} - r_{y,x2} r_{x1,x2}}{1 - (r_{x1,x2})^2} \right) \left(\frac{SD_y}{SD_{x1}} \right)$$

$$b_2 = \left(\frac{r_{y,x2} - r_{y,x1} r_{x1,x2}}{1 - (r_{x1,x2})^2} \right) \left(\frac{SD_y}{SD_{x2}} \right)$$

De berekening van de totale R^2

In de situatie van gecorreleerde onafhankelijke variabelen bekomen we de correcte waarde voor de verklaarde variantie door de volgende formule te hanteren:

$$R^2 = \beta_1 r_{y1} + \beta_2 r_{y2}$$

Als je de logica van de regressieanalyse snapt, dan begrijp je nu ook waarom de determinatiecoëfficiënt gedoemd is om een lagere waarde te hebben dan de optelsom van de twee bivariate correlaties (in geval van overlap tussen de twee onafhankelijke variabelen). Het gaat hier, zoals de formule zegt om het relatieve aandeel van de beide onafhankelijke variabelen: in de formule wordt gebruik gemaakt van de gestandaardiseerde richtingscoëfficiënten en de partiële correlatiecoëfficiënten. Hierdoor is elke gemeenschappelijke overlap verwijderd. Het gaat hem om de optelsom van wat uniek door X1 kan verklaard worden en wat uniek door X2 kan verklaard worden.

En hoe vinden we het intercept? Het intercept kan gevonden worden op dezelfde manier als we eerder hebben aangetoond voor de situatie waarbij we slechts één onafhankelijke variabele hadden.

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

Samengevat: wat hebben we nodig om de meervoudige regressieanalyse te kunnen uitvoeren?

- Gemiddelden van alle variabelen
- Standaardafwijkingen van alle variabelen
- Correlaties tussen alle variabele

7. Veronderstellingen bij het uitvoeren van een lineaire regressie analyse

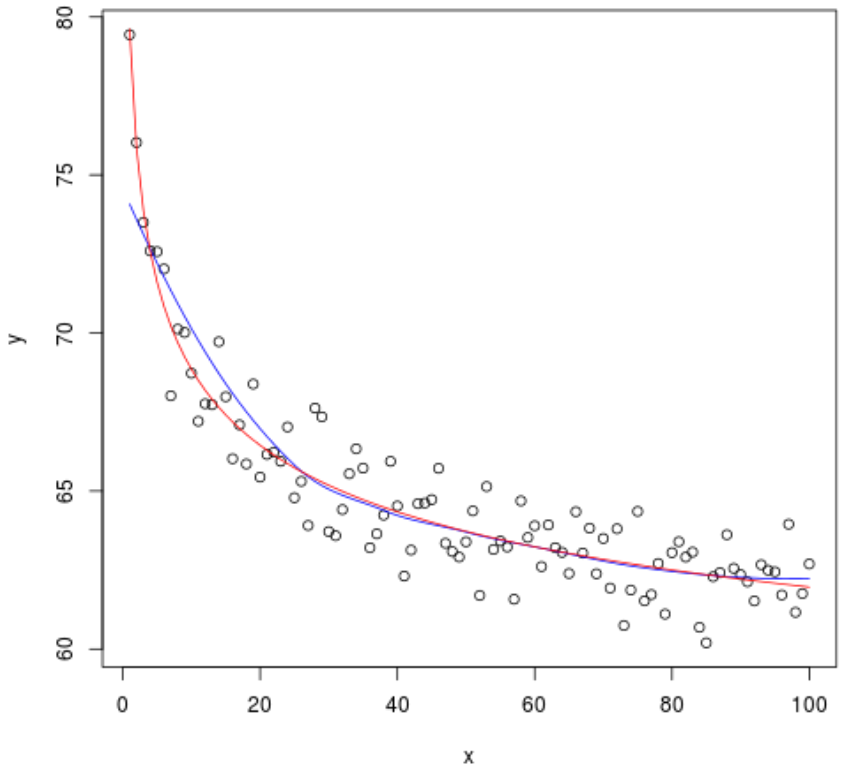
Het is belangrijk stil te staan bij een aantal niet eerder genoemde, maar toch belangrijke veronderstellingen waaraan dient voldaan te worden wanneer we een regressieanalyse uitvoeren.

Ten eerste wordt wat betreft het meetniveau verondersteld dat zowel de afhankelijke als de onafhankelijke variabele gemeten zijn op ***interval- of ratio-meetniveau***.

Een tweede veronderstelling is dat er een *lineair verband* bestaat tussen de onafhankelijke en afhankelijke variabele. Of beter nog, de lineaire regressie levert maar zinvolle informatie op voor zover er sprake is van een lineair verband tussen beide variabelen. Deze veronderstelling kan nagegaan worden door een grafische voorstelling te maken waar de residuele termen

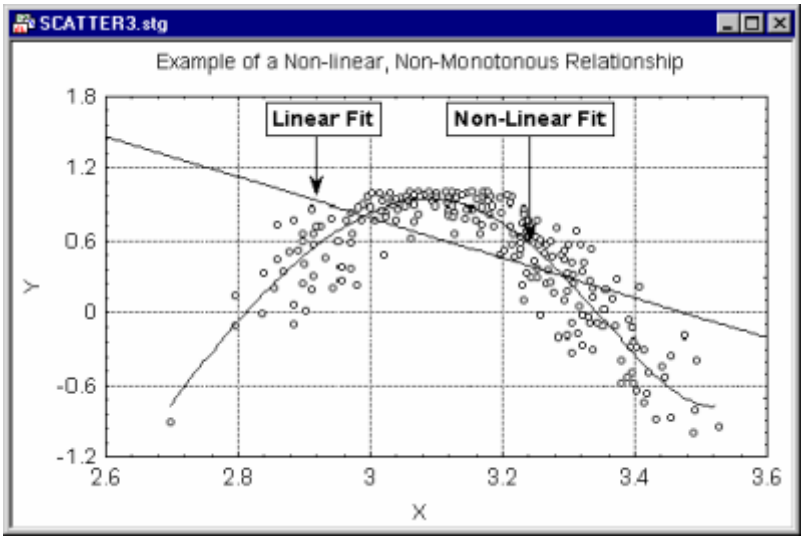
worden uitgezet tegenover de waarden van de onafhankelijke variabele. Indien deze residuen een (niet-lineair) patroon vertonen, is er sprake van schending van de assumptie van lineariteit. Ook het grafisch weergeven (“plotten”) van de afhankelijke variabele tegenover de onafhankelijke variabele kan je hierover al een idee geven.

Voorbeeld een: non-lineariteit



In het eerste voorbeeld is non-lineariteit weliswaar aanwezig, maar het verband tussen x en y is negatief, ongeacht je nu een lineaire associatiemaat of een non-lineaire associatiemaat berekent.

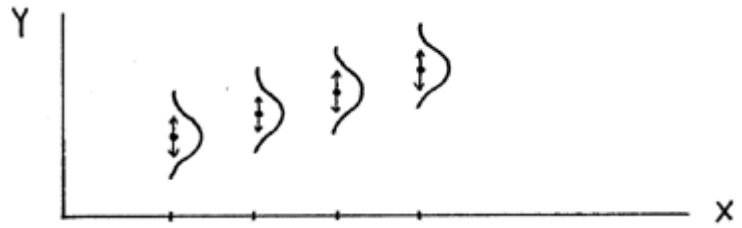
Voorbeeld twee: non-lineariteit



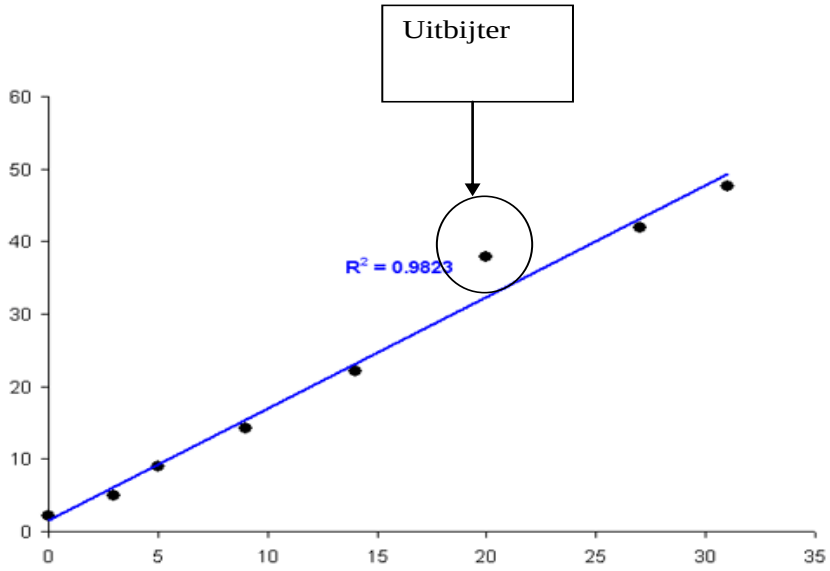
In dit voorbeeld wordt het wel erg duidelijk wat er kan gebeuren als je geen rekening houdt met de nonlineariteit. In dit voorbeeld is het zo dat een lineaire analyse een negatief verband suggereert en een non-lineaire analyse toont aan dat het verband tussen x en y eerst stijgt en dan vanaf een bepaalde waarde terug afneemt.

Ten derde wordt er verondersteld dat voor elke waarde van de onafhankelijke variabele, de afhankelijke variabele een ***normale verdeling*** kent. Bovendien is het noodzakelijk dat de variatie van de afhankelijke variabele voor elke waarde van de onafhankelijke variabele gelijk is. De technische term voor deze veronderstelling is *homoscedasticiteit*. Wanneer deze variaties verschillen naargelang de waarde van de onafhankelijke variabele spreekt men van *heteroscedasticiteit*. Mogelijke schendingen van deze assumpties kunnen eveneens nagegaan worden door het plotten van de residuen tegenover de onafhankelijke variabele. Het omgekeerde van homoscedasticiteit heet heteroscedasticiteit. Dit betekent dat de spreiding van geobserveerde waarden voor Y groter wordt naarmate dat de waarde op de onafhankelijke waarde toeneemt. Of de assumptie van homoscedasticiteit geschonden is, kan onder meer worden nagegaan door een aantal statistische toetsen. Een pionier was *White' s test*.

Figuur: Grafische illustratie van homoscedasticiteit



Ten vierde dienen we ook voorzichtig te zijn met uitbijters of "**outliers**". Dit zijn extreme waarden die de regressierechte beïnvloeden. Een puntenwolk toont je niet enkel de spreiding van observaties in het tweedimensionale vlak, maar als je de best passende lijn opvraagt, dan zie je ook welke punten sterk afwijken. We noemen deze punten uitbijters en men dient er voor waakzaam te zijn. Ze kunnen de relatie tussen twee variabelen ernstig beïnvloeden.



Ten vijfde mag er geen autocorrelatie zijn. **Autocorrelatie** (of **seriële correlatie**) betekent dat de waarden van een variabele gemeten op T2 niet beïnvloed worden door de waarden die een statistische eenheid had op T1. Dit probleem is eigen aan de tijdsreeksanalyse, die tot doel heeft voorspellingen te maken op basis van gegevens die doorheen de tijd werden verzameld. De analyse-eenheid is dan het jaar en de variabelen zijn bijvoorbeeld de werkloosheidsgraad en de criminaliteitsgraad. De criminaliteitsgraad van een eenheid op T1 is niet onafhankelijk

van de criminaliteitsgraad van diezelfde variabele op T2. Als er seriële correlatie is, moeten robuuste analysetechnieken worden gekozen. In de econometrie, de toepassing van de statistiek in het domein van de economie, wordt dit vaak gedaan en bestaan diverse manieren om aan tijdreeksanalyse te doen.

Ten zesde mag er geen statistische interactie zijn. Dit wordt later uitgelegd wanneer we het hebben over de grenzen van de bivariate statistiek. Een ander woord is conditionele causaliteit.

8. Controle op de regressievoorwaarden

In deze paragraaf wordt nagegaan in welke mate er een schending is van de regressievoorwaarden. De controles die zullen besproken worden, hebben betrekking op normaliteit, heteroscedasticiteit, lineariteit, en additiviteit. We zullen eveneens oog hebben voor uitbijters. Autocorrelatie laten we buiten beschouwing omdat dit probleem vooral een rol speelt in tijdreeksanalyses. Het probleem van multicollineariteit wordt niet meer afzonderlijk behandeld omdat we uit wat voorafging weten dat de onderlinge correlaties tussen de latente variabelen in geen geval problematisch zijn.

Normaliteit

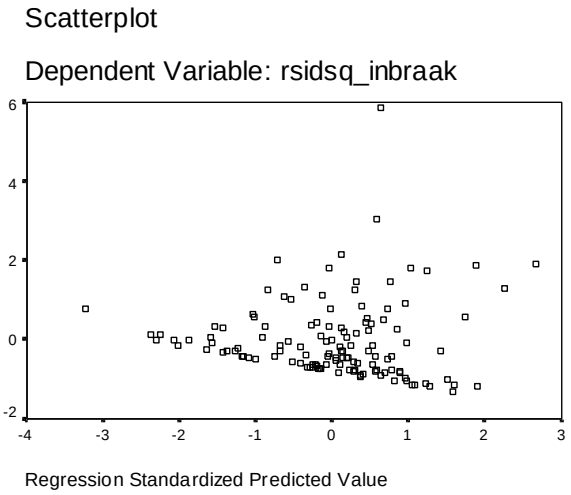
Het gaat hier zowel om univariate als multivariate normaliteit. Zowel de afhankelijke als de onafhankelijke variabelen dienen te voldoen aan deze eis.

Heteroscedasticiteit

Homoscedasticiteit betekent dat de spreiding in de error termen constant is. Wanneer de spreiding van de error termen toeneemt met de voorspelde waarden, komen we technisch gezien in de problemen bij de interpretatie van de significantietoets. In dit geval spreken we van **heteroscedasticiteit** en wordt de standaardfout van de parameters onderschat, met als gevolg dat effecten te snel als significant worden beschouwd.

Het waarnemen van heteroscedasticiteit met het blote oog is niet altijd vanzelfsprekend. Detectie van heteroscedasticiteit is op verschillende manieren mogelijk. Eén van de mogelijkheden om het probleem van de heteroscedasticiteit te onderzoeken, is het gebruik van

White's test. White ontwikkelde deze methode vanuit de econometrie[17]. In het geval van heteroscedasticiteit is OLS-regressie niet de meest adequate manier om populatieparameters te schatten. Het bestaan van heteroscedasticiteit kan eveneens worden afgeleid uit de scatterplot, waarbij de gestandaardiseerde residuele termen worden uitgezet als functie van de voorspelde waarden. We krijgen een patroon te zien van uitwaaiende residuele termen, m.a.w. er is geen random spreiding van de errortermen rond de verwachte waarde van Y. Hoe groter de verwachte waarde van Y, hoe groter het verschil tussen de geobserveerde waarde voor Y en de voorspelde waarde voor Y. Een manier om hiermee om te gaan is het gebruik van WLS-regressie, waarbij men een verschillend gewicht geeft aan de verschillende observaties, doch dit valt buiten het bestek van deze inleidende cursus.



Additiviteit

We vervolgen met een controle op **additiviteit**. We gaan hierbij als volgt te werk: een model wordt opgesteld dat bestaat uit alle mogelijke interactie-effecten en wordt vergeleken met een statistisch model met enkel de hoofdeffecten. Indien een regressiemodel met interactie een significante verbetering teweeg brengt in de verklaarde variantie, is dit een indicatie voor het feit dat niet elke variabele op zich een effect heeft op de criminaliteitsgraad, maar dat de verschillende variabelen elkaar versterken in hun effect op de graad. We spreken dan van conditionaliteit.

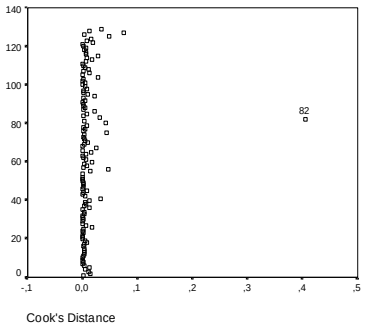
[17] Mc Clendon, J. (1994). *Multiple regression and Causal analysis*. New York: Peacock Publishers.

Lineariteit

Lineariteit betekent dat het verband tussen X en Y lineair is. Wanneer we geen speciaal curvilineair patroon ontwaren, behouden we de lineaire regressie. In het geval er curvilineariteit bestaat, moeten we curvilineaire regressietechnieken gebruiken. Deze vallen buiten het bestek van deze cursus.

Uitbijters of outliers

Tot slot is het belangrijk na te gaan of er in een regressiemodel **uitbijters** te detecteren vallen die meer dan twee standaardafwijkingen boven de gestandaardiseerde verwachte waarde liggen. Uitbijters zijn waarden die onmiddellijk in het oog springen omwille van hun geïsoleerde positie in de grafische weergave van de analyse. Het gaat om observaties met een wel erg groot verschil tussen de geobserveerde waarde en de voorspelde waarde op basis van een onafhankelijke variabele. De aanwezigheid van uitbijters hoeft op zich geen probleem te zijn. Het is perfect denkbaar dat bijvoorbeeld een bepaalde gemeente in een regressieanalyse een eenzame positie heeft, doch dit wil nog niet zeggen dat deze waarde de regressielijn zwaar beïnvloedt. Het probleem doet zich pas voor wanneer deze waarde de regressielijn extreem beïnvloedt. Dit kunnen we nagaan door **Cook's distance** op te vragen in een software programma zoals SPSS. Traditioneel wordt aangenomen dat waarden groter dan 1 een eerder problematisch karakter kennen.



9. De limieten van meervoudige regressie

Er wordt te vaak van uitgegaan dat de meervoudige regressie het probleem van de identificatie van de unieke causale impact van oorzaken overwint en ons in staat stelt hun relatieve belang ten aanzien van diverse uitkomstvariabelen af te wegen. Heel vaak moeten we vaststellen dat ook voorzichtige auteurs de volgende ideeën door elkaar halen:

1. A (een 'variabele', vb. IQ) correleert met B (een andere variabele, vb. inkomen)
2. A 'voorspelt' B
3. A 'verklaart de variantie' in B
4. A 'verklaart' B
5. A 'veroorzaakt' B

We kunnen punten 4 en 5 samen behandelen. We kunnen 'A verklaart B' zeggen, alleen als we kunnen zeggen dat 'A veroorzaakt B'. Maar, **ten eerste**, correlaties staven geen oorzaken. Oorzaken zijn 'mechanismen' die uitkomsten produceren. Zo kunnen we correlaties hebben waarbij er geen mechanisme denkbaar is, vb. tussen de prijs van eieren op een markt in Bejing en de prijs van Microsoft op de New Yorkse aandelenbeurs. **Ten tweede**, er bestaan heel veel oorzaken van een uitkomst. Zo is er, wanneer we vuur willen maken, zowel brandbaar materiaal nodig als een bron van warmte en zuurstof. Als er één van de voornoemde variabelen ontbreekt, kan er geen vuur zijn. Welke is nu meer belangrijk? Wel, we zullen enkel vuur krijgen bij een juiste combinatie. (Om een vinylen stof te laten ontbranden heb je meer warmte nodig dan voor het ontbranden van katoen). Als we een bron van warmte nemen als 'de oorzaak', is dat omdat we ervan uitgaan dat er zuurstof en brandbaar materiaal aanwezig zijn. We vergeten eigenlijk de zuurstof en zeggen dat een vonk het vuur heeft 'veroorzaakt'. Weber noemde dit ***'adequate causation'***, het verschil in de bestaande staat die het effect met zich meebracht. Dit is pragmatisch en niet verrassend, maar het blijft wel een feit dat alle factoren belangrijk zijn: je zal geen vuur hebben als er een factor ontbreekt. Neem nu de mogelijkheid van studenten uit de eerste BAC criminologie om hoog te scoren op statistiek. Wat is 'de oorzaak'? Welke van deze 'factoren' (oorzaken) zijn het belangrijkste? De ene student kan 'slim' zijn, maar de ander kan ook enorm gemotiveerd zijn, goed les gekregen hebben, een andere kan zich goed gevoeld hebben op de dag van de test. Met multivariate regressiemodellen tonen we aan welke effecten het sterkst samenhangen in

steekproeven of in populaties (wanneer we populatiegegevens hebben). Kortom, denken in termen van lineaire en additieve invloeden is niet steeds correct. Er moet gezocht worden naar relevante causale elementen die samen een voldoende oorzaak vormen voor de verklaring van een effect via het in gang zetten van een causaal mechanisme.

10. Testvragen

Hieronder vind je enkele uitspraken over de multivariate statistiek. Deze vragen kan je gebruiken om je parate kennis te toetsen over de multivariate analyse en pad-analyse. **In gewijzigde vorm kunnen dergelijke theorievragen ook op het examen voorkomen.** Deze vragen zijn afkomstig uit vroegere examens.

1. **De meervoudige regressieanalyse met X1 en X2 als onafhankelijke variabelen geeft hetzelfde resultaat als twee afzonderlijke bivariate regressieanalyses. Klopt dat?**
   - Dit klopt enkel in de situatie waarbij r(x1,x2) = 0.
   - Dit klopt nooit

2. **R (hoofdletter) staat in de output van een multipele regressie voor de correlatie tussen Y en de verwachte waarde voor Y op basis van de onafhankelijke variabelen**
   - Deze uitspraak is juist
   - Deze uitspraak is verkeerd

3. **Een meervoudige multivariate regressieanalyse is een**
   - Regressieanalyse met meerdere onafhankelijke variabelen en meerdere afhankelijke variabelen
   - Regressieanalyse met meerdere onafhankelijke variabelen en slechts een afhankelijke variabele
   - Regressieanalyse met meerdere afhankelijke variabele en een onafhankelijke variabele

4. **Heteroscedasciticiteit wil zeggen dat**
    - De waarde van de residuele termen toenemen naarmate X1 toeneemt
    - De waarde van de residuele termen gelijk blijft, naarmate X1 toeneemt

5. **Additiviteit betekent dat X1 en x2 onafhankelijke effecten hebben, ttz ze dragen elk bij tot de verklaring van Y**
    - Deze uitspraak is juist
    - Deze uitspraak is verkeerd

6. **Lineariteit wil zeggen dat Y als lineaire functie van X kan worden uitgedrukt**
    - Deze uitspraak is juist
    - Deze uitspraak is verkeerd

7. **Curvilineariteit wil zeggen dat het effect van X1 op Y toeneemt of afneemt naargelang X1 toeneemt.**
    - Deze uitspraak is juist
    - Deze uitspraak is verkeerd

8. **Een uitbijter kan de regressielijn ernstig beïnvloeden**
    - Deze uitspraak is juist
    - Deze uitspraak is verkeerd

9. **Interactie betekent dat het effect van X1 op Y conditioneel is op X2**
    - Deze uitspraak is juist
    - Deze uitspraak is verkeerd

10. Een mediatorvariabele is een variabele die het effect van een exogene variabele op een afhankelijke variabele medieert (dwz dat deze variabele het effect van de exogene variabele wegverklaart en dat het effect van de exogene variabele op Y via de mediator verloopt)

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

11. Een totaal effect is in een pad-analyse gelijk aan de rechtstreekse effecten min de onrechtstreekse effecten

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

11. Leerdoelen

De doelstelling van dit hoofdstuk ligt in het begrijpen en kunnen toepassen van de meervoudige regressieanalyse. Studenten dienen de meervoudige regressieanalyse zelf te kunnen toepassen. Deze inleiding is beperkt in die zin dat we enkel voorbeelden hebben gegeven van meervoudige regressieanalyses met twee onafhankelijke variabelen. Dit betekent dat studenten de parameters van de meervoudige regressieanalyse in de situatie met twee onafhankelijke variabelen zelf moeten kunnen uitrekenen. Het is cruciaal dat de studenten snappen waarom we niet zomaar twee afzonderlijke bivariate analyses bij elkaar kunnen optellen. De meervoudige regressieanalyse is echter een techniek die heel wat eisen stelt aan de data. Deze eisen zijn velerlei en werden slechts partieel besproken. Wat belangrijk is, is dat de lineaire OLS-regressieanalyse niet steeds bruikbaar is. In het geval van ernstige schendingen van de veronderstellingen (heteroscedasticiteit, normaliteit, lineariteit,…) moeten andere methoden gekozen worden. Deze andere methoden zijn ook regressieanalyses, maar deze zijn niet gebaseerd op het OLS-principe. Echter, de inhoudelijke interpretatie van de regressiecoëfficiënten is analoog. Deze andere varianten komen in dit handboek niet aan bod, maar wel in vervolgcursussen. Voorbeelden zijn de logistische regressie, de Poisson-regressie en de negatief-binomiaal regressie. Deze technieken stellen minder strenge eisen aan de structuur van de data. Zo heeft de scheefheid van de verdeling minder een invloed op de berekening van de parameters en de significantie van deze parameters in de laatst genoemde technieken. Tot slot zijn we heel concreet ingegaan op de statistische interactie door een

voorbeeld te geven van een analyse waarbij het verband tussen twee metrische kenmerken conditioneel is op een derde kenmerk.