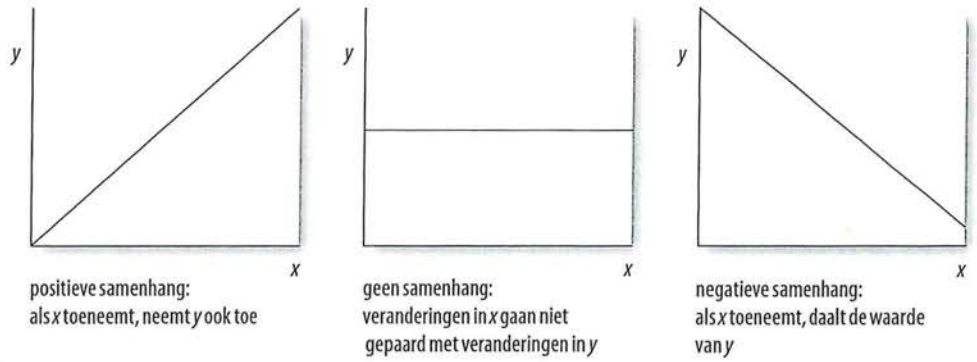
Associatiematen op interval- en rationiveau

8

In dit hoofdstuk staan de associatiematen op interval- en rationiveau centraal. We behandelen zowel bivariate analyses (correlatie en enkelvoudige regressie) als een multivariate analyse (meervoudige regressie).

In hoofdstuk 6 zagen we dat associatiematen voor variabelen die op minimaal ordinaal niveau zijn gemeten een waarde kunnen aannemen die varieert tussen –1 (perfecte negatieve samenhang) en +1 (perfecte positieve samenhang). Dit is ook het geval bij associatiematen voor variabelen die op interval- of rationiveau zijn gemeten. Er is ook op dit niveau een ordening in de waarden, waardoor je zowel een stijgende als een dalende lijn in de samenhang tussen twee variabelen kunt onderscheiden (zie figuur 8.1).



Figuur 8.1 Grafische weergave van samenhang

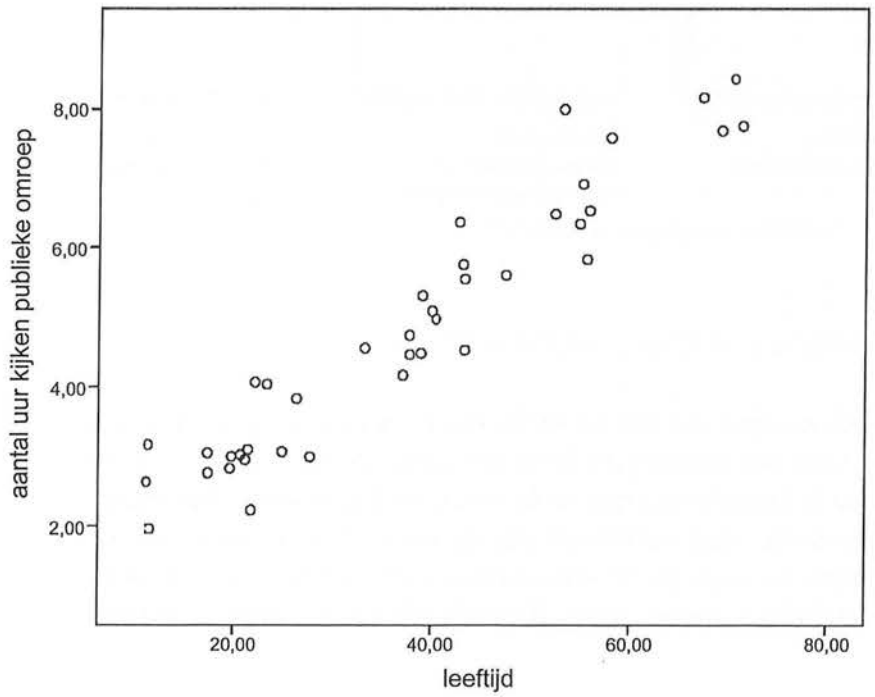
8.1 Pearsons correlatiecoëfficiënt

In hoofdstuk 6 zagen we dat Kendalls tau-b en gamma geschikte associatiematen zijn voor een symmetrische samenhang op ordinaal niveau, en dat we Kendalls tau-b kunnen weergeven in een correlatiematrix. Als maat voor de samenhang tussen twee variabelen die op interval- of rationiveau zijn gemeten, gebruiken we vaak de correlatiecoëfficiënt *r*. Deze correlatiecoëfficiënt is een symmetrische associatiemaat. Bij de berekening is de (on)afhankelijkheid van de variabelen niet van belang. De correlatiecoëfficiënt gebruik je alleen bij variabelen die beide op interval- of rationiveau zijn gemeten. Voluit heet de associatiemaat *Pearson productmoment correlatiecoëfficiënt*. Deze correlatiecoefficiënt duid je aan met de letter *r*.

8.1.1 Grafische weergave

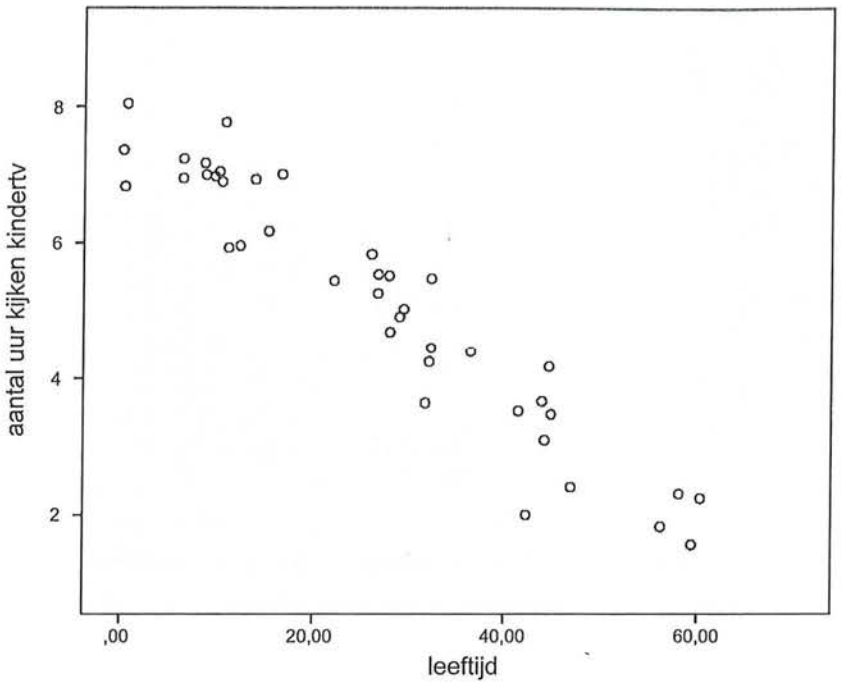
Een samenhang of correlatie is vaak zichtbaar in een *spreidingsdiagram* (*scatterplot*). We zagen deze al kort in de bespreking van Spearmans rho in paragraaf 6.5.1. Een spreidingsdiagram is een puntenwolk die is gebaseerd op de waarnemingen van twee variabelen. Als er sprake is van een onafhankelijke variabele, kiezen we voor deze onafhankelijke variabele de horizontale as, de x-as, en voor de afhankelijke variabele de verticale as, de y-as. Wanneer er geen sprake is van een onafhankelijke en afhankelijke variabele, maakt het niet uit welke variabele je op de x-as en welke je op de y-as plaatst. De waarden die een onderzoekseenheid op de twee variabelen heeft, bepaalt de positie van die onderzoekseenheid binnen dit assenstelsel. Alle onderzoekseenheden samen vormen een puntenwolk.

Wanneer je wilt weten wat de doelgroep van een omroep zou kunnen zijn, zou je bijvoorbeeld kunnen kijken naar de samenhang tussen leeftijd en het aantal uur dat (per week) naar die omroep wordt gekeken.[1] In figuur 8.2 is een spreidingsdiagram gegeven waarin de samenhang is te zien tussen leeftijd en het aantal uur dat iemand per week naar de publieke omroep kijkt. Uit dit spreidingsdiagram blijkt dat er sprake is van een positieve samenhang (positieve correlatie); bij toename van de leeftijd neemt ook de kijktijd naar de publieke omroep toe.
Op basis van deze grafiek zouden we al een voorzichtige conclusie kunnen trekken: naarmate mensen ouder zijn, wordt vaker naar de publieke omroep gekeken. Dit is wellicht voor de publieke omroep aanleiding om voornamelijk programma's voor ouderen uit te zenden.



Figuur 8.2 Spreidingsdiagram van leeftijd en aantal uur naar publieke omroep kijken (SPSS-output)

Op dezelfde manier zouden we kunnen kijken naar de samenhang tussen leeftijd en het kijken naar kindertelevisie. Daar verwachten we eerder een negatieve samenhang; dit is immers gericht op jongeren.

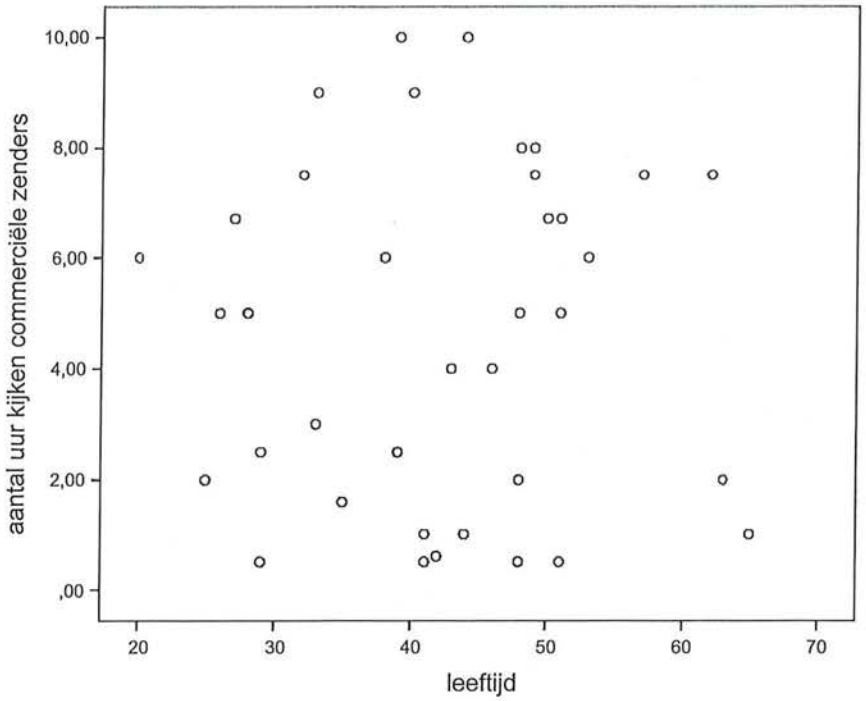


Figuur 8.3 Spreidingsdiagram tussen leeftijd en uren kijken naar kindertelevisie (SPSS-output)

Er blijkt (figuur 8.3) inderdaad een negatieve samenhang uit het spreidingsdiagram. Ouderen kijken minder naar kindertelevisie dan jongeren.

Tot slot een voorbeeld van een spreidingsdiagram waarin de samenhang minder duidelijk is (figuur 8.4). We kijken naar de samenhang tussen leeftijd en het aantal uur dat mensen naar commerciële zenders kijken. We zien dat de puntenwolk meer verdeeld is over alle leeftijden. In dit spreidingsdiagram zien we dus geen samenhang tussen leeftijd en het kijken naar commerciële zenders.

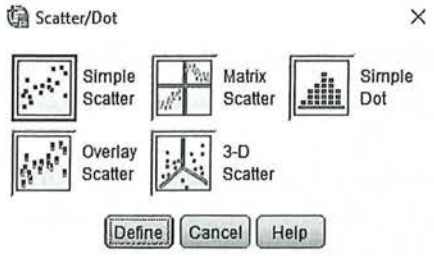
Sommige ouderen kijken veel, anderen weinig, en hetzelfde geldt voor jongeren. We verwachten op basis van dit diagram dus dat de waarde van de maat voor samenhang laag (dat wil zeggen: dicht bij het nulpunt) zal zijn.



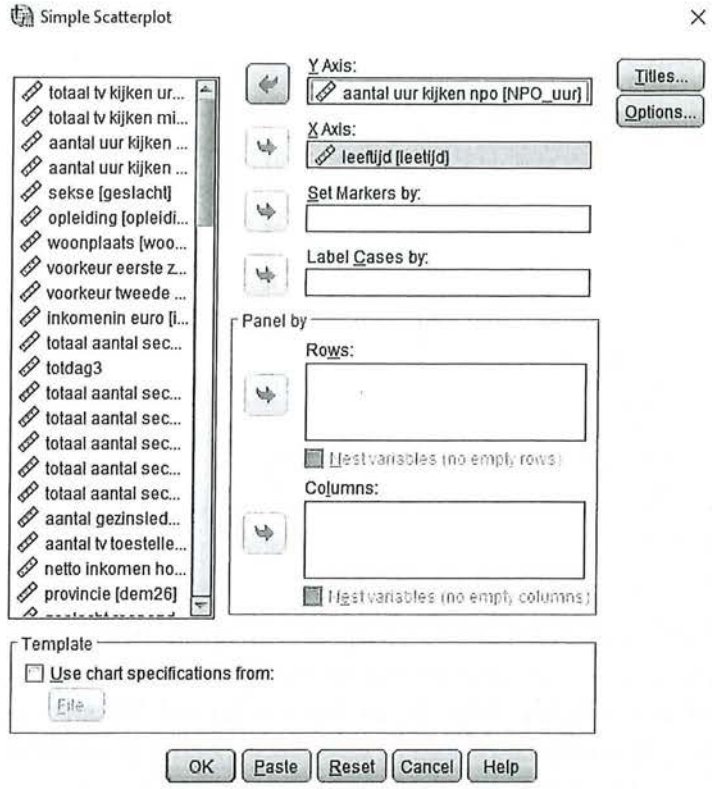
Figuur 8.4 Spreidingsdiagram tussen leeftijd en uren kijken naar commerciële zenders (SPSS-output)

SPSS Het maken van een spreidingsdiagram

Om een spreidingsdiagram te maken zoek je in de menubalk naar *Graphs* en ga je naar *Legacy Dialogs*. Vervolgens kies je in het *Scatter/Dot*-venster (figuur A) voor het eenvoudige spreidingsdiagram (*Simple Scatter*).
In het *Simple Scatterplot* (figuur B) kunnen we dan de variabelen kiezen die we op de x-as en de y-as van het spreidingsdiagram willen hebben. Zet de onafhankelijke variabele op de x-as en de afhankelijke variabele op de y-as.



Figuur A Scatter/Dot-venster



Figuur B Simple Scatterplot-venster

Kader 8.1

8.1.2 *Interpretatie*

Een spreidingsdiagram geeft een eerste indruk van de sterkte en de richting van de samenhang, de waarde van een correlatiecoëfficiënt geeft meer precieze informatie. We bekijken de correlatiecoëfficiënt van het eerste voorbeeld: de samenhang tussen leeftijd en het aantal uur per week dat mensen naar de publieke omroep kijken.

De correlatiecoëfficiënt tussen leeftijd en uren kijken naar de publieke omroep is 0,955 ($r = 0{,}96$). Tabel 8.1 geeft deze waarde tweemaal; één keer voor de samenhang tussen leeftijd en uren kijken naar de publieke omroep en één keer voor de samenhang tussen uren kijken naar de publieke omroep en leeftijd. Uiteraard is deze samenhang hetzelfde, want de berekening van de correlatiecoëfficiënt gaat uit van een symmetrische relatie tussen de twee variabelen.

Tabel 8.1 Correlatie tussen leeftijd en aantal uur kijken naar publieke omroep (SPSS-output)

Correlations

		leeftijd	NPO_uur aantal uur kijken publieke omroep
leeftijd	Pearson Correlation	1	,955**
	Sig. (2-tailed)		,000
	N	40	40
NPO_uur aantal uur kijken publieke omroep	Pearson Correlation	,955**	1
	Sig. (2-tailed)	,000	
	N	40	40

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

De interpretatie is gelijk aan de interpretatie van de overige associatiematen. We zien hier dus een zeer sterke, positieve samenhang tussen leeftijd en uren kijken naar de publieke omroep. Met de leeftijd neemt het kijken naar de publieke omroep toe. De interpretatie van de correlatiecoëfficiënt r is hetzelfde als de interpretatie van de associatiematen op ordinaal niveau.
De correlatiecoëfficiënt kan een waarde aannemen die ligt tussen –1 en +1:

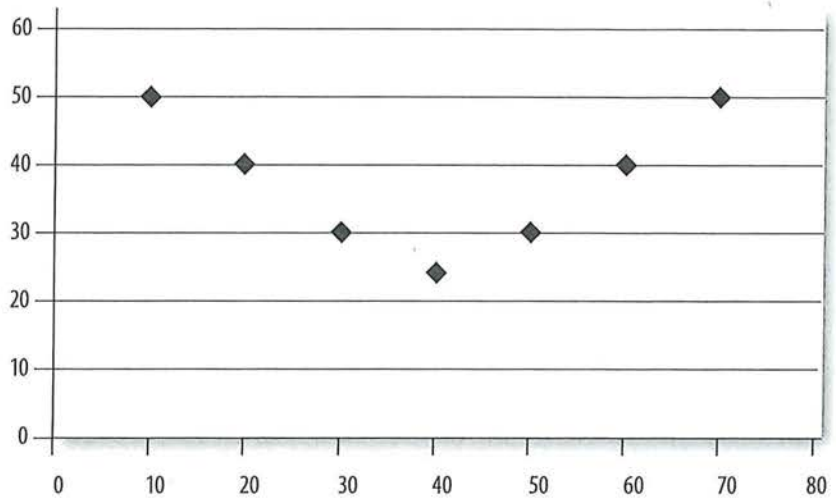
- Correlatiecoëfficiënt = 1: de twee variabelen hangen perfect samen, er is een positief verband, een stijgende lijn (als x stijgt, stijgt y ook, en andersom).
- Correlatiecoëfficiënt = 0: er is geen lineaire (= rechtlijnige) samenhang tussen de twee variabelen.
- Correlatiecoëfficiënt = –1: er is een perfect negatieve samenhang tussen de twee variabelen, er is een dalende lijn (als x stijgt dan daalt y, en andersom).

In dit geval concluderen we dus:

> *Er is een zeer sterke positieve samenhang ($r = 0{,}96$, $n = 40$) tussen leeftijd en het aantal uur dat naar de publieke omroep wordt gekeken. Hoe ouder mensen zijn, hoe vaker ze naar de publieke omroep kijken.*

In de vorige hoofdstukken keken we altijd naar de kruistabel om een indruk van de samenhang te krijgen. Als het gaat om interval- of ratiovariabelen is het verstandig om altijd even naar het spreidingsdiagram te kijken. Daarmee krijg je niet alleen een indruk van de samenhang, maar kun je ook voorkomen dat je op basis van de waarde van r denkt dat er geen samenhang is, terwijl dat wel het geval is. De correlatiecoëfficiënt r geeft namelijk alleen aan in welke mate er sprake is van lineaire samenhang. Als het verband kromlijnig is, is dat niet te zien aan de waarde van r (zie figuur 8.5). Op basis van de waarde van r ($r = 0$) zou je kunnen concluderen dat er geen samenhang bestaat tussen de twee variabelen uit figuur 8.5. Er is inderdaad geen lineair verband, maar uit

het spreidingsdiagram blijkt dat er wel een kromlijnig verband is. Dit zou je gemist hebben als je je alleen op de waarde van r had gebaseerd. De conclusie dat er geen samenhang is, is hier onjuist.



Figuur 8.5 Voorbeeld van een kromlijnig verband

We hebben al gezien in paragraaf 6.5.1 dat een kromlijnig verband ook minder 'krom' kan zijn dan in figuur 8.5 te zien is. Wanneer er veel extreme waarden zijn en daardoor de variabelen niet normaal verdeeld zijn, ontstaat er een sterke kromming die we ook kromlijnig noemen. In dat geval moet je niet de correlatiecoëfficiënt r berekenen maar is het beter Spearmans rho te gebruiken.

8.1.3 *Berekening*

De basis voor het berekenen van de correlatie is de covariantie en de standaarddeviatie. In hoofdstuk 3 hebben we gezien hoe je de standaarddeviatie berekent. De standaarddeviatie is een soort gemiddelde afstand ten opzichte van het gemiddelde. De variantie is het kwadraat van de standaarddeviatie (oftewel: de standaarddeviatie is de wortel uit de variantie). In hoofdstuk 3 keken we naar de variantie en de standaarddeviatie *binnen* één variabele (bijvoorbeeld in hoeverre de leeftijden van de onderzoekseenheden afweken van het gemiddelde). Het is ook mogelijk om te kijken naar de variantie *tussen* twee variabelen (tussen x en y). Deze vorm van variantie noem je de *covariantie.* De covariantie geeft de mate aan waarin twee variabelen tegelijk variëren. Wanneer er een positieve covariantie is, zullen twee variabelen positief met elkaar correleren. Als de onderzoekseenheden op de ene variabele hoog scoren, zullen ze dat op de andere variabele ook doen. Omgekeerd: wanneer er een negatieve covariantie is, zullen twee variabelen negatief met elkaar correleren. Dit betekent dat onderzoekseenheden die op de ene variabele hoog scoren, op de andere juist laag scoren.

Wanneer je de covariantie tussen de variabelen x en y berekent, noteer je dit als volgt: *Cov (x, y)*.

$$Cov(x,y)=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

Formule covariantie

Zowel de covariantie tussen x en y als de afzonderlijke standaarddeviaties van x en y komen terug in de formule voor de correlatie.

$$r_{xy}=\frac{Cov(x,y)}{s_x s_y}$$

Formule voor correlatie

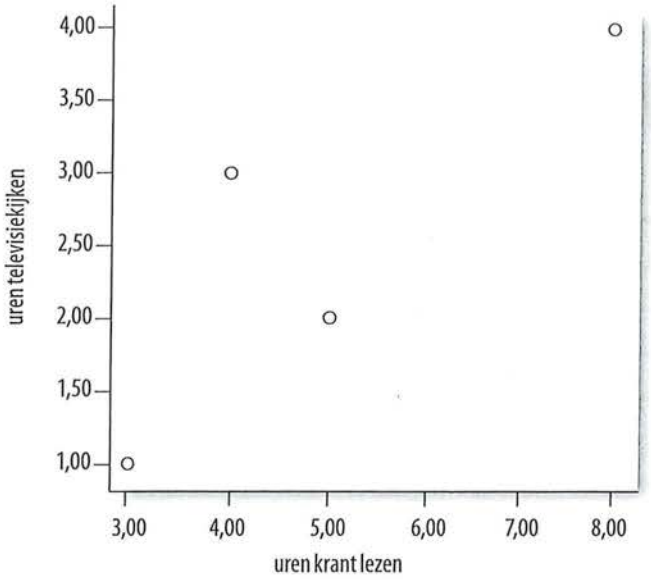
De letter r gebruiken we als symbool voor deze correlatie. De x en de y achter de r geven aan dat het om een correlatie tussen x en y gaat.

De formule van de covariantie lijkt erg op de formule voor variantie, maar nu wordt niet naar één maar naar twee variabelen gekeken (x en y). De basisingrediënten zijn, zoals te zien in de formule, de afwijkingen van x van het gemiddelde van x $(x-\overline{x})$ en de afwijking van y van het gemiddelde van y $(y-\overline{y})$. In de formule van r wordt de covariantie gedeeld door het product van de standaarddeviaties van de twee variabelen. Deze standaarddeviaties zijn zelf ook weer gebaseerd op afwijkingen van het gemiddelde. Dat is dan ook altijd stap 1: het berekenen van de gemiddeldes. Daarna kan de standaarddeviatie van zowel x (s_x) als y (s_y) worden uitgerekend en de covariantie. We zullen dit laten zien aan de hand van een voorbeeld.

We willen nagaan of mensen die vaak de krant lezen ook vaak naar het televisienieuws kijken, en andersom. Dit doen we op basis van de gegevens in tabel 8.2 (krant lezen en televisienieuws kijken is gemeten in uren per week).
Aan het spreidingsdiagram (figuur 8.6) is al te zien dat er een positieve samenhang is tussen het lezen van de krant en televisiekijken.

Tabel 8.2 Datamatrix uren krant en uren tv voor vier respondenten

Respondent	Uren krant (x)	Uren tv (y)
A	3	1
B	4	3
C	5	2
D	8	4



Figuur 8.6 Spreidingsdiagram tussen het lezen van de krant en televisiekijken

Eerst berekenen we het gemiddelde van x en het gemiddelde van y:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+4+5+8}{4} = \frac{20}{4} = 5$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1+3+2+4}{4} = \frac{10}{4} = 2{,}5$$

Gemiddeld lezen deze vier mensen vijf uur per week de krant, en kijken ze 2,5 uur per week televisie. Vervolgens berekenen we voor zowel het aantal uur krant (x) als het aantal uur televisie (y) de standaarddeviatie (die hebben we immers nodig om de formule van r in te vullen).
Om de standaarddeviatie te berekenen, moeten we eerst de variatie van x (de kwadratensom van het verschil tussen de afzonderlijke x'en en het gemiddelde van x) en de variatie van y (de kwadratensom van het verschil tussen de afzonderlijke y's en het gemiddelde van y) berekenen. Dit doen we op dezelfde manier als al eerder aan bod kwam in hoofdstuk 3. De M wordt in de tabel (tabel 8.3) gebruikt om de gemiddeldes (Mean) aan te geven.

Tabel 8.3 Berekenen van de variatie van x en y

	x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
A	3	1	$(3-5) = -2$	$-2^2 = 4$	$(1-2{,}5) = -1{,}5$	$-1{,}5^2 = 2{,}25$
B	4	3	$(4-5) = -1$	$-1^2 = 1$	$(3-2{,}5) = 0{,}5$	$0{,}5^2 = 0{,}25$
C	5	2	$(5-5) = 0$	$0^2 = 0$	$(2-2{,}5) = -0{,}5$	$-0{,}5^2 = 0{,}25$
D	8	4	$(8-5) = 3$	$3^2 = 9$	$(4-2{,}5) = 1{,}5$	$1{,}5^2 = 2{,}25$
Σ	20	10	0	14	0	5
M	5	2,5				

We kunnen nu de standaarddeviaties van beide variabelen berekenen. Daarvoor delen we de variaties door $n - 1$. Vervolgens trekken we daaruit de wortel (wanneer we niet zouden worteltrekken zouden we de variantie berekend hebben).

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{14}{3}} = 2{,}160$$

$$s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}} = \sqrt{\frac{5}{3}} = 1{,}291$$

We hebben nu de gegevens voor de noemer van de formule voor r. We hoeven alleen nog maar de covariantie te berekenen. Om deze uit te rekenen hebben we al veel werk verricht. Wat we nog wel moeten berekenen is $\sum(x-\bar{x})(y-\bar{y})$ voor de teller van de formule voor covariantie. We moeten dus voor elke onderzoekseenheid $(x-\bar{x})$ vermenigvuldigen met $(y-\bar{y})$, en deze producten vervolgens bij elkaar optellen.
Aan het sigmateken kunnen we zien dat we dit product voor elke respondent moeten uitrekenen en pas daarna over alle respondenten sommeren:

Hoe je dit op een systematische wijze kunt uitwerken, is te zien in tabel 8.4 (waarvan alleen de laatste kolom nieuwe informatie geeft, de eerdere kolommen bevatten berekeningen die we al in tabel 8.3 hebben gedaan).

Tabel 8.4 Berekenen van de covariantie

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
A	3	1	–2	–1,5	–2 * – 1,5 = 3
B	4	3	–1	0,5	–1 * 0,5 = –0,5
C	5	2	0	–0,5	0 * – 0,5 = 0
D	8	4	3	1,5	3 * 1,5 = –4,5
Σ	20	10	0	0	7
M	5	2,5			

Respondent A scoorde 2 uur minder dan het gemiddeld aantal uur krant lezen, 1,5 uur minder dan het gemiddeld aantal uur televisiekijken, en scoort daarom op $(x-\bar{x})(y-\bar{y}) = -2 * -1,5 = 3$. Dit doen we voor elk van de onderzoekseenheden, vervolgens tellen we deze sores bij elkaar op. De som van de producten is 7. We kunnen nu de rest van de formule van de covariantie invullen.

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{7}{3} = 2,333$$

We hebben nu alle gegevens die nodig zijn om de formule van r in te vullen.

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y} = \frac{2,333}{2,160*1,291} = 0,837$$

Oftewel: er is een zeer sterke positieve samenhang tussen het aantal uur dat iemand de krant leest en het aantal uur dat hij televisiekijkt. Wanneer het aantal uren dat iemand de krant leest toeneemt, neemt ook het aantal uren dat hij naar de televisie kijkt toe, en andersom.
SPSS laat zien dat onze berekening klopt.

Tabel 8.5 Correlatie tussen uur krant en uur tv (SPSS-output)

Correlations

		uurkrant	uurtv
uurkrant	Pearson Correlation	1	,837
	Sig. (2-tailed)		,163
	N	4	4
uurtv	Pearson Correlation	,837	1
	Sig. (2-tailed)	,163	
	N	4	4

Door de formule voor de covariantie in de formule van r in te vullen kun je r ook meer direct berekenen.

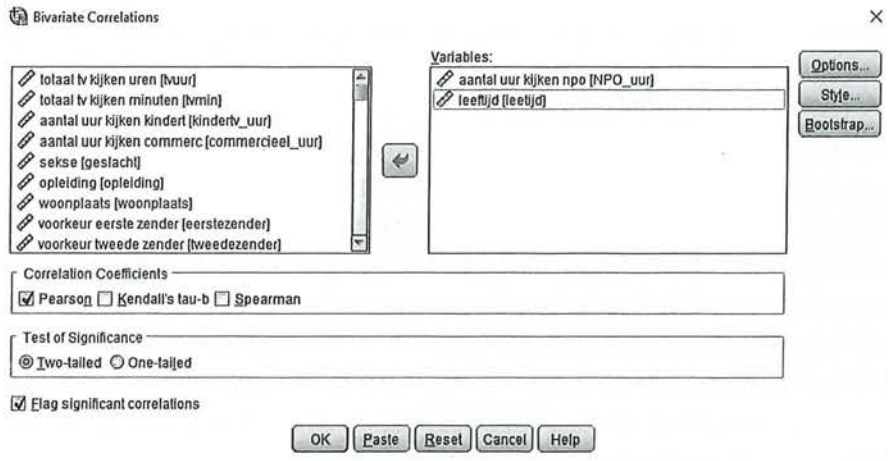
$$r_{xy} = \frac{Cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \div s_x s_y = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Dus je kunt ook de volgende formule voor r gebruiken:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

SPSS Het berekenen van de correlatie

Het berekenen van een correlatie in SPSS gaat via *Analyze* → *Correlate* → *Bivariate*. In het vak *Variables* worden de variabelen ingevuld die voor het berekenen van de correlatie nodig zijn. De *Pearson* correlatie staat automatisch aangevinkt.
NB: Via *Bivariate Correlations* kun je SPSS ook Kendalls tau-b en Spearmans rho laten uitrekenen. Je hoeft daarvoor alleen het desbetreffende hokje aan te vinken.



Figuur A Bivariate Correlations-venster: Pearson

Kader 8.2

8.1.4 *Partiële correlaties*

In hoofdstuk 7 zagen we al dat een samenhang tussen twee variabelen beïnvloed kan worden door het toevoegen van een derde variabele. Bij een tabelsplitsing voegde je een derde variabele toe in de *Layers* waardoor partiële associaties werden berekend per categorie van de derde variabele. Bij interval- en ratiovariabelen is het ook mogelijk om de bivariate samenhang te controleren voor een

derde variabele. Anders dan bij de associatiematen op nominaal of ordinaal niveau krijg je nu niet per waarde van de derde variabele een samenhangsmaat (interval- of ratiovariabelen hebben immers vaak erg veel waarden, daardoor zou je door de bomen het bos niet meer kunnen zien), maar wordt een nieuwe correlatie berekend tussen de twee variabelen terwijl je de derde variabele constant houdt.

We leggen dit uit aan de hand van het volgende voorbeeld. Stel je voor dat je onderzoek onder toneelacteurs doet naar het verband tussen plankenkoorts en tekstvastheid. Je vindt tussen de twee variabelen een sterke negatieve correlatie: $r = -0{,}54$ (tabel 8.6). Hoe meer plankenkoorts hoe minder tekstvast de acteurs zijn en andersom (hoe meer tekstvast de acteurs zijn, hoe minder plankenkoorts ze hebben). Vervolgens wil je kijken of een derde variabele een rol speelt in dit verband, en daarom ga je het verband nogmaals onderzoeken, maar nu onder constanthouding van de variabele 'concentratie'. Wanneer je nu eerst tussen deze drie variabelen bivariate correlaties berekent ziet dat er als volgt uit:

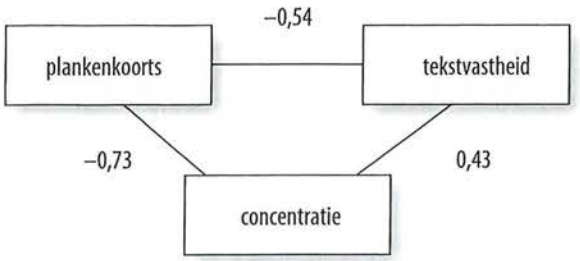
Tabel 8.6 Correlaties van tekstvastheid, plankenkoorts en concentratie (SPSS-output)

Correlations

		Tekstvastheid	Plankenkoorts	Concentratie
Tekstvastheid	Pearson Correlation	1	-,541**	,427**
	Sig. (2-tailed)		,000	,000
	N	103	103	103
Plankenkoorts	Pearson Correlation	-,541**	1	-,729**
	Sig. (2-tailed)	,000		,000
	N	103	103	103
Concentratie	Pearson Correlation	,427**	-,729**	1
	Sig. (2-tailed)	,000	,000	
	N	103	103	103

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

We zien dat concentratie een positieve correlatie heeft met tekstvastheid ($r = 0{,}43$) (dus hoe meer concentratie hoe meer tekstvast de acteur is en omgekeerd) en een negatieve correlatie met plankenkoorts: $r = -0{,}73$ (hoe meer concentratie, hoe minder plankenkoorts en omgekeerd). We kunnen dat nu in een *conceptueel model* tekenen, met concentratie als de mediator:



Figuur 8.7 Conceptueel model met concentratie als mediator

Vervolgens laten we SPSS de correlatie tussen plankenkoorts en tekstvastheid opnieuw berekenen, maar nu controleren we voor de variabele concentratie:

Tabel 8.7 Correlatie plankenkoorts en tekstvastheid waarbij gecontroleerd wordt voor concentratie (SPSS-output)

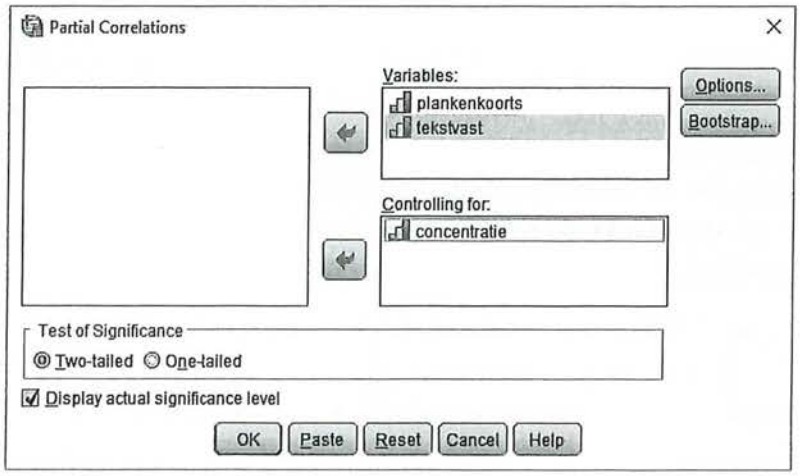
Correlations

Control Variables			Tekstvastheid	Plankenkoorts
Concentratie	Tekstvastheid	Correlation	1,000	-,247
		Significance (2-tailed)	.	,012
		df	0	100
	Plankenkoorts	Correlation	-,247	1,000
		Significance (2-tailed)	,012	.
		df	100	0

Waar de correlatie tussen plankenkoorts en tekstvastheid in eerste instantie −0,54 was, blijkt het verband minder sterk wanneer er gecontroleerd wordt voor de variabele concentratie ($r = -0{,}25$). Dat betekent dat het verband tussen de twee variabelen gedeeltelijk afhangt van de correlatie met concentratie. Er is dus sprake van mediatie. Wanneer het verband geheel zou verdwijnen zou hier sprake zijn van een spurieus verband tussen plankenkoorts en tekstvastheid.

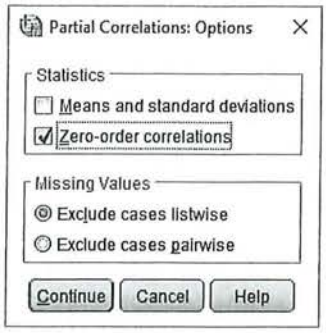
SPSS Het berekenen van partiële correlaties

Om te controleren voor een derde variabele in een correlatieanalyse ga je via *Analyze → Correlate* naar *Partial Correlation*. Bij *Variables* kun je de variabelen invoeren waar je in eerste instantie een correlatie tussen wilt berekenen, bij *Controlling for* zet je de variabele waarvoor je controleert.



Figuur A Partial Correlation-venster

Wanneer je naast de nieuwe, partiële correlatie, de correlaties wilt zien zonder dat gecontroleerd wordt voor een derde variabele, kun je onder *options* het vakje *Zero-order correlations* aanvinken (*Figuur B*).



Figuur B Options-venster

Kader 8.3

8.2 Enkelvoudige regressie

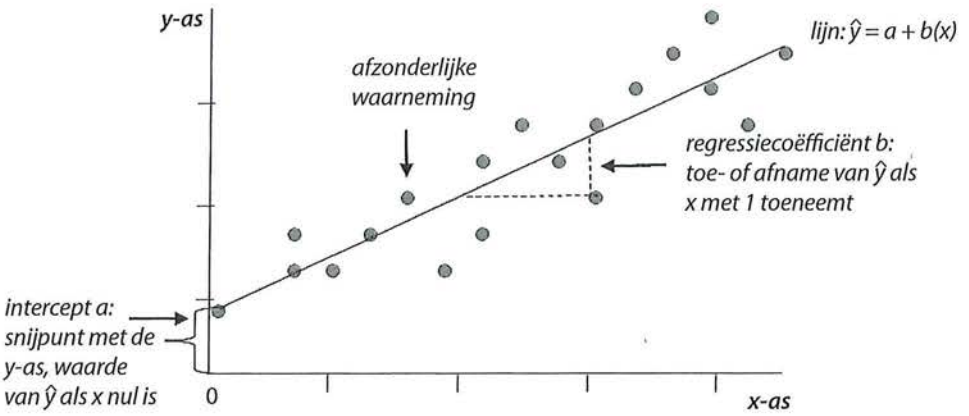
We hebben gezien dat de correlatiecoëfficiënt informatie geeft over de sterkte en de richting van de samenhang. Ook hebben we gezien dat je deze correlatie in een spreidingsdiagram kunt visualiseren. We kunnen op basis van een spreidingsdiagram al een (voorzichtige) conclusie trekken over het verband.
Een regressieanalyse gaat een stapje verder. Bij een regressieanalyse maak je onderscheid tussen afhankelijke en onafhankelijke variabelen. Naast de sterkte en richting van het verband geeft een regressieanalyse een voorspelling van de mate waarin de afhankelijke variabele verandert als gevolg van variatie in de onafhankelijke variabele.
Bij een regressieanalyse wordt dus een relatie tussen een onafhankelijke en afhankelijke variabele verondersteld, en deze is daarmee altijd asymmetrisch. We spreken over een *enkelvoudige regressie* als er één onafhankelijke variabele en één afhankelijke variabele is. De doelstelling van een regressieanalyse is het ontdekken van een patroon in de samenhang tussen x en y, zodat je voorspellingen (in de betekenis van schattingen) kunt doen over y indien x is gegeven. De regressieanalyses die we in dit boek bespreken zijn allemaal lineaire regressies. Met een lineaire regressie bereken je de best passende rechte (= lineaire) lijn door een puntenwolk van een spreidingsdiagram. Dat betekent ook dat, net als bij een Pearsons correlatiecoëfficiënt, er geen sprake mag zijn van een kromlijnig verband, en dat moet je dus altijd eerst nagaan door middel van een spreidingsdiagram. In formule ziet de regressielijn bij een enkelvoudige regressieanalyse er als volgt uit:

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

Formule voor enkelvoudige regressie

De lijn is een geschatte lijn. Een regressielijn houdt rekening met alle puntjes in een spreidingsdiagram (de afzonderlijke waarnemingen), en loopt daar zo tussen dat de afstand van alle puntjes tot de lijn minimaal is. Het dakje op de y

geeft aan dat het hier om een schatting of een voorspelling van de waarde van y gaat; $\hat{y}$ is de voorspelde waarde van y.



Figuur 8.8 Grafische weergave van regressielijn

De a in de formule noem je de *intercept* of de *constante*. Dit is het snijpunt van de regressielijn met de y-as, oftewel, a is de voorspelde waarde van y als $x = 0$. De b is de *ongestandaardiseerde regressiecoëfficiënt*. Deze coëfficiënt is bepalend voor de hellingshoek van de lijn. De ongestandaardiseerde regressiecoëfficiënt b geeft aan met hoeveel eenheden de voorspelde waarde van de afhankelijke variabele y verandert als de onafhankelijke variabele x met één eenheid toeneemt.

8.2.1 Berekening

Bij de behandeling van de regressieanalyse bespreken we eerst de wijze van berekenen en dan pas de interpretatie. De interpretatie is namelijk gemakkelijker te begrijpen wanneer je weet hoe je de berekening uitvoert.
Stel, we willen onderzoeken of leeftijd van invloed is op de waardering van het NOS *Journaal*. We vragen drie personen van verschillende leeftijden naar hun waardering van het NOS *Journaal*. Om de waardering van het NOS *Journaal* te meten is een meetinstrument ontwikkeld. De te meten variabele kan de waarde hebben van 0 tot en met 100. De waarde 0 betekent dat de persoon helemaal geen waardering heeft voor het NOS *Journaal* en 100 een zeer hoge waardering. Dit levert de data op die in tabel 8.8 staan.

Tabel 8.8 Datamatrix leeftijd – waardering NOS Journaal

Respondent	Leeftijd	Waardering
A	10	30
B	30	70
C	50	80

Op basis van deze gegevens kun je voorspellen welke waarde y (waardering voor het NOS *Journaal*) zal hebben bij een bepaalde waarde van x (leeftijd). Voor deze voorspelling gebruik je de formule voor de regressielijn: $\hat{y}=a+bx$.
Je hebt de gegevens van de datamatrix nodig om a (de intercept) en b (de ongestandaardiseerde regressiecoëfficiënt) te berekenen. Eén punt van de regressielijn is snel te bepalen, namelijk het punt $(\bar{x},\bar{y})$. Deze twee gemiddeldes liggen op de regressielijn. Van dat gegeven maak je gebruik om de intercept te berekenen. Je kunt de gemiddelden van x en y in de regressievergelijking invullen. Kijken we naar de formule, dan lijkt het logisch te beginnen met het uitrekenen van de intercept (a).

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

Formule voor de intercept

Maar om de intercept uit te rekenen, heb je de ongestandaardiseerde regressiecoëfficiënt (b) nodig. Je moet daarom wel met de berekening van b beginnen.

$$b=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

Formule voor de ongestandaardiseerde regressiecoëfficiënt

Voor de berekening van b heb je de gemiddeldes van x en y nodig. Je zet alle informatie die je nodig hebt om de formule voor b in te vullen in een tabel. Per onderzoekseenheid bereken je $(x-\bar{x})^2$ (de noemer) en $(x-\bar{x})\,(y-\bar{y})$ (de teller). Pas daarna tel je de gevonden getallen bij elkaar op (zie tabel 8.9). De teller van de ongestandaardiseerde regressiecoëfficiënt wordt hetzelfde berekend als de teller van de covariantie zoals we die in paragraaf 8.1.3 zagen.

Tabel 8.9 Berekenen van de ongestandaardiseerde regressiecoëfficiënt

Respondent	x_i	y_i	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(x-\bar{x})\,(y-\bar{y})$
A	10	30	(10 – 30) = –20	400	(30 – 60) = –30	–20 * –30 = 600
B	30	70	(30 – 30) = 0	0	(70 – 60) = 10	0 * 10 = 0
C	50	80	(50 – 30) = 20	400	(80 – 60) = 20	20 * 20 = 400
Σ	90	180		800		1000
M	30	60				

Alle informatie om b te berekenen staat nu in tabel 8.9, zodat je de formule kunt invullen.

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{1000}{800} = 1{,}250$$

Nu kun je ook a berekenen.
$a = \bar{y} - b\bar{x} = 60 - 1{,}25 * 30 = 22{,}500$

De regressievergelijking luidt dus: $\hat{y} = a + bx = 22{,}5 + 1{,}25(x)$

Wat wil dit nu zeggen? De intercept (a) is de voorspelde waarde van y (waardering voor het *NOS Journaal*) wanneer x (leeftijd) de waarde 0 heeft. Letterlijk betekent dit: wanneer iemand nul jaar oud is, zal de waardering voor het *NOS Journaal* 22,5 zijn (want dat was de waarde van de intercept). Een onzinnige voorspelling natuurlijk, wij weten wel beter dan die regressievergelijking. Toch is deze a in de regressievergelijking nodig om voorspellingen te kunnen doen voor elke x (leeftijd). De ongestandaardiseerde regressiecoëfficiënt geeft aan hoeveel de voorspelde waarde van y toeneemt als x met één eenheid toeneemt. Hier zeggen we: het model voorspelt dat wanneer iemand één jaar ouder wordt, de waardering voor het *NOS Journaal* stijgt met 1,25 (op een schaal van 0 tot 100).

Op basis van de intercept en de ongestandaardiseerde regressiecoëfficiënt kun je nu de waardering voor het *NOS Journaal* bij elke willekeurige leeftijd voorspellen door deze leeftijd (x) in de regressievergelijking in te vullen. Je zou bijvoorbeeld kunnen voorspellen hoe een 21-jarige het *NOS Journaal* waardeert:
$\hat{y} = 22{,}5 + 1{,}25 * 21 = 48{,}750$ (op een schaal van 0 tot 100).

We hebben in dit voorbeeld onder onze drie respondenten geen 21-jarige. Als we veel respondenten zouden hebben ondervraagd, met daarin wel een 21-jarige, zouden de scores van een 21-jarige in de puntenwolk die we dan krijgen waarschijnlijk niet precies op de regressielijn liggen. Er blijft variatie rond die regressielijn bestaan, een restvariatie die je niet kunt verklaren met de onafhankelijke variabele. Er zijn namelijk ook andere factoren dan alleen leeftijd die de waardering van het *NOS Journaal* verklaren, maar die hebben we niet met deze regressielijn gemeten. De mate waarin de regressielijn de variatie (of: variantie) verklaart, kunnen we berekenen. Bij regressieanalyse is de *proportie verklaarde variantie* R^2 een belangrijk begrip.[3] Deze lijkt op Goodman en Kruskals tau en lambda (zie hoofdstuk 5). Ook R^2 is gebaseerd op de proportie (of het percentage) voorspellingsverbetering. Ook hier wil je dus bepalen hoe goed de verschillen in de waarden van de onafhankelijke variabele (oftewel: de variantie in de onafhankelijke variabele) de verschillen in de waarden van de afhankelijke variabele verklaren. Het verschil is dat je tau en lambda gebruikt

bij nominale variabelen, en R^2 bij interval- of ratiovariabelen. R^2 is het symbool voor de proportie verklaarde variantie. De formule voor R^2 komt overeen met die van de tau en lambda:

$$R^2 = \frac{E_1 - E_2}{E_1}$$

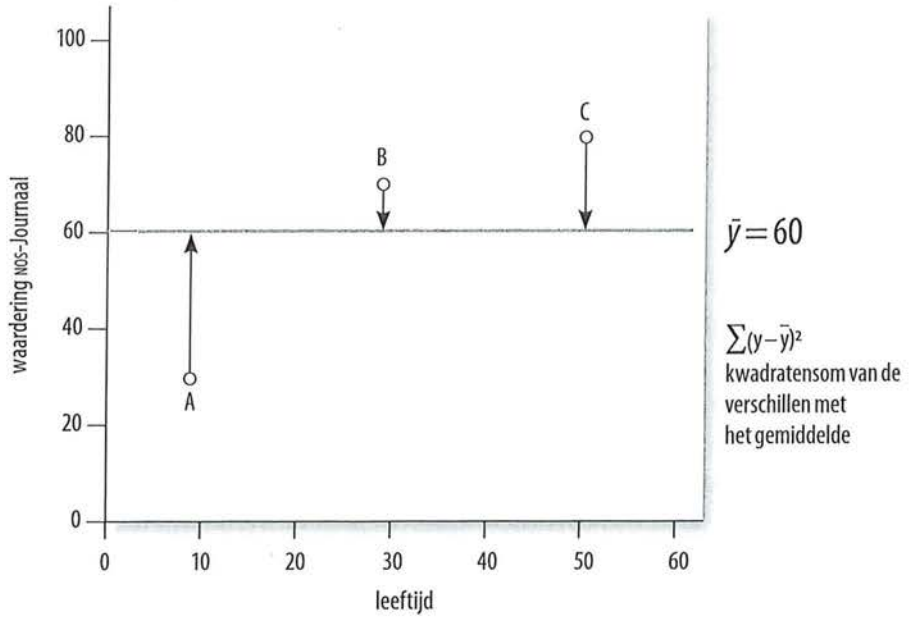
Formule voor proportie verklaarde variantie.

De manier waarop je E_1 en E_2 berekent, is wel anders dan bij tau en bij lambda. Als je bij een interval- of ratiovariabele een voorspelling wilt doen voor y en je daarvoor niet de informatie van x gebruikt maar alleen de gegevens over y, is de beste keuze het gemiddelde van die variabele y. Maar dan maak je voor de afzonderlijke onderzoekseenheden wel een fout die gelijk is aan de waarde van y minus $\bar{y}$. Als je voor alle onderzoekseenheden deze verschillen bij elkaar optelt, is de som 0. Daarom kwadrateren we de verschillen eerst. Die kwadratensom is de *totale variatie*;[4] dit zijn de voorspellingsfouten die je maakt als je het rekenkundig gemiddelde als voorspeller gebruikt.

$$E_1 = \sum (y_i - \bar{y})^2$$

Formule van E_1 bij een regressieanalyse

Een voorbeeld ter verduidelijking. We kijken naar de mate waarin we de waardering van het NOS *Journaal* kunnen voorspellen aan de hand van leeftijd. De waardering voor het NOS *Journaal* is hier de afhankelijke variabele, en bij het berekenen van de E_1 houden we alleen rekening met deze afhankelijke variabele. Van deze variabele kunnen we, omdat het meetniveau hier minimaal interval is, een gemiddelde berekenen. In ons voorbeeld is dat 60. Ongeacht de leeftijd is de gemiddelde waardering voor het Journaal 60 op een schaal van 0 tot en met 100. Persoon A scoorde op waardering van het Journaal 30. De afwijking ten opzichte van het gemiddelde (60) is dus –30. Respondent B scoort met een waardering van 70, 10 punten hoger dan het gemiddelde, en respondent C scoort 20 boven het gemiddelde. Zou je deze afwijkingen bij elkaar optellen, dan kom je uit op nul:
$-30 + 10 + 20 = 0$.
Om die reden kwadrateren we deze afwijkingen afzonderlijk en tellen ze daarna bij elkaar op, zoals we ook bij het berekenen van de variantie (zie hoofdstuk 3) deden. E_1 is dus $-30^2 + 10^2 + 20^2 = 1400$.



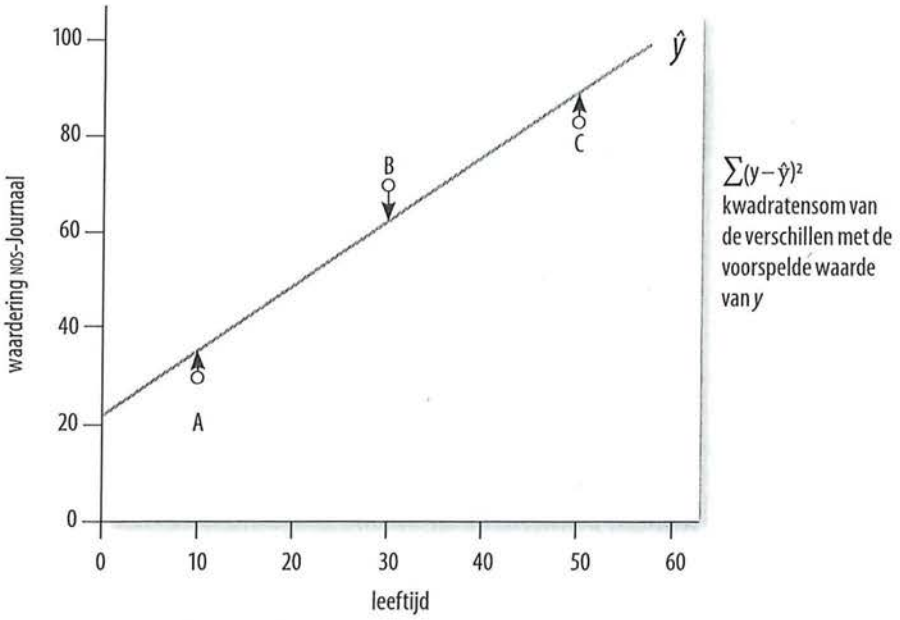
Figuur 8.9 Totale variatie (E_1), afstanden tot het gemiddelde

Bij het berekenen van E_2 maken we wel gebruik van de onafhankelijke variabele. Door de regressievergelijking te gebruiken, pas je informatie over x toe op je voorspelling van y. Ook dan maak je fouten. In dit geval is de voorspellingsfout per onderzoekseenheid y minus $\hat{y}$. De som van de kwadraten van deze verschillen is de *onverklaarde variatie*. De onverklaarde variatie is de voorspellingsfout als je $\hat{y}$ als voorspeller gebruikt. Deze resterende voorspellingsfout, de restvariatie, is gebaseerd op de verschillen met de voorspelde waarden:

$$E_2 = \sum(y-\hat{y})^2$$

Formule van E_2 bij regressieanalyse

De regressielijn in ons voorbeeld was $\hat{y}=22{,}5+1{,}25(x)$. Dat betekent in dit geval dat die lijn voorspelt dat een persoon van 10 jaar oud een waardering heeft van 35 voor het journaal: $\hat{y}=22{,}5+1{,}25*10=35$. Kijken we in onze datamatrix, dan zien we dat persoon A, die 10 jaar oud is, door de regressielijn wordt overschat, want deze persoon heeft een waardering van 30. Persoon A heeft dus een afwijking van 30 – 35 = -5 ten opzichte van de regressielijn. Voor een 30-jarige voorspelt de regressielijn: $\hat{y}=22{,}5+1{,}25*30=60$. Deze persoon wordt door de regressielijn dus onderschat, want de score van de 30-jarige in onze datamatrix is 70: 70 – 60 = 10. Tot slot zien we op dezelfde manier dat persoon C als 50-jarige –5 ten opzichte van de regressielijn scoort: $\hat{y}=22{,}5+1{,}25*50=85$ en 80 – 85 = –5. Ook deze waarden zijn bij elkaar opgeteld nul, dus kwadrateren we ze afzonderlijk voordat we deze bij elkaar optellen om tot E_2te komen: –52 + 102 + –52 = 150.



Figuur 8.10 Onverklaarde variatie (E_2), afstanden tot de regressielijn

We zien in figuur 8.10 dat de afstanden van de punten van de onderzoekseenheden tot de regressielijn kleiner zijn dan de afstanden van deze punten tot het gemiddelde in figuur 8.9. Wanneer alle waarnemingen precies op de lijn zouden liggen, zouden de verschillen in leeftijd perfect de verschillen in waardering verklaren. Nu is er echter een kleine afstand tussen A en C en de lijn, en een iets grotere afstand tussen B en de lijn. Er is dus een gedeelte dat niet wordt verklaard door de lijn: de onverklaarde variatie. Dit noem je ook wel het *residu*.

Wanneer we deze informatie samenvoegen, komen we tot de volgende formule voor de proportie verklaarde variantie.

$$R^2 = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{totale variatie} - \text{onverklaarde variatie}}{\text{totale variatie}}$$

Rekenformule voor de proportie verklaarde variantie

Net als tau en lambda heeft R^2 een minimale waarde van 0 en een maximale waarde van 1.

We laten de hele berekening van de R^2 nogmaals zien voor ons voorbeeld (leeftijd en waardering) in een overzichtstabel.

Tabel 8.10 Berekenen van de proportie verklaarde variantie (R^2)

Respondent	x_i	y_i	$(y-\bar{y})$	$(y-\bar{y})^2$	$\hat{y}=a+bx$	$(y-\hat{y})$	$(y-\hat{y})^2$
A	10	30	–30	900	22,5 + 1,25 * 10 = 35	(30 – 35) = –5	25
B	30	70	10	100	22,5 + 1,25 * 30 = 60	(70 – 60) = 10	100
C	50	80	20	400	22,5 + 1,25 * 50 = 85	(80 – 85) = –5	25
Σ	90	180	0	1400			150
M	30	60					

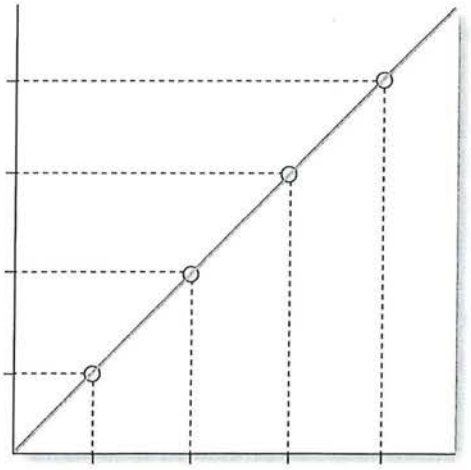
De totale variatie (E_1) is dus 1400, de onverklaarde variatie (E_2) is 150.

Nu we alle informatie hebben, kunnen we de formule invullen.

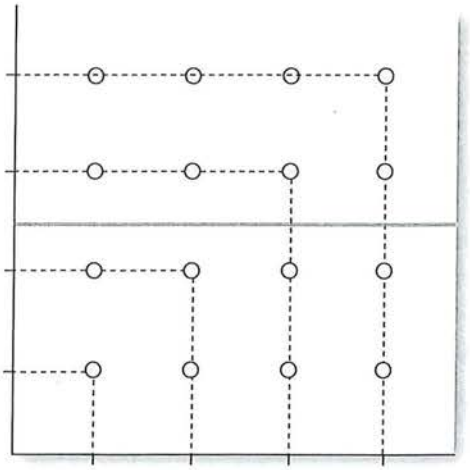
$$R^2=\frac{\sum(y-\bar{y})^2-\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}=\frac{1400-150}{1400}=0{,}893$$

We kunnen concluderen dat de varia(n)tie in de waardering voor het *NOS Journaal*, voor 89,3% verklaard wordt door de variantie in de variabele leeftijd. De regressievergelijking is dus een goed verklaringsmodel.

R^2 kan waarden aannemen die liggen tussen de 0 (0% verklaring) en de 1 (100% verklaring). Bij een perfecte verklaring liggen alle punten precies op de regressielijn; er is dan geen restwaarde of residu ($\sum(y-\hat{y})^2=0$). Als 0% wordt verklaard, valt de regressielijn samen met de gemiddelde waarde van y en zijn de verschillen met het gemiddelde gelijk aan de restwaarden $\sum(y-\bar{y})^2=\sum(y-\hat{y})^2$ (zie figuur 8.11 en 8.12).



Figuur 8.11 $R^2 = 1$



Figuur 8.12 $R^2 = 0$

8.2.2 *Interpretatie*

Nu we hebben gezien hoe je de berekeningen van een enkelvoudige regressieanalyse uitvoert, zal de interpretatie van de SPSS-output gemakkelijker zijn. We nemen nog steeds het voorbeeld van de samenhang tussen leeftijd en de waardering van het NOS *Journaal*. We bekijken in kleine stappen de tabellen die SPSS als output van een regressieanalyse geeft (zie kader 8.4 voor de wijze waarop je SPSS een regressieanalyse kunt laten uitvoeren).

Tabel 8.11 Regressieanalyse: intercept (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22,500	14,790		1,521	,370
	leeftijd	1,250	,433	,945	2,887	,212

a. Dependent Variable: waardering

De eerste tabel van SPSS die we hier bespreken is de coëfficiëntentabel. Deze zie je in tabel 8.11. De onafhankelijke variabele (in dit geval leeftijd) staat in de linkerkolom van deze coëfficiëntentabel. Wat de afhankelijke variabele (in dit geval waardering) is, wordt onder deze tabel aangegeven. De waarde die meteen onder de B staat, achter (Constant), is de intercept, het snijpunt met de *y*-as (*a*). Deze bedraagt 22,50, zoals we zelf al hadden berekend. De letterlijke betekenis van dit getal is: wanneer iemand nul jaar oud is, is de voorspelde waarde voor de waardering van het NOS *Journaal* 22,50.

Tabel 8.12 Regressieanalyse: ongestandaardiseerde regressiecoëfficiënt (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22,500	14,790		1,521	,370
	leeftijd	1,250	,433	,945	2,887	,212

a. Dependent Variable: waardering

De waarde van de ongestandaardiseerde regressiecoëfficiënt *b* wordt onder de intercept (*Constant*) gegeven, achter de onafhankelijke variabele (tabel 8.12). Dit is de waarde waarmee de voorspelling van *y* verandert als *x* met één eenheid toeneemt. Dit is het effect van *x* op *y*. Hier betekent deze waarde dus: wanneer de leeftijd met één jaar toeneemt, stijgt de waardering van het NOS *Journaal* met 1,25.

Tabel 8.13 Regressieanalyse: gestandaardiseerde regressiecoëfficiënt (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22,500	14,790		1,521	,370
	leeftijd	1,250	,433	,945	2,887	,212

a. Dependent Variable: waardering

In de kolom *Standardized Coefficients* staat bèta (β). De β is het zuivere effect van de onafhankelijke variabele leeftijd op de afhankelijke variabele waardering. Dit wordt ook wel de *gestandaardiseerde regressiecoëfficiënt* genoemd, en heet zo omdat deze niet afhankelijk is van de meeteenheden van de variabelen. Bèta neemt in de regel alleen waarden aan die tussen –1 en +1 liggen. Bij een enkelvoudige regressieanalyse (regressieanalyse met één onafhankelijke variabele) is bèta altijd gelijk aan de correlatiecoëfficiënt r. We zien dat er een zeer sterke, positieve samenhang is tussen leeftijd en waardering ($\beta = 0{,}95$).

In de tabel *Model Summary* geeft SPSS de proportie verklaarde variantie (R^2) en de multipele correlatie (tabel 8.14). De wortel uit R^2 is R. R is een multipele correlatiecoëfficiënt (zie paragraaf 8.3). Als er één onafhankelijke variabele is, is deze R gelijk aan $|\, r \,|$, de absolute waarde van de correlatie r. Bij meerdere onafhankelijke variabelen is R niet gelijk aan $|\, r \,|$. We zullen hier verder op ingaan in de volgende paragraaf.

Tabel 8.14 Regressieanalyse: proportie verklaarde variantie (SPSS-output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,945[a]	,893	,786	12,24745

a. Predictors: (Constant), leeftijd

Zoals we eerder al hadden berekend, blijkt de onafhankelijke variabele leeftijd 89,3% van de varia(n)tie in de afhankelijke variabele waardering te verklaren. SPSS geeft ook de waarde van de totale variatie (E_1) en van de onverklaarde variatie (E_2) in een aparte tabel, met als titel ANOVA (tabel 8.15). De totale variatie wordt aangeduid met *Total*, de onverklaarde variatie wordt met *Residual* aangeduid.

Aan de hand van tabel 8.15 kun je ook zelf R^2 uitrekenen.

$$R^2 = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{E_1 - E_2}{E_1} = \frac{\text{Total} - \text{Residual}}{\text{Total}} = \frac{1400-150}{1400} = 0{,}893$$

Tabel 8.15 Berekening proportie verklaarde variantie (SPSS-output)

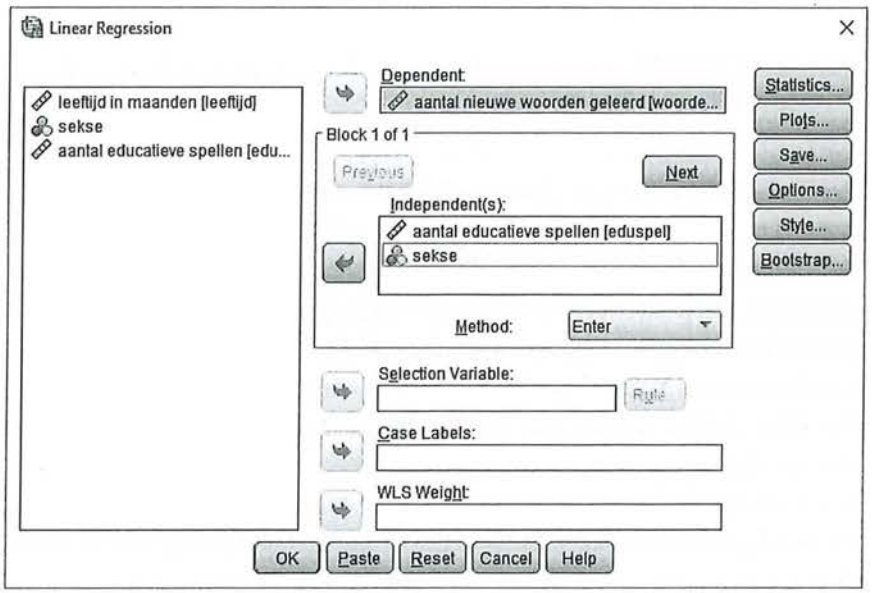
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1250,000	1	1250,000	8,333	,212[b]
	Residual	150,000	1	150,000		
	Total	1400,000	2			

a. Dependent Variable: waardering
b. Predictors: (Constant), leeftijd

SPSS Het uitvoeren van een regressieanalyse

Het uitvoeren van een regressieanalyse in SPSS doe je met *Analyze → Regression → Linear.* In het vakje *Dependent* voer je de afhankelijke variabele in en in het vakje *Independent(s)* één of meerdere onafhankelijke variabelen (voor een meervoudige regressieanalyse zie paragraaf 8.3).



Figuur A Linear Regression-venster

Kader 8.4

We behandelen nog een ander voorbeeld. We kijken weer naar de woordenschat van peuters wanneer ze educatieve spellen op een tablet spelen, maar meten deze variabelen nu allemaal op rationiveau. Een week lang wordt geteld hoe vaak de peuters een educatief spel spelen, en wordt het aantal nieuwe woorden dat zij zeggen in die week geteld. De verwachting is dat peuters die meer educatieve spellen hebben gespeeld (de onafhankelijke variabele) meer nieuwe woorden leren in de week (de afhankelijke variabele). Beide variabelen hebben een ratio meetniveau en we veronderstellen een asymmetrisch verband tussen

de twee variabelen. Een enkelvoudige regressieanalyse is daarom de meest geschikte analysetechniek om dit te onderzoeken. SPSS laat vier tabellen zien, waarvan er twee nodig zijn om een antwoord te geven op de verwachting (tabel 8.16).

Tabel 8.16 Regressieanalyse aantal educatieve spellen en woordenschat (SPSS-output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,536[a]	,287	,269	1,57891

a. Predictors: (Constant), eduspel aantal educatieve spellen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7,987	,839		9,516	,000
	eduspel aantal educatieve spellen	,160	,041	,536	3,915	,000

a. Dependent Variable: woordenschat aantal nieuwe woorden geleerd

Je kijkt eerst naar de intercept. Deze is 7,987. Letterlijk betekent deze waarde: wanneer x nul is, is de verwachte waarde van y 7,99. In ons voorbeeld concluderen we: het model voorspelt dat wanneer geen educatieve spellen op de tablet worden gespeeld, de peuter 7,99 nieuwe woorden per week leert.

De ongestandaardiseerde regressiecoëfficiënt (b) bedraagt 0,160. Letterlijk betekent dit: wanneer x met één eenheid stijgt, stijgt de verwachte waarde van y met 0,16. Hier stellen we dus dat het model voorspelt dat wanneer het aantal educatieve spellen spelen toeneemt met één keer, de woordenschat toeneemt met 0,16 woorden. Aan de gestandaardiseerde regressiecoëfficiënt (β) zien we dat dit verband positief en sterk is (β = 0,54). Deze is gelijk aan de R, wat weer de wortel is uit R^2.

Om te kijken of het aantal educatieve spellen spelen een goede verklaring is voor de woordenschat van peuters, kijken we naar de R^2. Deze is 0,287. De variantie in het aantal educatieve spellen spelen, verklaart voor 28,7% de variantie in de woordenschat. Met andere woorden: de mate waarin peuters educatieve spellen spelen op een tablet, verklaart voor 28,7% de mate waarin peuters wekelijks nieuwe woorden leren.

In een onderzoeksverslag zouden we dit als volgt kunnen beschrijven:

> *Uit een enkelvoudige regressieanalyse blijkt dat we vrij goed het aantal nieuwe woorden dat een peuter wekelijks leert kunnen voorspellen aan de hand van het aantal keer educatieve spellen op een tablet spelen (β = 0,54, n = 40).[5] De proportie verklaarde variantie is 28,7%. 71,3% van de*

variantie in het leren van nieuwe woorden kan dus niet verklaard worden door het spelen van educatieve spellen op een tablet. Peuters die geen educatieve spellen spelen, leren 7,99 nieuwe woorden per week. Er is een toename van 0,16 nieuwe woorden per week wanneer de frequentie van de educatieve spellen spelen toeneemt met één.[6]

8.3 Meervoudige regressieanalyse

In de vorige paragraaf keken we naar de invloed van één onafhankelijke variabele op één afhankelijke variabele. Bij een meervoudige regressieanalyse kijk je naar het voorspelde effect van meerdere onafhankelijke variabelen. Er is nog steeds maar één afhankelijke variabele. We zullen de meervoudige regressieanalyses niet met de hand berekenen zoals we dat bij de enkelvoudige hebben gedaan, maar ons voornamelijk richten op de interpretatie.
We hebben gezien dat het aantal nieuwe woorden dat geleerd wordt redelijk (voor 28,7%) wordt verklaard door het spelen van educatieve spellen. Daarbij hebben we geen rekening gehouden met andere variabelen. Misschien dat andere factoren ook een rol spelen. Het zou kunnen zijn dat meisjes sneller nieuwe woorden leren dan jongens, of misschien heeft leeftijd wel een grote verklarende invloed en gaat het leren van nieuwe woorden bij drie- en vierjarigen sneller dan bij tweejarigen. Bij een meervoudige regressieanalyse is het mogelijk om ook deze variabelen op te nemen in het verklaringsmodel. Een voorwaarde voor een regressieanalyse was echter wel dat alle variabelen minimaal op intervalniveau gemeten zouden zijn, en 'sekse' is een nominale variabele. Voordat we kijken naar de interpretatie van een regressieanalyse met meerdere onafhankelijke variabelen, bespreken we daarom eerst hoe je nominale variabelen toch in een regressieanalyse op kunt nemen.

8.3.1 Dummyvariabelen

We hebben al gezien dat je met nominale variabelen niet kunt rekenen. We willen echter vaak deze variabelen wel in onze analyses betrekken, zoals de variabele 'geslacht'. Het zou zonde zijn als we met de variabele geslacht alleen kruistabellen kunnen uitvoeren, terwijl het een variabele is die in veel onderzoek meegenomen zal worden. Rekenen met nominale variabelen is mogelijk wanneer we van deze nominale variabelen *dummyvariabelen* maken. Een dummyvariabele is een variabele die dichotoom is, of dichotoom is gemaakt. Een *dichotome variabele* is een variabele die slechts twee mogelijke waarden kan aannemen, bijvoorbeeld man-vrouw, of goed-fout, of ja-nee. In het geval van een dummyvariabele geven we deze waarden altijd de waarden 0 en 1. In het geval van sekse maakt het niet uit of je de man de waarde 1 geeft of de waarde 0. Bij een variabele waarbij je alleen met ja of nee kunt antwoorden ligt het voor de hand om 'nee' de waarde 0 te geven en 'ja' de waarde 1. De waarde

1 geeft dan aan dat de onderzoekseenheid het kenmerk wel heeft en de waarde 0 dat dit niet het geval is. Je zou bijvoorbeeld aan iemand de vraag kunnen stellen: 'ben je een man'? Is het antwoord nee (waarde 0), dan kan het niet anders dan dat deze persoon een vrouw is. We zullen in de volgende paragraaf zien dat wanneer we op deze manier (met nullen en enen) nominale variabelen coderen, we deze toch in een meervoudige regressieanalyse kunnen opnemen.

Sommige nominale variabelen zijn altijd dichotoom, hebben dus altijd slechts twee waarden (zoals sekse), maar een variabele met meerdere waarden kan ook worden gedichotomiseerd tot dummyvariabelen. Wanneer je hebt gevraagd op welke partij iemand stemt, resulteert dat in een nominale variabele met verschillende antwoordcategorieën. Partijvoorkeur is een nominale variabele en daarom niet geschikt voor gebruik in een regressieanalyse. We maken dan in SPSS nieuwe dummyvariabelen (met behulp van *Recode*, zie paragraaf 4.4), waarin we in feite steeds de vraag stellen: 'heeft de respondent op deze partij gestemd?' waarbij steeds geantwoord kan worden met 0 = nee of 1 = ja.

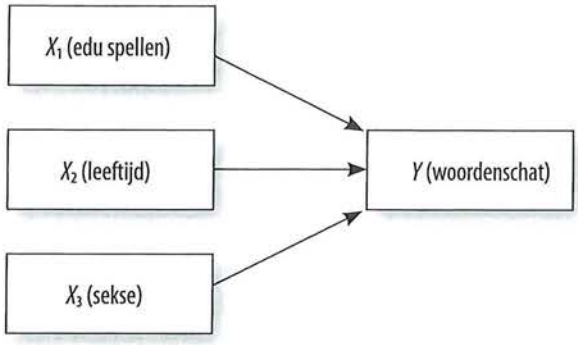
Stel, we hebben respondenten in een enquête de vraag voorgelegd: 'op welke partij zou u bij de komende verkiezingen stemmen?' en ze kunnen (voor de overzichtelijkheid van ons voorbeeld) kiezen uit vier mogelijke antwoorden: 1 = SP, 2 = PvdA, 3 = VVD, 4 = overige partijen. Deze variabele is nominaal, er zit geen rangordening in, er kan niet met de waarden van de antwoorden gerekend worden. We maken nu van deze nominale variabele met vier waarden, drie nieuwe dummyvariabelen. De eerste variabele noemen we 'SP', en heeft de waarden 0 = nee, ik zou niet op de SP stemmen, en 1 = ja, ik zou wel op de SP stemmen. De tweede variabele noemen we 'PvdA', en heeft ook weer twee waarden: 0 = nee, ik zou niet op de PvdA stemmen en 1 = ja, ik zou wel op de PvdA stemmen. En de derde variabele noemen we 'VVD', en heeft ook weer twee waarden: 0 = nee, ik zou niet op de VVD stemmen en 1 = ja, ik zou wel op de VVD stemmen. Voor de laatste categorie, overige partijen, hoeven we geen nieuwe dummyvariabele te maken, omdat wanneer op zowel 'SP' als 'PvdA' als 'VVD' 0 (nee) wordt geantwoord, de respondent automatisch op een van de overige partijen zal stemmen. Bij het maken van een dummyvariabele maak je dus altijd 'het aantal categorieën min 1' aantal dummyvariabelen.

Dummyvariabelen gebruiken we alleen in meervoudige regressieanalyses, en niet in enkelvoudige regressieanalyses. In hoofdstuk 9 zullen we bespreken welke analyse het meest geschikt is wanneer je één nominale onafhankelijke variabele hebt (al dan niet dichotoom) en een interval- of ratiovariabele als afhankelijke variabele. In de volgende paragraaf zullen we laten zien hoe je de dummyvariabelen in een meervoudige regressieanalyse kunt interpreteren.

8.3.2 *Interpretatie meervoudige regressieanalyse*

Een voorwaarde voor een meervoudige regressie is dat de afhankelijke variabele minimaal intervalniveau is, en dat er minimaal één onafhankelijke variabele op minimaal intervalniveau gemeten is. De andere onafhankelijke variabelen kunnen dummyvariabelen of ook interval- of ratiovariabelen zijn.

We gaan verder met het voorbeeld van de mogelijke toename van de woordenschat van peuters bij het spelen van educatieve spellen op een tablet. We voegen nu de variabelen 'leeftijd in maanden' en 'sekse' toe, waarbij we de waarde 0 hebben toegekend aan meisjes, en de waarde 1 aan jongens. Alleen op deze manier mogen we de nominale variabele immers in een meervoudige regressieanalyse gebruiken.



Figuur 8.13 Effect van meerdere onafhankelijke variabelen op een afhankelijke variabele

De eerdere formule voor de regressievergelijking is nu 'verdrievoudigd':

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$$

Formule voor meervoudige regressie bij drie onafhankelijke variabelen

We hebben drie onafhankelijke variabelen: educatieve spellen (x_1), leeftijd in maanden (x_2) en geslacht (x_3). Elke onafhankelijke variabele heeft zijn eigen ongestandaardiseerde regressiecoëfficiënt (b). Uit de formule blijkt dat er nog steeds maar één intercept is. Dit is het punt waarop de y-as wordt gesneden als alle x'en de waarde nul hebben.
De ongestandaardiseerde regressiecoëfficiënt van een onafhankelijke variabele geeft het effect van die variabele op y weer als de andere onafhankelijke variabelen niet veranderen (constant worden gehouden). Zo is b_1 het effect dat x_1 heeft op y, onder het constant houden van de overige x'en. Wanneer je kijkt naar b_1 om het effect van 'educatieve spellen' op de woordenschat vast te stellen, houd je leeftijd en sekse constant, zoals we dat ook al eerder deden bij tabelsplitsing en partiële correlatie. Hoe een meervoudige regressie er in SPSS-output uitziet, laat tabel 8.17 zien.

Tabel 8.17 Meervoudige regressie met drie onafhankelijke variabelen (SPSS-output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,729[a]	,531	,475	1,57232

a. Predictors: (Constant), sekse, leeftijd leeftijd in maanden, eduspel aantal educatieve spellen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,961	1,136		6,130	,000
	eduspel aantal educatieve spellen	,143	,048	,526	3,261	,002
	leeftijd leeftijd in maanden	2,014	,026	,181	,520	,606
	sekse	-1,687	,531	-,386	-1,293	,204

a. Dependent Variable: woordenschat aantal nieuwe woorden geleerd

De interpretatie begint bij de coëfficiëntentabel (met de titel *Coefficients*). De waarde van de intercept, achter *Constant*, is 6,961. Dat wil zeggen dat wanneer alle onafhankelijke variabelen nul zijn, het voorspelde aantal nieuwe woorden dat wordt geleerd, 6,96 is. Omdat sekse in deze analyse de waarden 0 en 1 heeft (meisjes hadden de waarde 0, jongens de waarde 1), kun je stellen dat meisjes (sekse = 0) van nul maanden (leeftijd = 0) die geen educatieve spellen spelen (eduspel = 0), 6,96 nieuwe woorden per week leren. Dit is dus het punt waarop de regressielijnen de y-as snijden.

We kijken nu naar de eerste ongestandaardiseerde regressiecoëfficiënt, b_1, achter de variabele 'eduspel'. Deze heeft een waarde van 0,143. Wanneer x met één eenheid stijgt, neemt de woordenschat toe met 0,14 als de andere onafhankelijke variabelen ongewijzigd blijven. Het model voorspelt dus dat peuters 0,14 meer nieuwe woorden leren wanneer ze één keer vaker een educatief spel op een tablet spelen, onder het constant houden van leeftijd en sekse. Dit verschilt niet veel met de toename van 0,16 die we zagen bij de enkelvoudige regressieanalyse (tabel 8.16). Het verschil is dat je nu rekening houdt met de andere twee onafhankelijke variabelen. Op basis van de regressievergelijking kun je bijvoorbeeld voorspellen dat een meisje van 25 maanden oud dat vier keer per week een spel speelt, 0,16 nieuwe woorden meer leert dan een meisje van 25 maanden oud dat drie keer per week een spel speelt.

Bij de ongestandaardiseerde regressiecoëfficiënt van leeftijd, b_2, zien we een waarde van 2,014. Dat wil zeggen: wanneer de leeftijd toeneemt met één, dus als een peuter één maand ouder wordt, stijgt het aantal nieuwe woorden met 2,01, onder constanthouding van het 'aantal educatieve spellen' dat gespeeld wordt

en 'sekse'. Een meisje van 25 maanden dat drie keer per week een educatief spel speelt, leert 2,01 nieuwe woorden per week meer dan een meisje van 24 maanden dat drie keer per week een educatief spel speelt.

De laatste ongestandaardiseerde regressiecoëfficiënt, b_3 van sekse, heeft een interpretatie die iets afwijkt van de eerdere twee. Omdat we hier te maken hebben met een dichotome variabele, kunnen we hier iets zeggen over het *gemiddelde* verschil tussen de twee categorieën (meisje-zijn of jongen-zijn). De waarde van b_3 is –1,687. Letterlijk staat er: wanneer sekse met één eenheid toeneemt, neemt het aantal nieuwe woorden dat geleerd wordt af met 1,69, onder het constant houden van het aantal educatieve spellen en leeftijd. Dat is natuurlijk een rare conclusie, want sekse kan niet toenemen (en net zo min afnemen). Omdat meisjes hier de waarde 0 hebben en jongens de waarde 1 hebben, en er geen andere waarden dan dat zijn, kunnen we hier daarom zeggen: jongens leren gemiddeld 1,69 nieuwe woorden minder dan meisjes, waarbij we controleren voor het aantal educatieve spellen dat gespeeld wordt op een tablet en de leeftijd. Een meisje van 22 maanden oud dat vijf keer per week een educatief spel speelt, zal gemiddeld 1,69 nieuwe woorden meer leren in de week dan een jongen van 22 maanden oud die vijf keer per week een educatief spel speelt.

Behalve dat er drie ongestandaardiseerde regressiecoëfficiënten zijn, zijn er nu ook drie gestandaardiseerde regressiecoëfficiënten (bèta's). Deze geven in een meervoudige regressieanalyse de *partiële zuivere effecten* aan. Omdat deze waarden gestandaardiseerd zijn, kun je de effecten van de verschillende onafhankelijke variabelen met elkaar vergelijken. Zonder standaardisatie is dat niet mogelijk, 'Sekse' heeft immers maar twee waarden (meisje en jongen), leeftijd heeft veel meer waarden en is gemeten in maanden (van 0 tot 60 maanden) en het aantal educatieve spellen dat gespeeld wordt kan misschien wel oplopen tot 30 keer per week. De bèta's variëren in de regel van –1 tot +1. In deze analyse zie je dat 'educatieve spellen spelen' de hoogste waarde van bèta heeft, namelijk 0,526. Je kunt daaruit concluderen dat deze onafhankelijke variabele het sterkste effect heeft op het aantal nieuwe woorden dat geleerd wordt en dat leeftijd het minst sterke effect heeft. Het maakt bij de bèta niet uit of de waarde positief of negatief is, een negatieve samenhang kan immers ook zeer sterk zijn.

Omdat er nu drie bèta's zijn, is R, de multipele correlatiecoëfficiënt, niet meer gelijk aan de correlatiecoëfficiënt $| r |$. In de tabel *Model Summary* zie je dat R 0,729 is. De proportie verklaarde variantie (R^2) is 0,531. Aantal educatieve spellen, leeftijd in maanden en sekse, verklaren samen 53,1% van de variantie in het aantal nieuwe woorden dat een peuter per week leert. Dat is vrij veel (er is immers een sterke multipele samenhang). Er blijft wel 46,9% van de variantie niet verklaard, er zijn dus ook nog andere factoren dan die wij hebben gemeten om de woordenschat van peuters te kunnen voorspellen. Door toevoeging van

de variabelen leeftijd en sekse hebben we de verklaarde variantie in het leren van nieuwe woorden kunnen verhogen van 28,7% naar 53,1%.

Op basis van deze gegevens kun je voorspellingen doen door waarden voor de verschillende onafhankelijke variabelen (*x*'en) in te vullen. Eerst stel je de regressievergelijking op:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 = 6{,}961 + 0{,}143(x_1) + 2{,}014(x_2) - 1{,}687(x_3)$$

Stel, je wilt voorspellen hoeveel nieuwe woorden een jongen van 24 maanden die zeven keer per week een educatief spel op de tablet speelt, leert. Je kunt die gegevens in de regressievergelijking invullen. Je vult voor educatieve spellen (x_1) de waarde 7 in, voor leeftijd de waarde 24, en voor sekse de waarde 1.

Wanneer je dit berekent, kom je op de volgende voorspelling uit:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 = 6{,}961 + 0{,}143 * 7 + 2{,}014 * 24 - 1{,}687 * 1 = 54{,}611$$

Het model voorspelt dat een jongen van 24 maanden oud die zeven keer per week een educatief spel op een tablet speelt, 54,61 nieuwe woorden per week leert.

8.3.3 *Schijnsamenhang in een meervoudige regressie*

Net als bij tabelsplitsing (hoofdstuk 7) houd je bij het uitvoeren van een meervoudige regressieanalyse andere onafhankelijke variabelen constant. Hierdoor is het mogelijk dat een eerder gevonden verband door toevoeging van een (of meerdere) onafhankelijke variabele(n) verdwijnt.

In een onderzoek onder basisscholieren is aan 129 kinderen gevraagd of zij het Jeugdjournaal leuk vonden (uitgedrukt in een rapportcijfer), en of zij wel eens over het nieuws praatten (gemeten in aantal keer per week). De verwachting daarbij is dat hoe leuker kinderen het Jeugdjournaal vinden (onafhankelijke variabele), hoe vaker zij over het nieuws zullen praten (afhankelijke variabele). Aangezien beide variabelen minimaal intervalniveau zijn, en er sprake is van een asymmetrische relatie, is een enkelvoudige regressieanalyse hier de meest geschikte analyse. Uit de regressieanalyse (tabel 8.18) blijkt een sterk positief verband tussen de twee variabelen ($\beta = 0{,}61$), de verwachting komt dus uit: hoe leuker zij het Jeugdjournaal vinden, hoe meer ze praten over het nieuws.

Tabel 8.18 Enkelvoudige regressieanalyse Jeugdjournaal leuk vinden en praten over het nieuws (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,154	,041		45,649	,000
	Jeugdjournaal leuk vinden	,654	,016	,614	21,758	,000

a. Dependent Variable: Praten over nieuws

Vervolgens is de variabele geslacht toegevoegd als controlevariabele, waarbij meisjes de waarde 0 kregen en jongens de waarde 1.

Tabel 8.19 Meervoudige regressieanalyse Jeugdjournaal leuk vinden, geslacht en praten over het nieuws (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,108	,057		36,723	,000
	Jeugdjournaal leuk vinden	,032	,016	,004	20,690	,000
	geslacht (0 = meisje)	-,121	,022	-,629	-5,471	,000

a. Dependent Variable: Praten over nieuws

Aan de gestandaardiseerde regressiecoëfficiënten is nu te zien dat het oorspronkelijke positieve, sterke verband bij Jeugdjournaal verdwijnt: $\beta = 0{,}004$. Bij toevoeging van de variabele geslacht blijkt dat het oorspronkelijke verband een schijnverband is, spurieus is, en dat het praten over het nieuws niet door het al dan niet leuk vinden van het Jeugdjournaal wordt veroorzaakt, maar door of het kind een meisje of een jongen is ($\beta = -0{,}63$). Aan de negatieve waarde van de ongestandaardiseerde regressiecoëfficiënt kunnen we aflezen dat jongens gemiddeld 0,12 keer minder vaak over het nieuws praten dan meisjes, onder constanthouding van of zij het Jeugdjournaal leuk vinden.

8.3.4 *Regressie- en correlatieanalyses in wetenschappelijke tijdschriften*

Regressieanalyses en correlaties worden veelvuldig in wetenschappelijke publicaties gebruikt om hypothesen te toetsen. Hierin wordt meer informatie gegeven dan wij in de voorgaande hoofdstukken hebben besproken, maar we denken dat met de informatie die je nu hebt, je al een heel eind zult komen bij

het kunnen interpreteren van wetenschappelijke resultaten. Als voorbeeld laten we een paar tabellen zien uit onderzoeken die gepubliceerd zijn in het Tijdschrift voor Communicatiewetenschap.

In een experiment van De Leeuw et al.[7] is bijvoorbeeld gekeken naar de invloed van zogenaamd prosociaal televisienieuws op kinderen. Kinderen in groep 7 en 8 van de basisschool werden verdeeld over een controlegroep en een experimentele groep, waarbij kinderen in de experimentele conditie een nieuwsprogramma te zien kregen waarin geld werd ingezameld voor UNICEF, en kinderen in de controlegroep een nieuwsuitzending te zien kregen die ook over UNICEF ging maar waarin het prosociale gedrag (geld inzamelen) niet werd getoond. Voor het experiment maakten de onderzoekers zelf een nieuwsprogramma met de naam *Newz Kids*. De onderzoekers geven eerst de beschrijvende statistieken van de variabelen door percentages, gemiddelden en standaarddeviaties te laten zien:

Tabel 8.20 Beschrijvende statistieken uit artikel van De Leeuw et al. (2014)

	Totaal (N = 372)	Experimentele conditie (n = 183)	Controleconditie (n =189)
Prevalentie			
Jongen	44.1%	43.7%	44.4%
Lijkt een project voor UNICEF leuk	93.7%	93.3%	94.1%
Bereidwilligheid om een project voor UNICEF op te zetten[1]	80.0%	84.8%	76.9%
Gemiddelde (Standaarddeviatie)			
Leeftijd	10.94 (.76)	10.94 (.76)	10.95 (.75)
Waardering van *NewzKids*	7.39 (1.55)	7.36 (1.53)	7.43 (1.57)
Initieel prosociaal gedrag[2]	2.63 (.34)	2.66 (.32)	2.60 (.36)
Mate waarin ouders goede doelen belangrijk vinden	3.57 (.36)	3.55 (.65)	3.59 (.61)
Donatie voor UNICEF[3]	62.16 (30.58)	64.98 (31.50)	59.41 (29.49)

Hieruit kunnen we onder andere aflezen dat kinderen (met een gemiddelde leeftijd van 10,94, SD = 0,76) uit de experimentele conditie, eerder bereid zijn om een project voor UNICEF op te zetten (en die dus meer prosociaal zijn), dan kinderen in de controleconditie, terwijl de waardering voor het fictieve programma in beide groepen ongeveer even hoog is.

In tabel 8.21 is een correlatiematrix te zien tussen alle variabelen die de onderzoekers hebben gemeten. Je ziet hierin dat in wetenschappelijke tijdschriften geen SPSS-tabellen worden gebruikt, maar dat aangepaste tabellen worden gemaakt. In dit boek zullen we niet ingaan op de sterretjes en kruisjes achter de waarden van de associatiematen, maar we kunnen wel de richting en de sterkte van de correlaties aflezen.

Tabel 8.21 Correlatiematrix uit artikel van De Leeuw et al. (2014).

	1	2	3	4	5	6
1. Sekse[1]						
2. Waardering van *NewzKids*	.27 **					
3. Initieel prosociaal gedrag	.28 **	.24 **				
4. Mate waarin ouders goede doelen belangrijk vinden	.22 **	.17 **	.38 **			
5. Conditie[2]	.01	.02	.10 †	-.03		
6.Donatie voor UNICEF	.06	.03	.13 *	.19 **	.09 †	
7. Bereidwilligheid om een project voor UNICEF op te zetten	.32 **	.23 **	.31**	.20 **	.10 †	.06

[1]0 = jongen; 1 = meisje
[2]0 = controleconditie; 1 = experimentele conditie; * $p < .05$, ** $p < .01$, † $p < .10$.

Hoewel wij in dit boek geen bivariate analyses met dummyvariabelen uitvoeren, is wel goed te zien in tabel 8.21 dat je als onderzoeker moet aangeven wat de codering is bij een dummyvariabele (anders weet de lezer niet wat een positieve of negatieve correlatie bij bijvoorbeeld sekse betekent). Zo is te zien dat meisjes op alle variabelen hoger scoren dan jongens (alle correlaties met sekse zijn positief, en meisjes hebben de waarde 1). We kunnen bijvoorbeeld ook aflezen dat hoe hoger de mate is waarin ouders goede doelen belangrijk vinden, hoe hoger het initiële prosociale gedrag is (r = 0,38), en dat dit verband redelijk sterk en positief is.

In een ander onderzoek dat in het Tijdschrift voor Communicatiewetenschap is gepubliceerd vinden we een regressieanalyse met zowel de waarden voor de correlaties als de gestandaardiseerde regressiecoëfficiënten (tabel 8.22). In dit onderzoek van Slot et al. (2014)[8] is onder andere onderzocht of kinderen in de leeftijd van 9 tot 12 jaar reclame in online sociale netwerken (met de naam *Habbo*) begrijpen en in welke mate zij gevoelig zijn voor de mening van leeftijdsgenoten met betrekking tot merknamen die in deze netwerken gebruikt worden, en of er een verlangen was naar het geadverteerde merk.

Ook bij deze tabel zullen we niet op alle statistieken ingaan (zoals de standaardfouten en de sterretjes achter de waarden), maar we kunnen al veel van de waarden interpreteren. We zien bijvoorbeeld dat meisjes minder verlangen hebben naar de geadverteerde merken dan jongens ($\beta = -0.11$) onder constanthouding van de overige onafhankelijke variabelen. Wanneer gekeken wordt naar het kunnen begrijpen van de persuasieve intentie, valt op dat onder constanthouding van de overige variabelen er bijna geen invloed is op het verlangen (β = 0,09), waar wel een matige samenhang bestond (r = 0,20) wanneer we de andere onafhankelijke variabelen niet in de analyse betrekken. We kunnen ook zien dat alle onafhankelijke variabelen samen voor 36% de variantie in het verlangen naar geadverteerde merken verklaren, en dat de gevoeligheid voor de mening van *peers* met betrekking tot de merken in *Habbo* het sterkste

effect heeft op het verlangen naar merken waar reclame voor wordt gemaakt (β = 0,37), gevolgd door de kritische houding ten opzichte van reclame (hoe kritischer de houding, hoe minder het verlangen, β = –0,34).

Tabel 8.22 Regressieanalyse in het artikel van Slot et al. (2013)

	Verlangen naar geadverteerde merken		
	β	*SE*	*r*
Controlevariabelen			
Leeftijd	-.08	(.07)	-.15
Geslacht (1 = meisjes)	-.11	(.13)	-.10
Speelfrequentie *Habbo*	.04	(.08)	.02
Reclamewijsheid			
Reclameherkenning	-.06	(.09)	-.09
Begrip commerciële bron	-.10	(.12)	-.13
Begrip persuasieve intentie	.09	(.09)	.20*
Kritische houding t.o.v reclame	-.34***	(.08)	-.46***
Gevoeligheid voor *peer*invloed			
*Peer*invloed merken i.h. algemeen	-.03	.11	.08
*Peer*invloed merken in *Habbo*	.37***	(.08)	.50***
N	148		
Totaal R^2 (aangepast)	.36		

β = genormaliseerde bètaregressiecoëfficiënten; *SE* = standaardfouten; *r* = correlaties verlangen naar geadverteerde merk

8.4 Samenvatting

De associatiematen die je gebruikt op interval- en rationiveau zijn de correlatiecoëfficiënt (*r*), de proportie verklaarde variantie (R^2) en de gestandaardiseerde regressiecoëfficiënt (β). Een correlatie geeft aan wat de sterkte en richting is van de samenhang tussen twee variabelen. De proportie verklaarde variantie gebruik je bij een regressieanalyse en geeft aan in welke mate de onafhankelijke variabele(n) de varia(n)tie in de afhankelijke variabele verklaart/verklaren. Bèta's geven het zuivere effect van de onafhankelijke variabele(n) op de afhankelijke variabele aan.

Een regressievergelijking geeft een voorspelling van de afhankelijke variabele *y* op basis van de onafhankelijke variabele(n) *x* (of meerdere *x'en*). Bij een meervoudige regressieanalyse kijk je naar het partiële effect van een onafhankelijke variabele, waarbij je de andere onafhankelijke variabelen constant houdt. Het constant houden van (controleren voor) een derde variabele is ook al eerder aan de orde geweest bij tabelsplitsing (zie hoofdstuk 7).

In meervoudige regressieanalyses kunnen ook dummyvariabelen worden gebruikt, waarbij een nominale variabele (indien nodig) wordt omgezet naar een dichotome variabele met de waarden nul en één. Bij een dichotome variabele

geeft de ongestandaardiseerde regressiecoëfficiënt het gemiddelde verschil tussen de nul- en de één-categorie aan.

Tabel 8.23 Overzicht associatiematen

	Nominaal	Ordinaal	Interval en ratio
Symmetrisch	Cramers V phi	Gamma Kendalls tau-b Spearmans rho	Correlatie (r)
Asymmetrisch	Goodman en Kruskals tau lambda	Somers' d	Proportie verklaarde variantie (R^2) Gestandaardiseerde regressiecoëfficiënt (β)

Ga naar de website om de opdrachten bij dit hoofdstuk te maken.

Noten

1 Leeftijd is hier de onafhankelijke variabele, want kijktijd kan nooit leeftijd beïnvloeden. Een asymmetrische relatie mag echter ook beantwoord worden met een symmetrische associatiemaat, hoewel deze misschien niet altijd het *meest geschikt* zal zijn.
2 We hadden ook kunnen kiezen voor uren tv als *x* en uren krant als *y*.
3 De begrippen variatie en variantie worden hier beide gebruikt om hetzelfde aan te duiden. In sommige publicaties zul je misschien het woord variatie in plaats van variantie zien staan, wij kiezen hier voor de term variantie.
4 Ook bij deze term worden variatie en variantie door elkaar gebruikt, ze duiden hier hetzelfde aan.
5 De *n* kan niet uit bovenstaande tabellen worden afgelezen, deze informatie moet je uit de datamatrix zelf halen.
6 Een regressievergelijking beschrijven kan op verschillende manieren en is minder eenduidig dan bij de vorige associatiematen. Zorg in ieder geval dat de onderdelen intercept, (on)gestandaardiseerde regressiecoëfficiënt, proportie verklaarde variantie, (aantal) onderzoekseenheden en de variabelen worden besproken.
7 De Leeuw, N.H., Rozendaal, E., Kleemans, M., Anschütz, D.J. & Buijzen, M. (2014). 'Prosociaal nieuws en prosociaal gedrag in kinderen', *Tijdschrift voor Communicatiewetenschap (42)*4, 342-357.
8 Slot, N., Rozendaal, E., Van Reijmersdal E.A., & Buijzen, M. (2013). 'Hoe kinderen reageren op reclame in online sociale netwerken: reclamewijsheid en de invloed van leeftijdsgenoten', *Tijdschrift voor Communicatiewetenschap (41)* 1, 19-40.