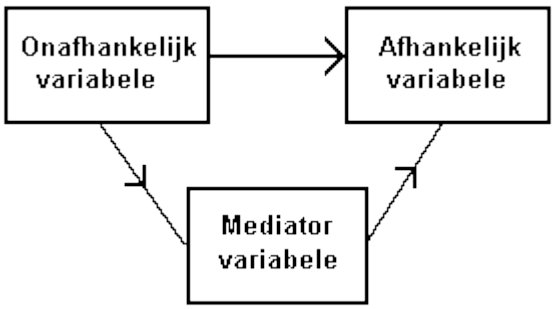
HOOFDSTUK IX

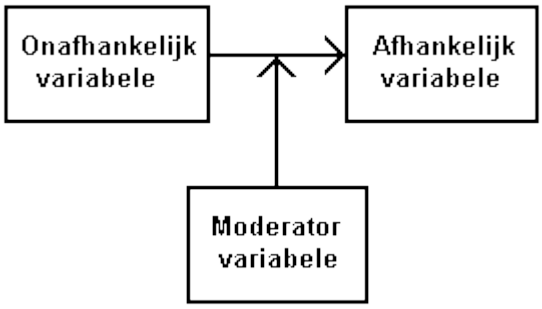
MULTIPLE REGRESSIE ANALYSE

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk begrijpen studenten de meervoudige regressieanalyse en kunnen zij deze zelf toepassen (zelf uitrekenen) in de situatie met twee onafhankelijke variabelen. Studenten begrijpen waarom niet zomaar twee afzonderlijke bivariate analyses bij elkaar kunnen opgeteld worden. Studenten begrijpen dat de lineaire OLS-regressieanalyse niet steeds bruikbaar is en dat bij schendingen van assumpties voor andere methoden dient gekozen te worden. Studenten kunnen de regressiecoëfficiënten inhoudelijk interpreteren. Tot slot kunnen studenten statistische interactie uitleggen aan de hand van een voorbeeld.

2. TE ONTHOUDEN KERNBEGRIPPEN

Confounder	Controlevariabele
Interactie-effect	De sterkte van een relatie tussen twee kenmerken wordt beïnvloed door een derde kenmerk (een moderatorvariabele)
Mediatorvariabele	Statistische variabele die de relatie tussen twee andere variabelen X en Y verklaart 
Meervoudige of multiple regressieanalyse	Statistische techniek waarbij een lineair verband wordt berekend tussen een afhankelijke variabele van het metrische niveau op basis van meerdere onafhankelijke variabelen eveneens van het metrische niveau
Moderatorvariabele	Verandert het effect dat een variabele X heeft op een variabele Y afhankelijk van de waarde van de moderatorvariabele Z



Multicollineariteit	De mate van overlap/correlatie die bestaat tussen de onafhankelijke variabelen van een regressieanalyse. De samenhang tussen de onafhankelijke variabelen moet zo klein mogelijk zijn.

3. STATISTISCHE SYMBOLEN EN FORMULES

Determinatiecoëfficiënt	$R^2 = \beta_1 r_{y1} + \beta_2 r_{y2}$
Gestandaardiseerde regressiegewichten	$\beta_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$ $\beta_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}$
Intercept	a= gemiddelde Y - b1 * gemX1 - b2 * gemX2
Ongestandaardiseerde regressiegewichten	$b_1 = \left(\frac{r_{y,x1} - r_{y,x2} r_{x1,x2}}{1 - (r_{x1,x2})^2}\right)\left(\frac{SD_y}{SD_{x1}}\right)$ $b_2 = \left(\frac{r_{y,x2} - r_{y,x1} r_{x1,x2}}{1 - (r_{x1,x2})^2}\right)\left(\frac{SD_y}{SD_{x2}}\right)$
Regressievergelijking met 2 onafhankelijke variabelen	$Y = a + b_1 X_1 + b_2 X_2 + e$

4. OEFENINGEN

1. *Helpt het als studenten een privébegeleider inschakelen om een hogere eindscore op hun statistiekexamen te behalen ?*

   Op basis van gegevens over 3492 studenten (van wie er 573 een privébegeleider hadden) werd het model E(y)= a + b1*x1 + b2*x2 geschat, waarin y = de eindscore op het examen statistiek, x1= score op een tussentest statistiek en x2= de mate waarin de student een privébegeleider had.

- Het geschatte model heeft een R^2-waarde van 0.76. Interpreteer dit resultaat.

- De schatting voor b2 is 19 met een standaardfout van 3. Gebruik deze informatie om een 95% betrouwbaarheidsinterval voor b2 te construeren. Geef een interpretatie van het gevonden interval.

- Wat kun je, uitgaande van voorgaande, zeggen over het effect van een privébegeleider op de eindscore van het statistiekexamen ?

2. Een onderzoeker wil de impact van sociale bindingen (SOCBIND) en extraverte persoonlijkheid (EXTRAVERT) op tevredenheid op het werk (TEVREDEN) analyseren. In de tabel hieronder vind je de gegevens op deze drie variabelen voor 10 respondenten. De drie variabelen zijn van het metrische meetniveau.

Observaties	SOCBIND	EXTRAVERT	TEVREDEN
1	20	15	20
2	10	30	15
3	4	5	5
4	17	16	20
5	10	14	15
6	11	8	10
7	7	7	8
8	4	4	5
9	15	10	17
10	17	5	17

a. Wat is de afhankelijke variabele? Wat zijn de onafhankelijke variabelen?

b. Bereken met de hand de regressie-coëfficiënten (intercept, ongestandaardiseerde en gestandaardiseerde regressieparameters)

c. Interpreteer elke regressie-coëfficiënt.

d. Hoeveel bedraagt de totale verklaarde variantie in de afhankelijke variabele ?

e. Welke onafhankelijke variabele heeft het sterkste relatieve netto-effect?

3. Onderstaande uitspraken over de meervoudige lineaire regressie zijn FOUT. Leg uit wat er fout is en waarom.

FOUTE UITSPRAKEN	WAT IS FOUT EN WAAROM?
De meervoudige correlatie geeft de verhouding weer van de variatie in de te verklaren variabele dat is gegeven door de verklarende variabelen.	
Eén van de assumpties in meervoudige lineaire regressie is dat de verdeling van elke verklarende variabele normaal moet zijn.	
Een hoge R-kwadraat waarde betekent een causaal verband tussen de onafhankelijke en afhankelijke variabelen.	
Alle onafhankelijke variabelen moeten sterk gecorreleerd zijn met de afhankelijke variabele om een goed regressiemodel te hebben.	

4. Veronderstel onderstaande hypothetische gegevens van 20 ex-gedetineerden.

Respondenten	Aantal arrestaties na gevangenisstraf	Aantal jaren gevangenisstraf	Aantal arrestaties voor gevangenisstraf
1	0	2	4
2	0	3	2
3	1	1	2
4	1	2	3
5	1	3	3
6	1	3	2
7	2	4	3
8	2	2	3
9	2	2	1
10	3	3	2
11	3	3	3
12	3	3	3
13	4	3	4
14	4	4	3
15	4	4	4
16	4	4	5
17	5	4	4
18	6	4	5
19	7	5	5
20	9	4	6

- Definieer een meervoudige lineaire regressie waarin *het aantal arrestaties na gevangenisstraf* wordt gedefinieerd in functie van het *aantal jaren gevangenisstraf* en het *aantal arrestaties voor gevangenisstraf*.

- Bereken de regressiecoëfficiënten.
- Interpreteer elke regressiecoëfficiënt.
- Welke onafhankelijke variabele heeft het sterkste netto-effect?
- Hoeveel bedraagt de verklaringskracht van het model op basis van de twee onafhankelijke variabelen.
- Wat is het verwachte *aantal arrestaties na gevangenisstraf* bij een gevangenisstraf van 4 jaar en 3 arrestaties voor gevangenisstraf?

5. In onderstaande tabel worden de ruwe data weergegeven van 10 atleten. Op basis van deze scores willen we predicties maken van prestaties op basis van uren training en scores op een motivatietest.

Observaties	Prestatie-scores	Uren training	Motivatie-scores
1	67	6	36
2	87	8	32
3	87	8	43
4	56	5	26
5	72	7	31
6	57	5	38
7	60	6	42
8	92	9	48
9	56	5	33
10	67	6	30

a) Wat is de afhankelijke variabele? Wat zijn de onafhankelijke variabelen?

b) Bereken met de hand de regressie-coëfficiënten (intercept, ongestandaardiseerde en gestandaardiseerde regressieparameters)

c) Hoeveel bedraagt de totale verklaarde variantie in de afhankelijke variabele?

d) Wat is de verwachte prestatie bij een training van 7 en een motivatie van 48?

e) Wat is de training bij een verwachte prestatie van 95 en een motivatie van 45?

6. De European Social Survey (ESS) of het Europees Sociaal Onderzoek, is een cross-nationaal onderzoek dat in verschillende Europese landen beoogt attitudes, meningen en gedragspatronen van de bevolking in kaart te brengen. Sinds 2002 wordt de enquête iedere twee jaar georganiseerd. Ook België neemt hieraan deel. Aan 1704 respondenten werd gevraagd *'Hoe tevreden bent u met de democratie in uw land?'* (Y).

   We voerden een meervoudige lineaire regressieanalyse uit met als onafhankelijke variabelen:

   In blok 1 : gender, leeftijd, opleidingsniveau

   In blok 2 : attitude_immigratie, vertrouwen_politieke_instellingen.

   Attitude_immigratie : peilt naar de attitude van de respondent ten aanzien van immigratie. Hoge waarden verwijzen naar een positieve attitude ten aanzien van immigratie

   Vertrouwen_politieke_instellingen : peilt naar de mate van vertrouwen in politieke instellingen in België bij de respondent. Hoge waarden verwijzen naar veel vertrouwen in politieke instellingen in België.

   Hieronder vind je de SPSS-output. Beantwoord de onderstaande vragen.

Descriptive Statistics

	Mean	Std. Deviation	N
hoe tevreden bent u met de werking van de democratie in uw land	5,20	2,215	1.640
gender	,49	,500	1.640
leeftijd	54,0628	18,67611	1.640
Opleidingsniveau	8,13	4,359	1.640
attitude ten aanzien van immigratie	14,6884	5,33924	1.640
Vertrouwen_politieke_instellingen	12,1994	5,89235	1.640

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	,175[a]	,031	,029	2,183	,031	17,206	3	1.636	,000
2	,549[b]	,302	,300	1,854	,271	317,343	2	1.634	,000

a. Predictors: (Constant), Opleidingsniveau, gender, leeftijd

b. Predictors: (Constant), Opleidingsniveau, gender, leeftijd, vertrouwen_politieke_instellingen, attitude_immigratie

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	245,960	3	81,987	17,206	,000
	Residual	7.795,424	1.636	4,765		
	Total	8.041,385	1.639			
2	Regression	2.426,805	5	485.361	141.2535	
	Residual	5.614,579	1634	3.4361		
	Total	8.041,385	1.639			

a. Dependent Variable: hoe tevreden bent u met de werking van de democratie in uw land

b. Predictors: (Constant), Opleidingsniveau, gender, leeftijd

c. Predictors: (Constant), Opleidingsniveau, gender, leeftijd, vertrouwen_politieke_instellingen, attitude_immigratie

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5,206	,215		24,234	,000	4,785	5,627
	gender	,301	,108	,068	2,794	,005	,090	,513
	leeftijd	-,012	,003	-,098	-3,995	,000	-,017	-,006
	Opleidingsniveau	,058	,013	,115	4,663	,000	,034	,083
2	(Constant)	2,335	,224		10,435	,000		
	gender	,167	,092	,038	1,819	,069		
	leeftijd	-,006	,002	-,052	-2,483	,013		
	Opleidingsniveau	-,003	,011	-,005	-,242	,809		
	attitude_immigratie	,076	,010	,183	7,864	,000		
	vertrouwen_politieke_instellingen	,166	,008	,441	19,637	,000		

a. Dependent Variable: hoe tevreden bent u met de werking van de democratie in uw land

- Op basis van hoeveel respondenten is de analyse uitgevoerd ?

- Hoeveel bedraagt de multiple determinatiecoëfficiënt en hoe wordt deze maat geïnterpreteerd ?

- Hoeveel bedraagt de *F*-waarde in blok 2 in de ANOVA-tabel ?

- Is de *F*-waarde statistisch significant op het niveau α = .001 ? (zoek in de F-tabel de kritische F-waarde op gegeven het aantal df in teller en noemer). Interpreteer.

- Welke variabele in blok heeft het sterkste relatieve effect ? Is er sprake van een statistisch significant effect ? Beargumenteer.

- Hoe interpreteer je dit relatieve effect ?

- Welke variabelen hebben in blok 2 geen statistisch significant effect ?

- Geef voor de variabele 'vertrouwen_politieke_instellingen' in blok 2 het 95% betrouwbaarheidsinterval. Interpreteer.

7. In onderstaande tabellen vindt u de output van een multipele regressieanalyse met als afhankelijke variabele '*dagsalaris van een topmodel'* en als onafhankelijke variabelen '*leeftijd', 'jaren ervaring als model'* en '*gepercipieerde schoonheid'* (mate waarin een model door een panel van experten als 'aantrekkelijk' wordt gepercipieerd).
Interpreteer de output.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2
1	,429[a]	,184	,173	14,57213	,184	17,066	3	227

a. Predictors: (Constant), gepercipieerde schoonheid, Jaren ervaring, leeftijd

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.871,964	3	3.623,988	17,066	,000[b]
	Residual	48.202,790	227	212,347		
	Total	59.074,754	230			

a. Dependent Variable: dag-salaris

b. Predictors: (Constant), gepercipieerde schoonheid, Jaren ervaring, leeftijd

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-60,890	16,497		-3,691	,000	-93,396	-28,384
	leeftijd	6,234	1,411	,942	4,418	,000	3,454	9,015
	Jaren ervaring	-5,561	2,122	-,548	-2,621	,009	-9,743	-1,380
	Gepercipieerde schoonheid	-,196	,152	-,083	-1,289	,199	-,497	,104

a. Dependent Variable: dag-salaris

a) Hoeveel bedraagt de multipele correlatiecoëfficiënt? Interpreteer.

b) Hoeveel bedraagt de determinatiecoëfficiënt? Interpreteer.

c) Hoeveel respondenten zijn in de analyse betrokken?

d) Hoeveel bedraagt de toetsstatistiek *F*? Is deze significant ? Interpreteer.

e) Hoeveel bedraagt het intercept en wat betekent dit?

f) Hoeveel bedragen de ongestandaardiseerde richtingscoëfficiënten?

g) Hoeveel bedragen de gestandaardiseerde richtingscoëfficiënten?

h) Welke variabele heeft het sterkste relatieve effect? Welke parameter interpreteer je dat? Waarom?

i) Zijn de regressieparameters significant?

j) Wat betekent het 95% betrouwbaarheidsinterval voor B?

k) Wat is het verwachte dag-salaris van een topmodel van 18jaar, zonder ervaring en met een gepercipieerde schoonheid van 81?