

EBM & Statistiek III

Robby.DePauw@Ugent.be

2021-12-19

Contents

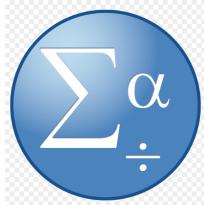
1	Introductie	1
2	Correlaties	1
2.0.1	Pearson correlatie (ρ)	5
2.0.2	Spearman correlatie (r_s)	5
3	Lineare regressie	1
4	Logistische regressie	1
4.0.1	Performantie van het regressiemodel	10
5	Meervoudige regressie	1
6	Voorbeeldexamens	1

Over deze cursus

De cursus EBM & Statistiek III bouwt verder op de leerstof uit EBM & Statistiek I en EBM & Statistiek II. De situering, vereiste begincompetenties en eindcompetenties zijn terug te vinden op de vak-specifieke pagina op Ufora. Binnen deze cursus bespreken we meer geavanceerde statistische modellen. De belangrijkste van deze modellen zijn:

- Bivariate correlaties
- Lineaire regressie
- Logistische regressie
- Mixed models
- Overlevingsanalyses

Binnen deze cursus maken we gebruik van enkele oefeningen in SPSS



Probeer voor de oefeningen versie 26 van SPSS te gebruiken indien deze beschikbaar is op Athena. Er kan ook steeds een stand-alone versie van SPSS aangevraagd worden via volgende link. **Opgelet:** de website met instructies om SPSS te installeren op jouw computer is alleen beschikbaar vanuit het UGent netwerk of via VPN.

De lesgevers

De lessen binnen het vak worden georganiseerd door *dr. Robby De Pauw* en *prof. dr. Pascal Coorevits*. Vragen kunnen steeds gepost worden in de discussieruimtes op Ufora.

Voor vragen over het stuk statistiek kunnen jullie steeds mailen naar Robby.DePauw@Ugent.be, voor vragen over het stuk datamanegement kunnen jullie mailen naar Pascal.Coorevits@Ugent.be.

Uitprintversie boek

Bovenaan de balk van deze website kunnen jullie een pdf-versie van dit boek downloaden.

```
knitr::asis_output('\\\\mainmatter\\n\\n')
```

Chapter 1

Introductie

In de huidige cursus EBM & Statistiek III wordt er verder gebouwd op de kennis die opgedaan is tijdens de lessen van EBM & Statistiek I en EBM & Statistiek II. Binnen deze lessen werden concepten zoals **steekproef**, **beschrijvende statistiek** en **inferentiële statistiek**. Alle kennis en vaardigheden die jullie binnen de leerlijn **EBM en Statistiek** opbouwen, moet jullie in staat stellen om zelfstandig een onderzoeksproces van het begin tot het einde te begrijpen, te beoordelen op kwaliteit en te vertalen naar de klinische praktijk.

Een onderzoekproces bestaat uit verschillende belangrijke stappen die nodig zijn om tot een duidelijk en rechtlijnige conclusie te komen. Enkele basisbegrippen die noodzakelijk zijn om te begrijpen zijn *hypothese*, *fouten* bij hypothesetesten en statistische *toetsen*. Hieronder bespreken we kort deze belangrijke concepten opnieuw. Voor meer informatie verwijzen we naar de cursussen uit EBM & Statistiek I en EBM & Statistiek II.

Hypothesetoetsen

Een hypothesetoets bestaat uit twee delen, de nulhypothese H_0 en de alternatieve hypothese H_1 of H_a . Het uitgangspunt van een hypothesetoets is vaak H_0 , waarbij we veronderstellen dat een effect (uitgedrukt als een verschil, associatie, ...) niet bestaat. Na het opstellen van een hypothese en de vertaalslag van deze hypothese naar H_a en H_0 , wordt er data verzameld binnen een steekproef. Op basis van deze verzamelde data proberen we H_0 te verwerpen in het voordeel van H_a . De beslissing om H_0 te verwerpen zal uiteindelijk gebeuren op basis van de *p*-waarde. Als cut-off wordt hier vaak 5% (of 0.05) genomen. We noemen deze cut-off ook wel het significantieniveau of α .

Aangenomen wordt dat geen effect geassocieerd wordt met H_0 , wat niet betekent dat er geen alternatieve mogelijkheden zijn. We kunnen bijvoorbeeld ook stellen

dat we onder H_0 veronderstellen dat een effect een specifiek getal moet zijn (bijvoorbeeld dat het verschil tussen twee groepen exact 2 moet bedragen).

Fouten bij hypothesetoetsen

Op basis van een hypothesetoets die we uitvoeren op basis van verzamelde data uit een steekproef van een bepaalde populatie proberen we een besluit te vormen over de hypothese binnen de populatie. Aangezien een steekproef en metingen gepaard gaan met variabiliteit en onzekerheid, kunnen er fouten optreden. In onderstaande tabel vinden jullie een grafische weergave van deze mogelijke fouten:

Null hypothesis is...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β

Figure 1.1: Visuele samenvatting van fouten bij hypothesetoetsen.

Voorbeeld: Veronderstel dat er binnen de populatie van kinesitherapeuten geen verschil is in het aantal uren die gepresteerd worden tussen mannen en vrouwen. We organiseren binnen deze populatie ($n = 30$) en evaluaeren het aantal uren. Uiteindelijk voeren we een statistische toets uit, waaruit we finaal besluiten of er al dan geen verschil is in gepresteerde uren. Hiervoor veronderstellen we volgende hypothesen:

- H_0 : Er is geen verschil in gepresteerde uren tussen mannen en vrouwen.
- H_1 : Er is een verschil in gepresteerde uren tussen mannen en vrouwen.

Indien we binnen de populatie weten dat er **geen** verschil is (H_0 is dus juist), zijn er twee mogelijke uitkomsten:

- We verwerpen H_0 en besluiten dus dat er wel een verschil is. In dit scenario maken we een fout en binnen de inductieve statistiek wordt deze fout

aangeduid als α (Type I error). Algemeen wordt aanvaard om deze α op 5% te zetten.

- We aanvaarden H_0 en besluiten dus dat er wel een verschil is. In dit scenario maken we een correcte beslissing.

Indien we binnen de populatie weten dat er **wel** verschil is (H_0 is dus fout), zijn er twee mogelijke uitkomsten:

- We verwerpen H_0 en besluiten dus dat er wel een verschil is. In dit scenario maken we een correcte beslissing.
- We aanvaarden H_0 en besluiten dus dat er geen verschil is. In dit scenario maken we een fout en binnen de inductieve statistiek wordt deze fout aangeduid als β (Type II error). Over de grootte van β is er minder consensus, maar 80% wordt vaak algemeen aanvaard.

In bovenstaande figuur 1.1 kan je de **vier** verschillende scenario's zoals hierboven beschreven herkennen.

Keuze statistische toets

In figuur 1.2 vinden jullie een schema met de specifieke keuze van statistische toets op basis van de gestelde voorwaarden.

Aantal steekproeven	Aard variabele	Gepaard / Ongepaard	Parametrisch / niet-parametrisch	Test
1	Categorisch			Chi-kwadraat test
	Continu		Parametrisch	z-test of t-test
	Continu/ordinaal		Niet-parametrisch	Wilcoxon signed-rank test
2	Dichotoom	Gepaard		McNemar test
	Categorisch	Ongepaard		Chi-kwadraat test of Fisher's Exact test
	Continu	Gepaard	Parametrisch	Gepaarde t-test
	Continu/ordinaal	Gepaard	Niet-parametrisch	Wilcoxon matched-pairs signed-rank test
	Continu	Ongepaard	Parametrisch	Ongepaarde t-test
	Continu/ordinaal	Ongepaard	Niet-parametrisch	Mann-Whitney U-test
> 2	Categorisch	Ongepaard		Chi-kwadraat test of Fisher's Exact test
	Continu/ordinaal	Gepaard	Niet-parametrisch	Friedman test
	Continu	Ongepaard	Parametrisch	One-way ANOVA
	Continu/ordinaal	Ongepaard	Niet-parametrisch	Kruskal-Wallis test

Figure 1.2: Keuze van statistische toetsen.

Studie design

Er zijn verschillen manieren waarop een studie kan worden uitgevoerd. Bij wijze van voorbeeld nemen we een studie waarbij we 100 sporters includeren ($n = 100$) waarbij we willen nagaan of lenigheid een risicofactor kan zijn voor het oplopen van een blessure. Bij een cross-sectionele studie voeren we de volledige studie uit op één tijdspunt, m.a.w. is er geen opvolging voorzien van de sporters. Hierbij is het dus belangrijk om al een idee te hebben van welke sporters een blessure hebben opgelopen en welke niet. In een cross-sectionele studie zal de onderzoeker dus beslissen om een aantal geblesseerde sporters ($n = 50$) en niet geblesseerde sporters ($n = 50$) te includeren, waarbij men vaak kiest voor een gelijke verdeling van het aantal sporters per groep. Binnen een cross-sectionele studie is het uiteindelijke doel om de lenigheid van de geblesseerde groep te vergelijken met de lenigheid van de niet-geblesseerde groep. Bij een longitudinale cohorte is er wel een opvolgperiode voorzien. Wanneer we een longitudinale studie uitvoeren, dan includeren we eerst 100 niet-geblesseerde sporters, die we opvolgen voor een specifieke follow-up periode (bijvoorbeeld 1 jaar). Tijdens een eerste testmoment (vaak baseline genoemd) evalueren we de lenigheid bij alle geïncludeerde sporters. Tijdens de opvolgperiode houden we bij welke sporters geblesseerd raken en welke niet. Op het einde van de studie kunnen we de baseline-lenigheid vergelijken tussen sporters die zich al dan niet blesseerden. Elk van deze specifieke studie-designs hebben zowel voordelen als nadelen. Zo staat een longitudinale prospetieve cohorte hoger aangeschreven in de evidentiepyramide, aangezien er een kleinere kans op confounding is bij deze studies. Anderzijds zijn studies met een opvolgperiode aanzienlijk duurder en moeilijker op uit te voeren.

Epidemiologie

Om een aandoening epidemiologisch te beschrijven bestaan er twee veelgebruikte indicatoren, namelijk de **incidentie** en **prevalentie**. De **prevalentie** is een inschatting van het aantal individuen die op een gedefinieerd tijdspunt lijden aan een bepaalde aandoening. Wanneer we de **prevalentie** willen bepalen van blessures bij 100 sporters bij de start van 2011, bekijken we het aantal sporters die leiden aan een blessure zoals visueel weergegeven in figuur 1.3. Hierbij merken we op dat er 5 sporters een blessure hebben en de geschatte prevalentie dus $\frac{5}{100}$ of 5% zijn. De **incidentie** is een inschatting van het aantal nieuwe individuen die leiden aan de aandoening over een welgedefinieerde periode. Een belangrijk verschil met de **prevalentie** is dat we steeds starten met een groep van sporters die aan het begin van de periode waarbij we de **incidentie** gaan berekenen. Wanneer we de **incidentie** willen bepalen van blessures in het jaar 2011 starten we van de groep zonder de aandoening ($100 - 5 = 95$). In de groep van 95 sporters zonder blessure, zijn er 6 die een blessure hebben opgelopen. De incidentie in 2011 wordt dus geschat op $\frac{6}{95}$ of 6%.

Om een inschatting te maken van de **prevalentie** van een aandoening volstaat informatie over het al dan niet aanwezig zijn van een aandoening op een bepaald tijdstip (cross-sectionele studie). Om een inschatting te maken van de **incidentie** van een aandoening is een follow-up periode noodzakelijk (prospectieve follow-up cohorte).

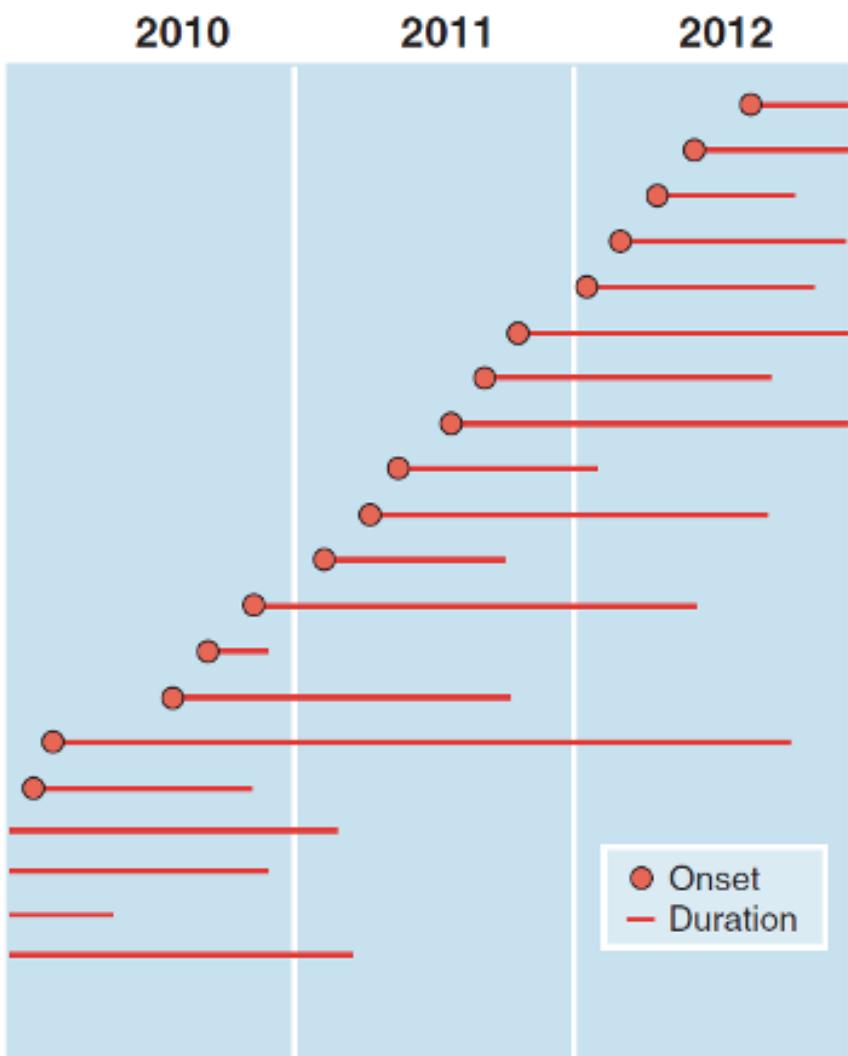


Figure 1.3: Incidentie en prevalentie.

Chapter 2

Correlaties

In de paper van Al-Hadidi et al. (2019) vermeld men volgende stelling:

“A significant positive correlation was also found between the duration of use for studying and the duration of pain ($p < 0.001$, $r = 0.212$).”

—Al-Hadidi et al. (2019)

- Kunnen we hieruit besluiten dat verhoogd smartphonegebruik de oorzaak is van langdurigere nekpijnervaringen?
- Hoe interpreteren we de r en p -waarde? Kunnen we hieruit besluiten dat er een sterke link is?

Inleiding

Er zijn verschillende manieren om een associatie tussen twee variabelen in te schatten. Eén van de meest gebruikte statistische maten om een verband tussen twee variabelen aan te tonen is een correlatie. Er is er een uitgebreid gamma aan correlaties beschikbaar die gebruikt worden om een verband tussen verschillende specifieke types variabelen aan te tonen. Binnen deze cursus focussen we op correlatiecoëfficiënten die het verband tussen twee **continue** variabelen trachten in te schatten.

Belangrijk om te onthouden is dat correlaties nooit een oorzaak-gevolg relatie kunnen inschatten binnen een cross-sectionele studie. Oorzaak-gevolg relaties (of causale relaties) kunnen enkel aangetoond worden bij gerandomiseerde studies (vb. RCT). Dit is één van de voornaamste redenen waarom een RCT bovenaan de evidentietafballen terug te vinden is (Figuur 2.1). Binnen een niet-gerandomiseerde studie kan een schatting van de correlatie vervuild/verstoord

Table 2.1: Scores op wiskundetoets bij studenten na behalen diploma per geslacht.

Gender	Geslaagd	Niet Geslaagd
Man	40	10
Vrouw	20	10

zijn door onderliggen de factoren die een belangrijke rol spelen, maar niet in rekening werden gebracht of niet geobserveerd werden.

Levels of Evidence



Figure 2.1: Evidentiepiramide.

Volgend voorbeeld verduidelijkt het probleem van vervuiling van een associatie. In tabel 2.1, worden de resultaten weergegeven van deze studie. Zoals aangegeven in deze tabel is het slaagpercentage 80% binnen de groep van mannen en is het slaagpercentage 67% binnen de groep van vrouwen. Hieruit zouden we kunnen besluiten dat er significant minder vrouwen slagen op een wiskundetoets in vergelijking met mannen.

Wanneer we echter kijken naar de slaagpercentages per studie-achtergrond (Tabel 2.1), dan blijkt het slaagpercentage in elke groep hetzelfde te zijn. Binnen de groep van wiskundestudenten slaagt iedereen (100%) en binnen de

Table 2.2: Scores op wiskundetoets bij studenten na behalen diploma per geslacht en achtergrond.

Faculty	Gender	Geslaagd	Niet Geslaagd
Wiskunde	Man	30	0
Wiskunde	Vrouw	10	0
Psychologie	Man	10	10
Psychologie	Vrouw	10	10

groep van studenten met een achtergrond binnen de psychologie slaagt 50%. De vervuiling van de associatie, waardoor het lijkt dat de slaagpercentages hoger zijn bij mannen in vergelijking met vrouwen, treedt op wanneer we geen rekening houden met de studie-achtergrond van de studenten. Dit is een type-voorbeeld van het optreden van **confounding** binnen een studie.

Een correlatie geeft binnen een cross-sectionele studie nooit een causaal verband weer. Er kunnen steeds andere factoren onrechtstreeks betrokken zijn die ertoe leiden dat er een correlatie tussen twee variabelen ontstaat, zonder dat dit verband effectief aanwezig is.

Soorten correlaties

Een veelgebruikte maat voor het beschrijven van een associatie tussen twee variabelen is een correlatie. Een correlatie is een schatting die de sterkte van het verband weergeeft tussen twee variabelen en de kan variëren tussen -1 en 1 ($\rho \in [-1, 1]$). Wanneer een correlatie dicht bij 0 ligt, wijst dit om geen associatie tussen twee variabelen, terwijl een correlatie dicht bij ± 1 , wijst op een sterk verband tussen twee variabelen. In dedze cursus worden twee correlaties besproken, de Pearson correlatie coëfficiënt (ρ) en de Spearman correlatie coëfficiënt (r_s), die elk een specifiek verband weergeven.

In figuur 2.2, staan een aantal correlaties weergegeven. In deze figuur worden een aantal spreidingsdiagrammen weergegeven waarbij de relatie tussen X en Y steeds verschillend is. De punten op de grafiek geven de effectieve waarden weer voor elke observatie, die men kan weergeven als punt (x_i, y_i) . De blauwe lijn op de grafiek geeft steeds de best passende rechte weer door de puntenwolk. Het valt op dat elke Spearman correlatie (r_s) gelijk is aan 1 of -1 en de Pearson correlatie varieert. Dit fenomeen kan je als volgt interpreteren: een Pearson correlatie het lineaire of rechtlijnige verband nagaat tussen twee variabelen, terwijl de Spearman correlatie nagaat of het om een monotone relatie gaat. Een monotone relatie is een relatie waarbij bij een toename van X , ofwel Y altijd toeneemt ofwel Y altijd afneemt. De eerste 4 figuren geven een positief verband weer en ook een positieve correlatie, waarbij een toename in X samengaat met

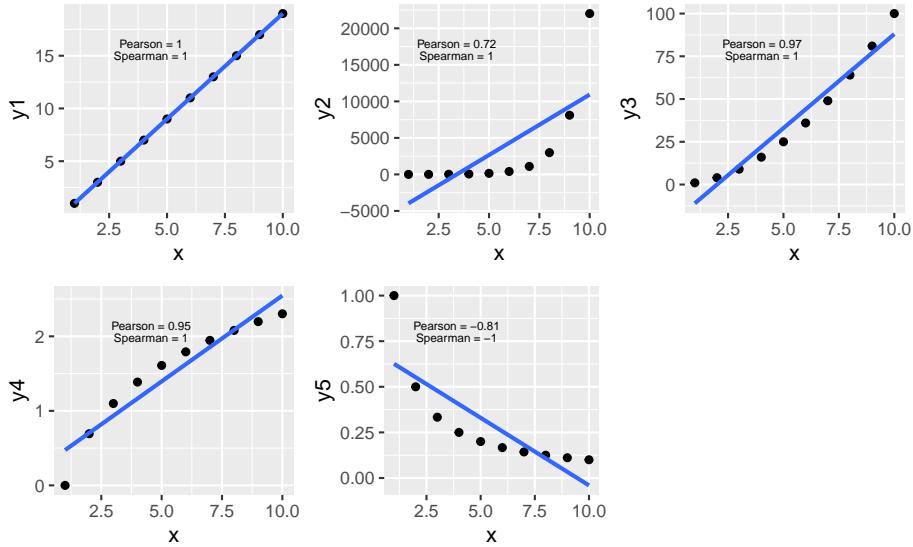


Figure 2.2: Different types of associations and their respective correlations

Table 2.3: Overzicht van karakteristieken en eigenschappen van correlaties.

Karakteristieken	Pearson correlatie	Spearman correlatie
Voorwaarde	Normaal verdeling beide variabelen	Geen
Soort relatie	Lineaire relatie	Monotone relatie
Mogelijke waarden	$[-1, 1]$	$[-1, 1]$
Formule	Gebaseerd op de (co)variantie	Gebaseerd op de rank

een toename in Y . De laatste figuur geeft een negatieve relatie weer, waarbij een toename in X samengaat met een afname in Y .

In tabel @ref(tab: corrtab) staan alle eigenschappen van beide correlatiecoëfficiënten vermeld. De formules voor beide correlaties worden hieronder weergegeven.

Interpretatie: Een correlatie die dichter bij 1 of -1 ligt, geeft aan dat er een sterker lineair of monotoon verband is. De observaties liggen in dit geval meer op één lijn. Bij een correlatie < 0 zal bij een toename van X de waarde van Y dalen, terwijl bij een correlatie > 0 een toename van X gelijklopen met een toename Y .

Table 2.4: Steekproef naar tevredenheid bij kinesitherapeuten.

X	Y
25	8
34	6
34	7
36	5
42	3
44	4
45	4
50	2
62	0

2.0.1 Pearson correlatie (ρ)

Bij de formule voor ρ , vinden we in de teller kenmerken terug van de formule voor covariantie en in de noemer kenmerken voor de variantie.

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

2.0.2 Spearman correlatie (r_s)

Bij de formule voor r_s , vinden we in de teller de rang terug van de verschillende correlaties en in de noemer kenmerken voor de steekproefgrootte.

$$r_s = \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Studie naar tevredenheid

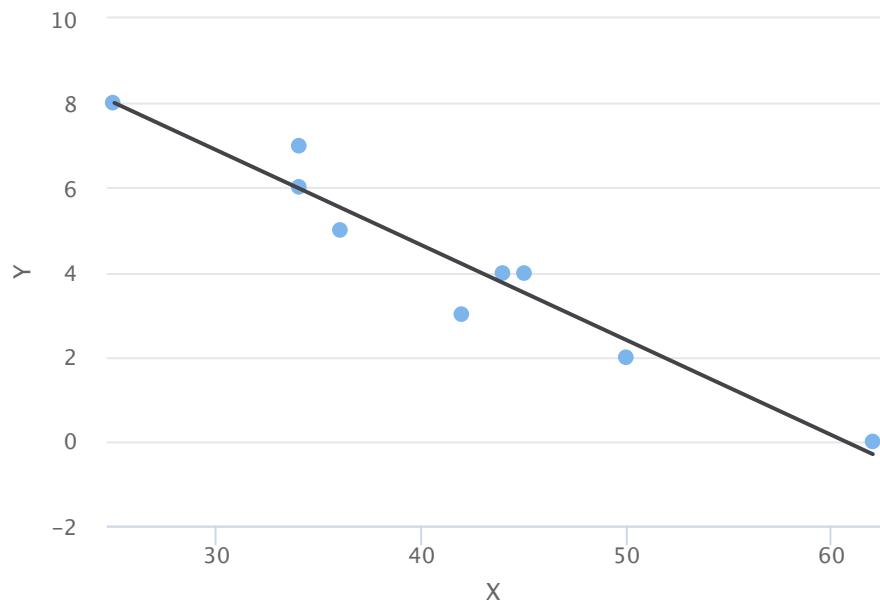
Gegeven is de volgende dataset **tevreden** met informatie over een bevraging bij kinesitherapeuten over hun tevredenheid van honoraria.

- Y : tevredenheid (schaal 0-10)
- X : leeftijd (jaren)

We beschikken in deze specifieke steekproef over een totaal van 9 observaties ($n = 9$).

de dataset ziet er als volgt uit:

Als we de relatie visueel willen weergeven tussen de leeftijd en tevredenheid over de honoraria, kunnen we gebruik maken van een spreidingsdiagram, welk er als volgt uitziet:



De hypothese kunnen we als volgt opstellen:

- H_0 : Er is geen correlatie (ρ of $r_s = 0$) tussen de leeftijd van de therapeut en de gegeven tevredenheidsscore.
- H_1 : Er is een correlatie (ρ of $r_s \neq 0$) tussen de leeftijd van de therapeut en de gegeven tevredenheidsscore.

Een significante correlatie, wil dus zeggen significant verschillend van 0. Dit geeft **geen** indicatie over hoe sterk de relatie is! In figuur 2.3 kan je een voorbeeld van een indeling voor de sterkte van correlaties.

SPSS

In dit hoofdstuk maken we gebruik van de dataset **tevredenheid**, waarin $n = 9$ kinesitherapeuten werden bevraagd.

Stap 1: Voer de data in in de **dataview** van SPSS.

Stap 2: Hierna kunnen we de correlaties binnen SPSS laten berekenen via **Analyze > Correlate > Bivariate**.

Table. Example of a Conventional Approach to Interpreting a Correlation Coefficient	
Absolute Magnitude of the Observed Correlation Coefficient	Interpretation
0.00–0.10	Negligible correlation
0.10–0.39	Weak correlation
0.40–0.69	Moderate correlation
0.70–0.89	Strong correlation
0.90–1.00	Very strong correlation

Several stratifications (with different cutoff points) have been previously published.

Figure 2.3: Correlatiesterkte

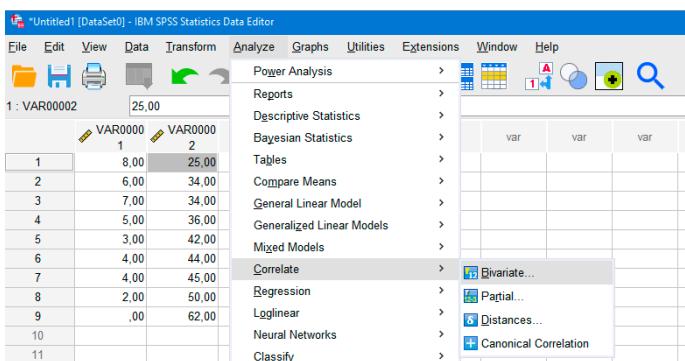


Figure 2.4: SPSS

Stap 3: Hierna krijgen we de pop-up zoals weergegeven in figuur 2.5, waarin we zowel Pearson als Spearman kunnen aanduiden en de verschillend evariabelen waartussen we een correlatie wensen te berekenen. We dienen hierbij minstens 2 variabelen aan de duiden, maar kunnen meerdere variabelen toevoegen. SPSS zal automatisch alle 2x2 correlaties berekenen en weergeven in een matrix met op de diagonaal $\rho = 1$ or $r_s = 1$.

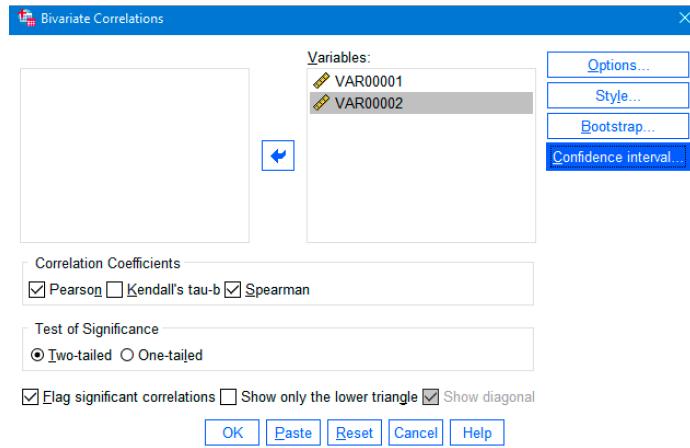


Figure 2.5: SPSS

Stap 4: Vergeet niet om de Syntax te gebruiken binnen SPSS.

```

DATASET ACTIVATE DataSet0.
CORRELATIONS
/VARIABLES=VAR00001 VAR00002
/PRINT=TWOTAIL NOSIG FULL
/CI(LEVEL95)
/MISSING=PAIRWISE.
NONPAR CORR
/VARIABLES=VAR00001 VAR00002
/PRINT=SPEARMAN TWOTAIL NOSIG FULL
/CI(METHOD(FP)) CILEVEL(95)
/MISSING=PAIRWISE.

```

Figure 2.6: SPSS

Stap 5: Tot slot kunnen we de correlaties bekijken en interpreteren. In figuur 2.7 vinden we de resultaten voor ρ , terwijl we in figuur 2.8 de resultaten voor r_s . We vinden in deze figuren ook de p -waarde terug en een 95% betrouwbaarheidsinterval (95%CI).

In figuur 2.6 zijn de resultaten weergegeven wanneer we de Pearson correlatie

berekenen. In figuur 2.7 zijn de resultaten weergegeven wanneer we de Spearman correlatie berekenen.

		Correlations	
		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	-.967**
	Sig. (2-tailed)		<.001
	N	9	9
VAR00002	Pearson Correlation	-.967**	1
	Sig. (2-tailed)	<.001	
	N	9	9

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

Confidence Intervals				
Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed) ^a		
		Lower	Upper	
VAR00001 - VAR00002	-.967	<.001	-.993	-.845

a. Estimation is based on Fisher's r-to-z transformation.

Figure 2.7: SPSS

		Correlations		
		VAR00001	VAR00002	
Spearman's rho	VAR00001	Correlation Coefficient	1,000	-,941 **
		Sig. (2-tailed)	.	<,001
		N	9	9
VAR00002	VAR00002	Correlation Coefficient	-,941 **	1,000
		Sig. (2-tailed)	<,001	.
		N	9	9

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

Confidence Intervals of Spearman's rho				
Spearman's rho	Significance (2-tailed)	95% Confidence Intervals (2-tailed) ^{a,b}		Upper
		Lower	Upper	
VAR00001 - VAR00002	-,941	<,001	-,988	-,728

a. Estimation is based on Fisher's r-to-z transformation.
b. Estimation of standard error is based on the formula proposed by Fieller, Hartley, and Pearson.

Figure 2.8: SPSS

Oefeningen

In dit onderdeel vinden jullie alle oefeningen over correlaties en associatiaties. De oplossingen worden later gepost op **Ufora**.

Exercise 2.1. Binnen een onderzoek naar het oplopen van letsel bij sporters, zijn onderzoekers op zoek gegaan naar mogelijke risicofactoren. Ze includeerde 10 sporters, allemaal met een leeftijd van 15 jaar en noteerden de volgende aantal letsels: 5, 0, 0, 2, 2, 3, 1, 2, 1, 0. De onderzoekers waren geïnteresseerd in het verband tussen leeftijd en het aantal letsels. Welke van onderstaande stellingen is WAAR:

- Met een correlatie kunnen we niet aantonen wat een risicofactor is en wat niet.
- We kunnen binnen deze studie geen correlatie berekenen.
- Binnen deze studie kunnen we het verband nagaan tussen aantal letsels en leeftijd door een Pearson correlatie te berekenen.
- Binnen deze studie kunnen we het verband nagaan tussen aantal letsels en leeftijd door een Spearman correlatie te berekenen.

Exercise 2.2. Een correlatie kunnen we steeds weergeven aan de hand van een lijn.

- Waar
- Onwaar

Table 2.5: Steekproef naar TUG bij opuderen.

Aantal keer gevallen	TUG (s)
0	9.69
4	17.38
2	13.87
3	15.40
6	21.25
1	6.69
8	22.65
1	13.79
1	11.53
0	14.32
0	7.29
1	10.82
1	11.95
0	13.77
4	20.01

De volgende oefeningen zijn gebaseerd op de volgende klinische studie:

In een cross-sectionele studie wordt er onderzoek gedaan naar het verband tussen valrisico bij ouderen en de Timed Up and Go (TUG) test. In het totaal zijn er 15 deelnemers binnen de studie waarbij de TUG test wordt afgenummerd en de tijd (in s) wordt opgenomen. Verder wordt er bij elke deelnemer ook gevraagd hoeveel ze gevallen zijn het afgelopen jaar.

In onderstaande tabel 2.5 vinden jullie de resultaten van deze studie. Voer deze gegevens ook in binnen de DATA omgeving van SPSS en gebruik het programma waar nodig.

Exercise 2.3. Binnen deze studie zullen we kunnen aantonen dat de tijd voor het uitvoeren van de TUG al dan niet een oorzaak is voor meer/minder vallen binnen de oudere populatie.

- Waar
- Onwaar

Exercise 2.4. Welke grafiek past het best voor een visualisatie van deze gegevens?

Exercise 2.5. Welke correlatie zouden jullie selecteren voor het berekenen van een correlatie en waarom?

Exercise 2.6. Geef de schatting weer van de door jou geselecteerde correlatiecoëfficiënt tot 2 cijfers na de komma en geef een correcte interpretatie aan deze correlatie.

Conclusies

- Correlatie \neq een causaal verband.
- Een correlatie kan gebruikt worden om de sterkte van een lineair ρ of monotoon r_s verband uit te drukken.
- De p -waarde van een correlatie zegt niets over de sterkte van de correlatie.

Chapter 3

Lineare regressie

In de paper van Chester et al. (2016) vermeld men volgende stelling:

“At 6 weeks only, a better outcome for both measures was associated with no previous compared to a previous major operation (shoulder surgery excluded) (SPADI, $\beta = -3.66$ to -12.56).”

—Chester et al. (2016)

- Kunnen we hieruit besluiten dat er een statistisch significant verschil is in uitkomst na kinesitherapie tussen patiënten met een vroegere operatie vs geen operatie?

In dezelfde studie lezen we:

“There was no significant difference in mean age or sex between consenters (57 years, SD = 15, 44% male) and non-consenters (56 years, SD = 16, 47% male).”

—Chester et al. (2016)

- Kunnen we hieruit besluiten dat er geen baseline verschil is tussen mensen die deelnamen aan de volledige studie vs mensen die vroegtijdig de studie hebben verlaten?

Inleiding

Lineaire regressie kan zowel toegepast worden in een cross-sectioneel design als prospectieve cohorte en wordt vaak gebruikt voor volgende doelstellingen:

- Het voorspellen van een uitkomstmaat op basis van één of meerdere factoren.
- Het beschrijven van een verband tussen de uitkomstmaat en andere factoren, waarbij rekening wordt gehouden met verstoorende variabelen.

In essentie is een lineair regressie model opgebouwd uit twee componenten:

- De afhankelijke uitkomstvariabele (Y), welke een continue variabele dient te zijn.
- De onafhankelijke variabele(n) (X), welke zowel continu als categorisch van aard kunnen zijn.

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Wanneer er slechts één onafhankelijke variabele X_1 spreken we van een enkelvoudig lineair regressiemodel. Indien er meerdere onafhankelijke variabelen $X_1, X_2, X_3, \dots, X_i$ spreken we van een meervoudig lineair regressiemodel.

Een regressiemodel bestaat uit twee delen, een afhankelijk deel (de uitkomst of wat er geschat moet worden) en een onafhankelijk deel (de verklarende variabelen).

Studie naar tevredenheid

Gegeven is de volgende dataset **tevreden** met informatie over een bevraging bij kinesitherapeuten over hun tevredenheid van honoraria.

- Y : tevredenheid (schaal 0-10)
- X : leeftijd (jaren)
- Z : geslacht (M/V)

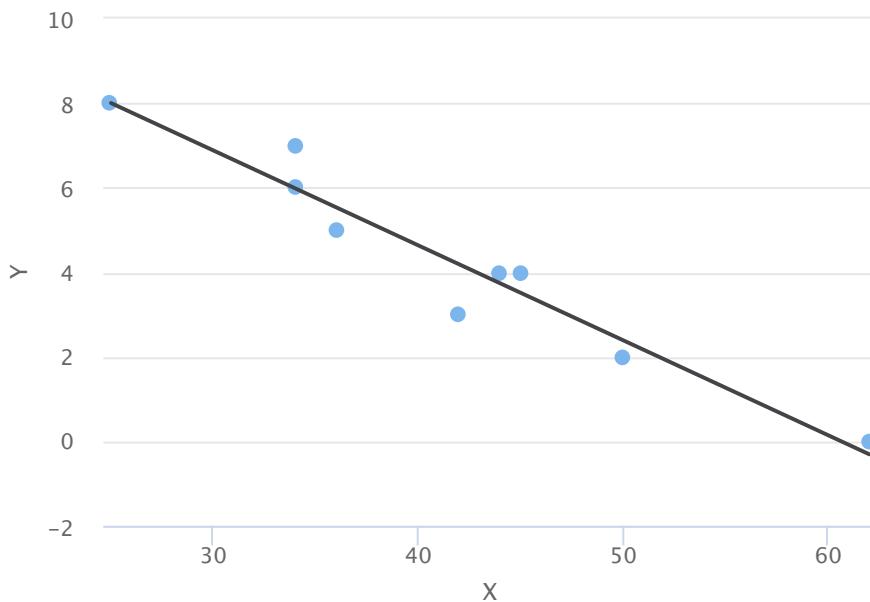
We beschikken in deze specifieke steekproef over een totaal van 9 observaties ($n = 9$).

de dataset ziet er als volgt uit:

Als we de relatie visueel willen weergeven tussen de leeftijd en tevredenheid over de honoraria, kunnen we gebruik maken van een spreidingsdiagram, welk er als volgt uitziet:

Table 3.1: Steekproef naar tevredenheid bij kinesitherapeuten.

X	Y	Z
25	8	M
34	6	M
34	7	V
36	5	V
42	3	V
44	4	M
45	4	V
50	2	V
62	0	M



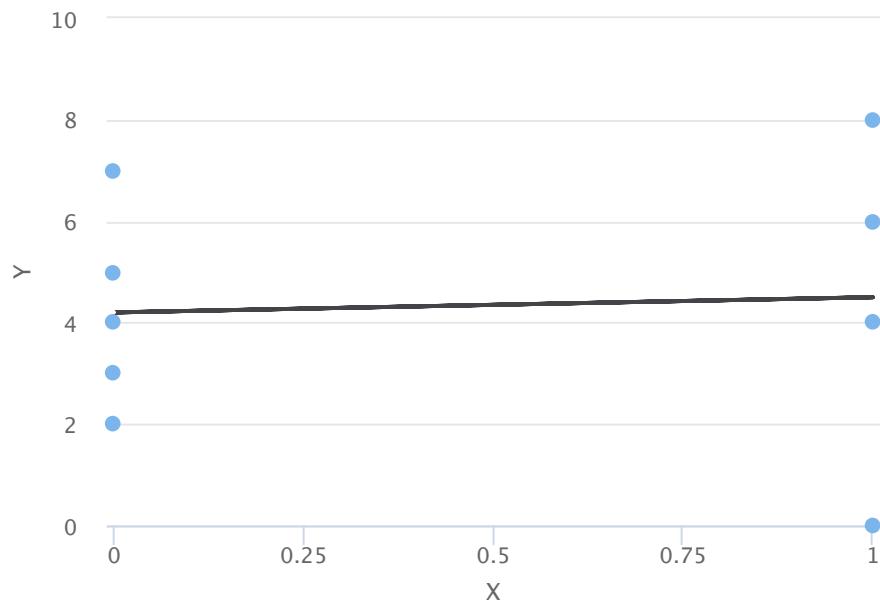
Lineaire regressie geeft net als ρ het lineaire verband weer (in tegenstelling tot r_s , waarbij het monotone verband geschat wordt). Op basis van X en Y uit de **tevredendataset** werd volgend regressiemodel geschat:

$$Y = 13.6 + (-0.2)X$$

Interpretatie: Wanneer de leeftijd X toeneemt met één jaar, dan daalt (−) de tevredenheidsscore die gegeven werd door de kinesitherapeuten gemiddeld met 0.2 (β_1). De interpretatie van het getal 13.6 β_0 heeft vaak geen reële betekenis. In deze studie geeft de waarde 13.6 weer wat de gemiddelde score zou zijn voor kinesitherapeuten met een leeftijd van 0 jaar.

Table 3.2: Gemiddelde tevredenheid bij kinesitherapeuten per geslacht.

	<i>z</i>	<i>y</i>
M	4.5	
V	4.2	



Op basis van Z en Y uit de `tevredendataset` werd volgend regressiemodel geschat:

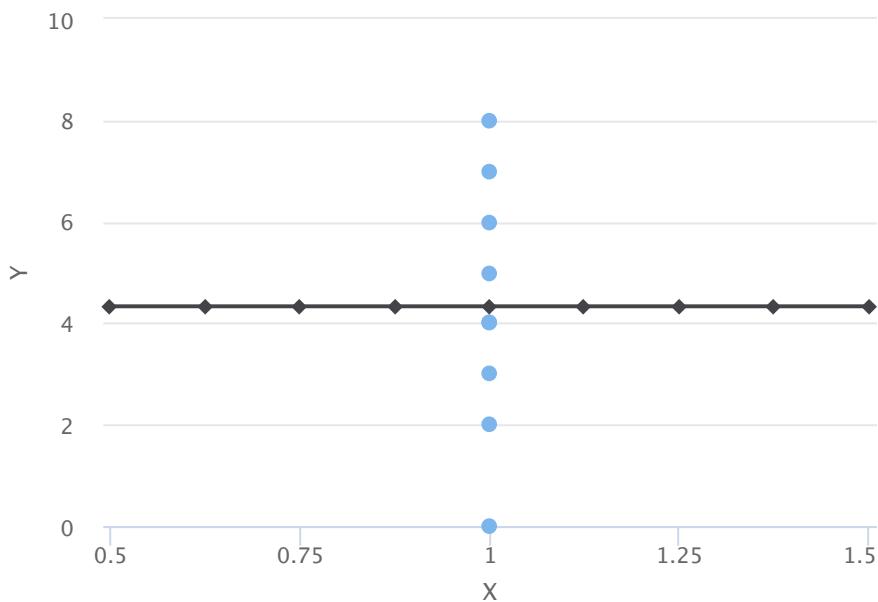
$$Y = 4.5 + (-0.3)Z$$

Interpretatie: Wanneer we de gemiddelde tevredenheid uitrekenen voor mannen en vrouwen bekomen we volgend resultaat zoals weergegeven in tabel 3.2. Om tot een numerieke oplossing te komen werden de `labels` gehercodeerd naar 0 voor V en 1 voor M . De waarde van β_1 bedraagt nu -0.3, wat het gemiddelde verschil weergeeft tussen mannen en vrouwen voor tevredenheidscore. β_0 geeft in dit geval de score weer voor $Z = 0$ (score gegiveen door vrouwen).

Eigenschappen van een lineair regressiemodel

Schatten van lineaire regressie

Eén van de belangrijkste evaluatiecriteria bij een lineair regressiemodel is de meerwaarde van X in het verklaren van Y . Indien we geen informatie zouden hebben over leeftijd en geslacht, kunnen we Y het best beschrijven aan de hand van het gemiddelde \bar{Y} .



Om de waarde van de X variabele in te schatten in het verklaren van Y , kijken we naar de afstand tussen de geschatte regressielijn en de observaties (punten op de grafiek). Wanneer de punten dichter tegen de lijn liggen in vergelijking met een schatting op basis van het algemene gemiddelde \bar{Y} , dan lijkt het dat X een deel van Y lijkt te verklaren. Wanneer we Y willen schatten op basis van de regressielijn, dan duiden we deze schatting aan met \hat{Y} . Zo zal voor $X = 30$, $\hat{Y} = 13.6 + (-0.2)30 = 7.6$.

Het verklaarde deel wordt uitgedrukt als verklaarde variantie. Het becijferen van deze verklaarde variantie gebeurt aan de hand van ANOVA tabellen. Zoals weergegeven in figuur 3.1 varieert Y over verschillende waarden. De blauwe bollen geven de observaties weer, terwijl de blauwe horizontale lijn het gemiddelde van Y weergeeft. De afstand van een observatie (blauwe bol) tot de horizontale lijn, maakt onderdeel uit van de totale variantie. Deze totale variantie wordt ook wel de kwadratensom (Sum of Squares) genoemd. Wiskundig wordt de totale spreiding gedefinieerd als de Total Sum of Squares (SST) ofwel $\sum(Y_i - \bar{Y})^2$. Deze formule geeft aan dat het de som van alle afstanden betreft

tussen de blauwe bollen en de blauwe horizontale lijn. Daarnaast hebben we ook de regressielijn of best passende rechte. De afstand van de blauwe bollen tot deze lijn, noemen we de resterende fout van het model of residu (SSE) wat wiskundig $\sum(Y_i - \hat{Y})^2$ wordt ofwel het verschil tussen de blauwe bollen en de regressielijn. Tot slot hebben we de verklaarde variantie van het model, welke uitdrukt in welke mate het model meer verklaard van de variantie van Y t.o.v. \bar{Y} . Dit wordt weergegeven door $\sum(\hat{Y}_i - \bar{Y})^2$.

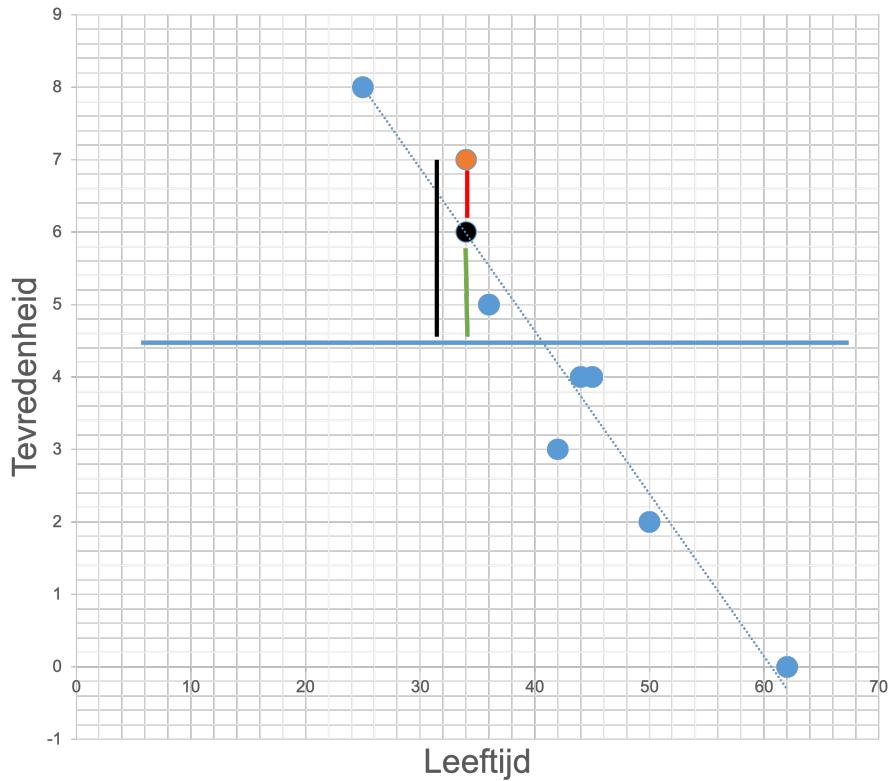


Figure 3.1: Grafische weergave van variantiebronnen

Uitgewerkt voor de oranje bol ($X = 34, Y = 7$) in figuur 3.1:

- SST: de afstand tussen de oranje bol Y_i en de blauwe horizontale lijn \bar{Y} ofwel $Y_i - \bar{Y} = 7 - 4,3 = 2,7$.
- SSE: de afstand tussen de oranje bol Y_i en de regressielijn \hat{Y} voor $X = 34$ ofwel $Y_i - \hat{Y} = 7 - (13,6 + (-0,2) \times 34) = 7 - 6 = 1$. **(opmerking:** hier werden alle cijfers na de komma meegenomen, het volledige model is: $\beta_0 = 13,6177$ en $\beta_1 = -0,02246$).
- SSR: de afstand tussen \hat{Y} en \bar{Y} .

Table 3.3: Opbouw kwadratensom (1)

SST	SSE	SSR
3.6666667	-0.0021598	3.6688265
1.6666667	0.0194384	1.6472282
2.6666667	1.0194384	1.6472282
0.6666667	-0.5313175	1.1979842
-1.3333333	-1.1835853	-0.1497480
-0.3333333	0.2656587	-0.5989921
-0.3333333	0.4902808	-0.8236141
-2.3333333	-0.3866091	-1.9467243
-4.3333333	0.3088553	-4.6421886

Table 3.4: Opbouw kwadratensom

SST	SSE	SSR
13.4444444	0.0000047	13.4602878
2.7777778	0.0003779	2.7133608
7.1111111	1.0392547	2.7133608
0.4444444	0.2822983	1.4351661
1.7777778	1.4008742	0.0224245
0.1111111	0.0705746	0.3587915
0.1111111	0.2403752	0.6783402
5.4444444	0.1494666	3.7897354
18.7777778	0.0953916	21.5499152

Wanneer we deze oefening voor alles observaties zouden herhalen en kwadrateren en optellen, komen we tot de uiteindelijke kwadratensom.

Wanneer we al deze waarden kwadrateren komen we tot volgende tabel:

Tot slot tellen we alles op:

Zoals jullie kunnen opmerken uit de laatste tabel, geldt voor de variantiebronnen dat $SST = SSE + SSR$.

Table 3.5: Opbouw kwadratensom (3)

	x
SST	50.000000
SSE	3.278618
SSR	46.721382

Table 3.6: Overzicht van alle variantiebronnen

Bron	Kwadratensom (SS)	Vrijheidsgraden (df)	Kwadratengemiddelde (MS)
Regressie	$\sum (\hat{Y}_i - \bar{Y})^2$	m	$\frac{\sum (\hat{Y}_i - \bar{Y})^2}{m}$
Residu/Fout	$\sum (Y_i - \hat{Y}_i)^2$	$n-m-1$	$\frac{\sum (Y_i - \hat{Y}_i)^2}{n-m-1}$
Totaal	$\sum (Y_i - \bar{Y})^2$	$n-1$	$\frac{\sum (Y_i - \bar{Y})^2}{n-1}$

Performantie van het regressiemodel

De performantie of verklarende kracht van het regressiemodel wordt uitgedrukt aan de hand van de *determinatiecoëfficiënt* (R^2). Deze geeft weer hoeveel van de variantie in Y verklaard kan worden aan de hand van één of meerdere onafhankelijke X variabelen. Meer bepaald is $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$. In het voorbeeld hierboven gegeven is $R^2 = 0.93$ ofwel 93. De determinatiecoëfficiënt varieert tussen 0 en 1 ($R^2 \in [0, 1]$) en hoe dichter bij 1, hoe meer van de variantie verklaard kan worden.

Deze berekening wijkt af van de slides door afronding. In de slides werd er geen afronding meegenomen, maar in dit boek werd dit wel meegenomen.

Een volledig overzicht van alle variantiebronnen is weergegeven in onderstaande tabel:

De uiteindelijke statistische grootheid op basis waarvan de uiteindelijke p -waarde berekend wordt, is gebaseerd op een afgeleide van de kwadratensom, namelijk het kwadratengemiddelde. Hiervoor wordt elke kwadratensom gedeeld door de bijhorende vrijheidsgraden. Deze vrijheidsgraden zijn $n - 1$ voor de SST (zoals in de formule van variantie uit eerste/tweede Bachelor), m voor de SSR (waarbij m het aantal onafhankelijke variabelen weergeeft) en $n - m - 1$ voor de SSE. De F-statistiek wordt uiteindelijk berekend als $F = MS_{regressie}/MS_{residu}$. Hoe groter de waarde van F , hoe meer de regressie in staat is om variantie te verklaren en hoe sneller een statisch significant resultaat gevonden kan worden. Voor het **tevreden** voorbeeld is $n - 1 = 8$, $m = 1$ en $n - m - 1 = 8 - 1 - 1 = 6$. Hieruit kunnen we F berekenen als $F = \frac{46.7}{1}/\frac{3.3}{6} = 84.9$

Op basis van de berekende F -waarde en de nul-distributie zoals weergegeven in figuur 3.2, zal de p -waarde < 0.001 . Het regressiemodel is met andere waaarden

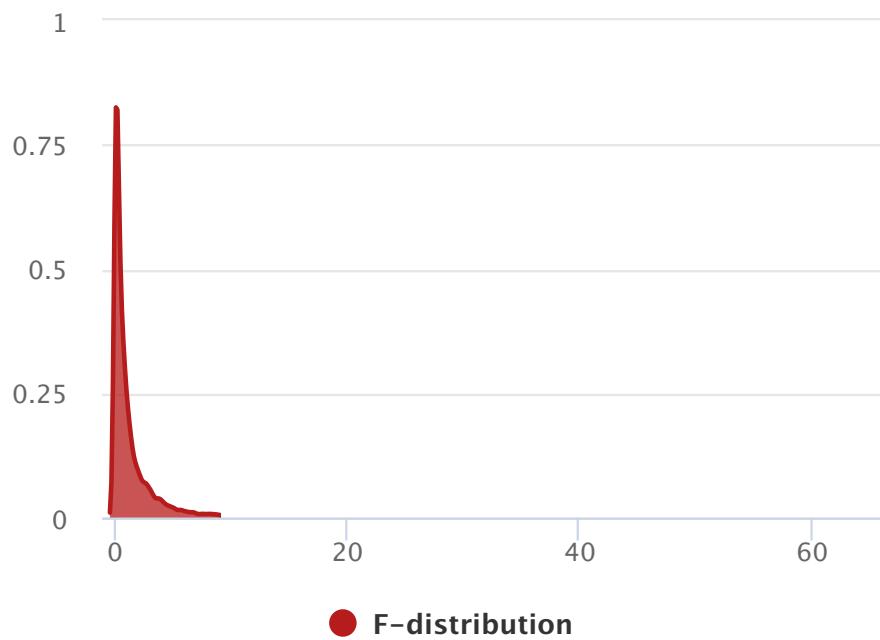


Figure 3.2: Grafische weergave van variantiebronnen

in staat om een significant deel van de variantie van Y te verklaren op basis van waarden van X .

Bij uitbreiding naar meervoudige lineaire regressie wordt een model als volgt weergegeven: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \epsilon$, waarbij Y de afhankelijke continue variabele is, X_i de onafhankelijke variabele is, β_i een inschatting is van het verband tussen X_i en Y en ϵ de fout is die gemaakt wordt door het model ($Y_i - \hat{Y}_i$).

Hypothesen

Binnen regressie kunnen er twee hypothesen worden opgesteld: (1) een hypothese over de regressie-analyse zelf en (2) een hypothese over de β coëfficiënten. De hypothese kunnen we algemeen als volgt opstellen:

- H_0 : Er is geen verband ($\beta_i = 0$) tussen de leeftijd (X_i) van de therapeut en de gegeven tevredenheidsscore (Y_i).
- H_1 : Er is een verband ($\beta_i \neq 0$) tussen de leeftijd (X_i) van de therapeut en de gegeven tevredenheidsscore (Y_i).

Een significante β_i wil dus zeggen significant verschillend van 0 en dus een meerwaarde om iets te zeggen over Y . **Let op:** binnen deze hypothesen kan het ook over meervoudige regressie gaan met verschillende β coefficiënten.

H_0 kan in dit geval $\beta_1 = \beta_2 = \beta_3 = 0$ zijn.

Er wordt nooit een hypothese gevormd over β_0 . Kan je zelf bedenken waarom dit het geval is?

Voorwaarden van een lineair regressiemodel

Er zijn drie belangrijke voorwaarden vooraleer we regressie-analyse kunnen toepassen:

- Lineariteit
- Onafhankelijke waarnemingen
- Residuen normaal verdeeld met gemiddelde 0 en constante variantie

De eerste voorwaarde, **lineariteit**, houdt in dat een regressieanalyse enkel in staat is om een correcte inschatting te maken van een verband tussen één continue afhankelijke variabele en één of meerdere afhankelijke variabelen als dit verband lineair is. Wanneer het verband niet lineair is, zal de inschatting op basis van lineaire regressie een foutief beeld geven, m.a.w. de schattingen zullen niet valide zijn en er zal *bias* optreden.

De tweede voorwaarde, **onafhankelijke waarnemingen**, geeft aan dat de klassieke lineaire regressie een inschatting kan maken wanneer er slechts één observatie is per variabele per persoon. Indien eenzelfde variabele meerdere malen gemeten of gevraagd werd bij eenzelfde persoon zal de inschatting van de regressie niet correct zijn. De schatting zal nu niet per se onderhevig zijn aan bias, maar de berekeningen van het 95%CI zal niet correct kunnen gebeuren.

De derde voorwaarde, **normaal verdeelde residuen**, geeft aan dat de fouten die door het model gemaakt zijn normaal verdeeld moeten zijn met een gemiddelde van 0. Dit is belangrijk aangezien de regressielijn een gemiddelde weergeeft van Y (\bar{Y}) voor elke X_i . Indien de residuen niet normaal verdeeld zijn en het gemiddelde van de residuen niet gelijk zou zijn aan 0, gaat de veronderstelling van een gemiddelde inschatting niet meer op.

Selectie van een regressiemodel

Er bestaan verschillende automatische selectiemethoden om een regressiemodel te definiëren:

- Forward
- Backward
- Enter
- Stepwise

In deze cursus gaan we hier niet verder op in.

Dummy coding

Wanneer we gebruik maken van een onafhankelijke categorische X variabele die bestaat uit > 2 categoriën, moeten we dummy coding uitvoeren, aangezien we een variabele met bijvoorbeeld 5 categoriën niet kunnen weergeven aan de hand van één β coëfficiënt. Neem bijvoorbeeld tabel 3.7, waarbij we de variabele **Rookstatus** met drie categoriën gaan opdelen in $m - 1$ variabelen, waarbij m het aantal categoriën weergeeft. Op basis van rookstatus wensen we in dit voorbeeld een inschatting te maken van het gewicht (in kg).

Hier: $3 - 1 = 2$ dummy categoriën.

Uit tabel 3.7 kunnen we op basis van slechts twee nieuwe variabelen de drie oorspronkelijke categoriën steeds opnieuw terughalen. Voor elk van deze nieuwe variabelen wordt er in het regressiemodel een β ingeschattet. We krijgen hierdoor volgende regressievergelijking:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
, waarbij X_1 de associatie is met **Actieve roker** en X_2 de associatie met **Vroegere roker**.

Table 3.7: Dummy coding

Rookstatus	Actieve roker	Vroegere roker
Niet roker	0	0
Vroegere roker	0	1
Actieve roker	1	0

- $Y = \beta_0 + \beta_1(\text{actieveroker}) + \beta_2(\text{vroegegeroker})$
- $Y = 64 - 2 \times (\text{actieveroker}) + 0.5 \times (\text{vroegegeroker})$

Voor een actieve roker komen we dan volgende schatting uit: $\hat{Y} = 64 - 2 \times (\text{actieveroker} = 1) + 0.5 \times (\text{vroegegeroker} = 0)$ ofwel $\hat{Y} = 64 - 2 \times 1 = 62\text{kg}$. Voor elk label binnen de variabele Rookstatus, kunnen we dus de inschatting van het gewicht gaan berekenen.

Exercise 3.1. Kan je zelf voor de andere categoriën deze inschatting maken? Hoe bepaal je het gewicht voor de categorie die is weggevallen?

Ook in wetenschappelijke artikels zal je binnen categorische variabelen merken dat er aan dummy coding werd gedaan. Hierbij wordt de referentiecategorie vaak gedefinieerd als 0.

Table 3 Multivariable linear regression for SPADI at 6-month follow-up: n=804, adjusted $R^2=0.31$ (baseline SPADI and QuickDASH not included in the model)

Independent predictor	Subgroup comparisons	Parameter estimate	95% CI	p Value
Patient expectation of change	Completely recover Much improve Slightly improve No change/worse	0 5.21 12.43 -0.94	1.80 to 8.61 8.20 to 16.67 -8.53 to 6.66	0.003 <0.001 0.809

Figure 3.3: Dummy coding in wetenschappelijk artikel

Multicollineariteit

Wanneer we gebruiken maken van meervoudige lineaire regressie, kan er multicollineariteit optreden. Hierbij zijn de verschillende onafhankelijke variabelen onderling gecorreleerd. Multicollineariteit heeft invloed op schatten van de regressiecoëfficiënten, betrouwbaarheidsintervallen, etc...

Binnen SPSS kunnen we hiervoor **Collinearity diagnostics** aanduiden. De twee meest gebruikte schattingen voor collineariteit zijn:

- Tolerance (waarde dicht bij 0 wijst op multicoll.) = % variantie in de onafh. dat niet kan worden verklaard door de andere onafh. variabelen
- Variance Inflation Factor (VIF) = reciproke waarde van de ‘Tolerance’ (hoge waarde wijst op multicoll.)

Table 3.8: Overzicht van data binnen obesitaskliniek

BMI (kg/m ²)	6MWT (meter)
23	120
28	108
26	105
22	115
31	95
26	95
32	82
36	87
42	40

SPSS en oefeningen

In dit deel kunnen jullie voor het eerst kennis maken met lineaire regressie binnen het statistisch softwareprogramma SPSS. Probeer op een gestructureerde manier alle stappen in dit document te volgen en hierna ook de vragen te beantwoorden.

Als kinesitherapeut binnen een obesitaskliniek wil je het verband tussen het BMI (kg/m²) van de patiënten en de uitkomst van de 6-minuten wandeltest (6MWT) in kaart brengen. Gegeven is de volgende dataset:

Exercise 3.2. Bereken de SST voor uitkomst van de 6MWT. Vergeet niet dat de SST berekend wordt door enkel rekening te houden met het gemiddelde.

Exercise 3.3. Het verband tussen de 6MWT en het BMI wordt weergegeven aan de hand van volgende formule: $Y = 195.5 - 3.4X$ OF: $6MWT = 195.5 - 3.4 \times BMI$ Kan je voorspellen wat de uitkomst op de 6MWT zou zijn voor een persoon met een BMI van 30?

Exercise 3.4. Bereken nu de SSE voor de uitkomst van de 6MWT op basis van bovenstaand regressie-model. Kan je ook de determinatiecoëfficiënt (R^2) berekenen?

Open SPSS en voeg de data in die in de les gebruikt werd voor het beschrijven van het verband tussen de tevredenheid van kinesitherapeuten over hun honoraria en hun leeftijd. Uiteindelijk zou je volgende dataset moeten hebben ingegeven:

Ga naar **analyze > regression > linear** en vul daar volgende gegevens aan:

Let er op dat we tevredenheid onder dependent (afhankelijk) plaatsen en leeftijd onder independent (onafhankelijke). Klik hierna op **Statistics...** en selecteer onder de statistieken “confidence intervals”. Klik hierna op **Paste** en voer de Syntax uit.

	Tevredenheid	Leeftijd
1	8	25
2	6	34
3	7	34
4	5	36
5	3	42
6	4	44
7	4	45
8	2	50
9	0	62

Figure 3.4: Data voor tevredenheid

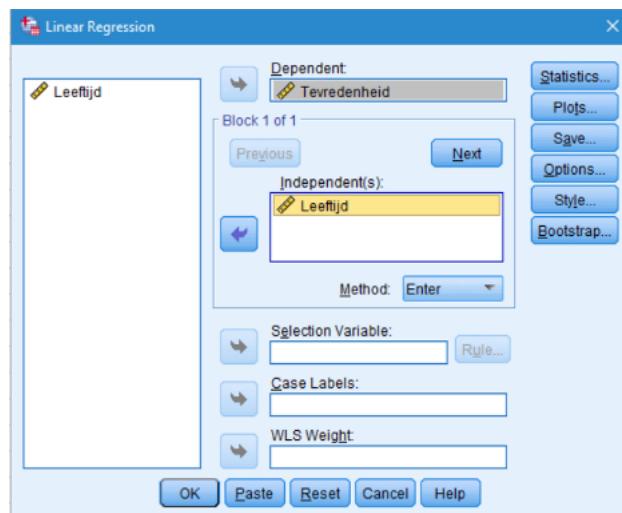


Figure 3.5: Model invullen

Exercise 3.5. Kan je in de output het regressiemodel terugvinden?

Exercise 3.6. Is er een significant verband tussen de leeftijd en de score voor tevredenheid? Waaruit leid je dit af?

Exercise 3.7. Waarvoor staat het 95% CI? Kan je dit interpreteren?

Open SPSS en voeg de data in die in de les gebruikt werd in oefening 1. Uiteindelijk zou je volgende dataset moeten hebben ingegeven:

10 : Afstand			
	BMI	Afstand	var
1	23	120	
2	28	108	
3	26	105	
4	22	115	
5	31	95	
6	26	95	
7	32	82	
8	36	87	
9	42	40	
10			
11			

Figure 3.6: Data voor obesitas

Ga naar `analyze > regression > linear` en vul daar volgende gegevens aan:

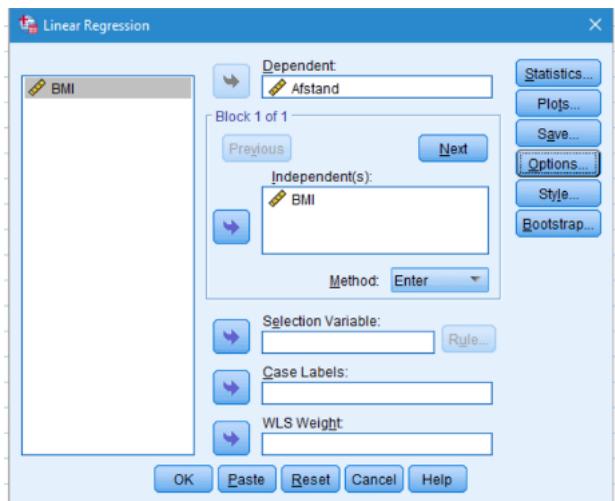


Figure 3.7: Model invullen obesitas

Klik hierna op **Statistics...** en selecteer onder de statistieken “confidence intervals”. Klik hierna op **Paste** en voer de Syntax uit.

Exercise 3.8. Kan je in de output de fouten terugvinden die het model maakt (SSE en SST)? Wat is de waarde van de SSE?

Exercise 3.9. Is er een significant verband tussen de BMI en de 6MWT? Wat is de p-waarde voor deze associatie?

Exercise 3.10. Waarvoor staat het 95% CI? Kan je dit interpreteren?

Probeer ook nog even volgende meerkeuzevragen te beantwoorden:

Exercise 3.11. Ik heb een categorische variabele over het opleidingsniveau welke bestaat uit 5 mogelijke antwoorden. Selecteer het correcte antwoord.

- Ik kan deze variabele zo toevoegen in het lineaire model in SPSS.
- Ik zal deze moeten hercoderen naar 5 nieuwe variabelen.
- Ik zal deze moeten hercoderen naar 4 nieuwe variabelen.
- Ik zal deze moeten hercoderen naar een continue variabele met schaal 1 t.e.m. 5.

Exercise 3.12. Als ik een lineair model maak in SPSS, dan...

- Is het belangrijk dat ik steeds de lineariteit, multicollineariteit en normaalverdeling van de residuen bekijk.
- Dien ik eerst te kijken naar de VIF om multicollineariteit na te gaan.
- Dien ik gebruik te maken van een spearman correlatiecoëfficiënt om lineariteit na te gaan.

Exercise 3.13. Wanneer je verschillende onafhankelijke variabelen hebt, is het vaak moeilijk om een finel model te selecteren. Op basis van jouw opgedane kennis, hoe denk je dat het finaal model het best geselecteerd wordt? Op basis van...

- De klinische relevantie.
- Een automatische methode voor variabelen selectie.
- Het veranderen van R^2 , hoe hoger deze wordt hoe beter mijn model.
- De significantie van de variabelen in mijn model.

Je maakt een lineair regressiemodel met extensiekraft ter hoogte van de knie als afhankelijke variabele en leeftijd, lage rugpijn, geslacht en fysieke fitheid als onafhankelijke variabelen. Gegeven is volgende onvolledige ANOVA-tabel. Probeer deze zo goed mogelijk aan te vullen (de zwarte vakken dienen niet te worden aangevuld).

	SS (Sum of Squares)	df	MS (Mean square)	F	Sig
Regression	2200				< .001
Residual					
Total	2800	65			

Figure 3.8: Invullen ANOVA tabel

Conclusies

Enkele belangrijke stappen die je steeds moet doornemen:

- ANOVA-tabel en p-waarde
- Determinatiecoëfficiënt (R^2)
- Voorwaarden vervuld (Lineariteit, onafhankelijk en residuen)
- Residuenanalyse
- Klinische relevantie!
- Regressie \neq een causaal verband.
- Regressie is mogelijk met één of meerdere onafhankelijke variabelen.
- De R^2 zegt iets over hoe goed de regressielijn in staat is om Y te verklaren.
- Regressie geeft steeds een lineair verband weer en is dus gerelateerd aan ρ .
- De afhankelijke variabele Y bij lineaire regressie is steeds continu.

Chapter 4

Logistische regressie

In de paper van Smith et al. (2021) vermeld men volgende stelling:

“Player age greater than 25 years (OR = 2.9, $p < 0.05$), was a factor significantly associated with subsequent hamstring injury.”

—Smith et al. (2021)

- Wat kunnen we hieruit besluiten?

Risico relatie

Een risico op een aandoening kan vaak ingeschat worden als een kans, waarbij we beroep kunnen doen op de **prevalentie** en **incidentie**. Wanneer we weten dat de **prevalentie** van lage rugpijn 50% is, dan is binnen de populatie de kans dat iemand lage rugpijn heeft 50%. De **prevalentie** (en de **incidentie**) kan dus gezien worden als een risico op lage rugpijn. Om een risico op een binaire gebeurtenis (waarbij het optreden van bijvoorbeeld lage rugpijn vaak gecodeerd wordt als 1 en het niet optreden van lage rugpijn als 0) statistisch te beschrijven wordt er vaak gebruik gemaakt van twee statistische grootheden, een **risico** en een **odds**.

Risico en risico ratio (RR)

Gegeven is de volgende dataset **tevreden** met informatie over een bevraging bij kinesitherapeuten over hun tevredenheid van honoraria.

- Y : tevredenheid (0-niet tevreden; 1-tevreden)

Table 4.1: Steekproef naar tevredenheid bij kinesitherapeuten.

X	Y
< 40	1
< 40	1
< 40	1
< 40	1
< 40	0
< 40	0
40+	0
40+	0
40+	0
40+	1

Table 4.2: Kruistabel van tevredenheid.

	0	1
< 40	2	4
40+	3	1

- X : leeftijd (in categorieën)

We beschikken in deze specifieke steekproef over een totaal van 9 observaties ($n = 10$).

de dataset ziet er als volgt uit:

Om het risico op tevredenheid na te gaan (**risico** kan ook een positieve quotatie hebben), dienen we te identificeren hoeveel individuen aangegeven hebben tevreden te zijn over hun honorarium. In deze studie gaven 5 van de 10 therapeuten aan tevreden te zijn. Het risico (of kans) op tevredenheid is in dit geval dus 50 %.

Wanneer we het risico in één groep willen uitzetten ten opzichte van een andere groep, kunnen we gebruik maken van het relatieve risico (of risicoratio).

De risicoratio wordt vaak verkozen boven bijvoorbeeld het risicoverschil, aangezien een klein risico verschil een grote impact kan hebben wanneer het bijvoorbeeld een zeldzame aandoening gaat. Stel dat we het effect van een therapie willen nagaan bij een zeldzame aandoening (waarbij de prevalentie 2% is). Wanneer onze therapie in staat is de prevalentie te verlagen naar 1%, waren we in staat om met onze therapie een verschil te bewerkstelligen van 1% (risicoverschil). Als we echter kijken naar de risicoratio (relatief risico), dan waren we in staat om de prevalentie te halveren $\frac{1\%}{2\%} = 0.5$. Wanneer we

Table 4.3: Steekproef naar tevredenheid bij kinesitherapeuten.

X	Y
< 40	1
< 40	1
< 40	1
< 40	1
< 40	0
< 40	0
40+	0
40+	0
40+	0
40+	1

dezelfde oefening maken voor een ziekte met een hogere prevalentie (50%), dan is een effect van 1% door therapie nog steeds een effect van 1%, maar het relatieve risico is wel gewijzigd naar $\frac{49\%}{50\%} = 0.98$. Omwille van deze verschillen wordt er vaak verkozen om risico's te vergelijken op basis van ratio's in plaats van verschillen.

Om het verschil in risico op tevredenheid tussen jongere en oudere therapeuten in te schatten, dienen we eerst het risico op tevredenheid in elke groep te berekenen. Voor de jongste groep < 40 is het risico $\frac{4}{6} = 66.7\%$, terwijl voor de oudere groep 40+ het risico $\frac{1}{4} = 25.0\%$ is. De risico ratio of relatieve risico (RR) voor tevredenheid in de jongere t.o.v. de oudere groep kan dan geschat worden door $\frac{0.667}{0.250} = 2.67$. De RR is direct en letterlijk interpreteerbaar als een toegenomen risico op tevredenheid waarbij het risico bij kinesitherapeuten jonger dan 40 jaar 2.6 keer hoger is dan in de groep 40 jaar of ouder.

Odds en odds ratio (OR)

Om de odds en odds ratio (OR) te berekenen gebruiken we dezelfde dataset:

Om de odds op tevredenheid na te gaan (**odds** kan ook een positieve quotatie hebben), dienen we te identificeren hoeveel individuen aangegeven hebben tevreden te zijn over hun honorarium. In deze studie gaven 5 van de 10 therapeuten aan tevreden te zijn. Het risico (of kans) op tevredenheid is in dit geval dus 50 %. De odds is de verhouding van de kans op het plaatsvinden van een gebeurtenis over de kans dat de gebeurtenis niet plaats vindt. De odds wordt berekend door $\frac{p}{1-p}$ of $\frac{0.5}{1-0.5} = 1$. De odds op tevredenheid is in dit voorbeeld dus 1, aangezien de kans op een tevreden kinesitherapeut even groot is als de kans op een ontevreden kinesitherapeut.

Wanneer we de odds in één groep willen uitzetten ten opzichte van een andere

Table 4.4: Kruistabel van tevredenheid.

	0	1
< 40	2	4
40+	3	1

groep, kunnen we gebruik maken van de relatieve odds (of odds ratio). Om de OR op tevredenheid van jongere t.o.v. oudere therapeuten in te schatten, dienen we eerst de odds op tevredenheid in elke groep te berekenen. Voor de jongste groep < 40 is de odds $\frac{4}{2} = \frac{4}{2} = 2.00$, terwijl voor de oudere groep 40+ de odds $\frac{1}{3} = \frac{1}{3} = 0.33$ is. De odds ratio (OR) voor tevredenheid in de jongere t.o.v. de oudere groep kan dan geschat worden door $\frac{2.00}{0.33} = 6.1$. De OR is moeilijk en niet letterlijk interpreteerbaar aangezien het weergeeft dat de odds bij kinesitherapeuten jonger dan 40 jaar 6.1 keer hoger zijn dan in de groep 40 jaar of ouder.

Odds vs Risico

De OR geeft is stabiel en verondersteld geen oorzaak-gevolg relatie, terwijl de RR verandert wanneer rijen en kollomen (oorzaak-gevolg) verwisseld worden.

Hypothese en sterkte risico-relatie

De hypothese van een OR of RR wordt als volgt geformuleerd:

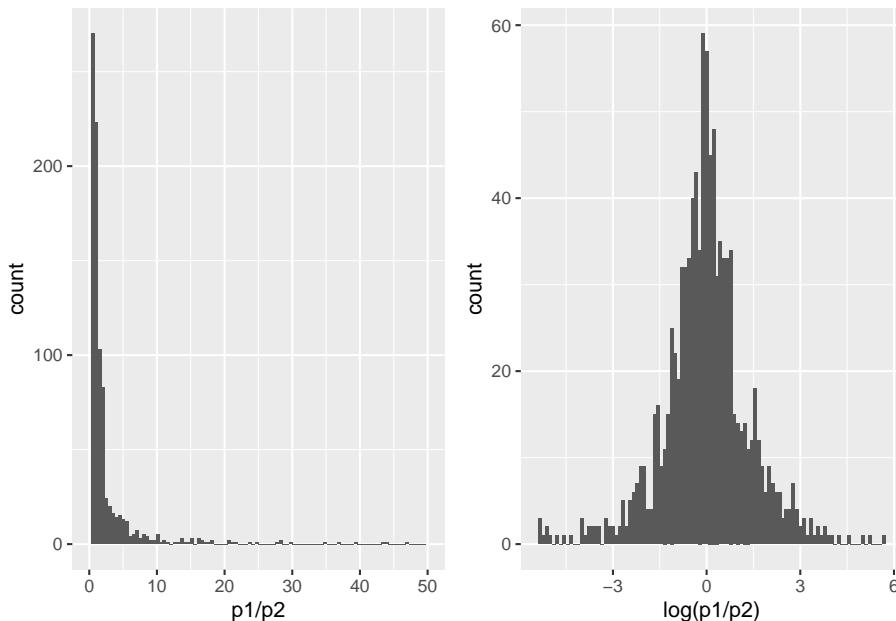
H_0 : Leeftijd is geen risicofactor (RR of OR = 1) gerelateerd aan de tevredenheid van de therapeut. H_1 : Leeftijd is een risicofactor (RR of OR $\neq 1$) gerelateerd aan de tevredenheid van de therapeut.

Een significante risicofactor, wil dus zeggen significant verschillend van 1. Belangrijk om te vermelden is dat een OR (of een RR) > 1 een hoger risico inhoudt ten opzichte van de referentiecategorie. Een OR (of een RR) < 1 houdt een lager risico in ten opzichte van de referentiecategorie. Alle ORs (of RRs) groter dan (kleiner dan) één kunnen onderling vergeleken worden qua grootte, maar RRs (of ORs) groter dan één kunnen niet vergeleken met RRs (of ORs) kleiner dan één. Wanneer we de odds ratio's berekenen van risicofactoren voor blessures bij sporters, vinden we een $OR_{geslacht:man} = 2$ en een $OR_{geenopwarming} = 8$, dan kunnen we besluiten dat het risico gerelateerd aan een opwarming groter is dan het risico gerelateerd aan geslacht, aangezien $8 > 2$.

We kunnen de referentiecategorie omdraaien voor een OR of RR door de inverse te berekenen (vb. $\frac{1}{OR}$ of $\frac{1}{RR}$). Zo kan de OR voor geen opwarming versus een opwarming gelijk zijn aan 8 ($OR_{geenopwarming} = 8$), dan is de OR

voor een opwarming tegenover geen opwarming: $OR_{opwarming} = \frac{1}{8} = 0.125$. Deze nieuwe OR is zoals in voorgaand voorbeeld moeilijk te vergelijken met de $OR_{geslacht:man} = 2$, aangezien $0.125 < 2$, maar het risico gerelateerd aan opwarming aanzienlijk groter is.

De odds, OR en RR wordt vaak weergegeven na een log-bewerking (met het natuurlijke logaritme: ln). Deze worden dan de log(odds), log(OR) or log(RR), aangezien we door deze transformatie een normaal verdeling van de uitkomst kunnen krijgen.



Inleiding logistische regressie

Binaire logistische regressie kan zowel toegepast worden in een cross-sectioneel design als prospectieve cohorte en wordt vaak gebruikt voor volgende doelstellingen:

- Het voorspellen van een uitkomstmaat op basis van één of meerdere factoren.
- Het beschrijven van een verband tussen de uitkomstmaat en andere factoren, waarbij rekening wordt gehouden met verstoorende variabelen.

In essentie is een logistisch regressie model opgebouwd uit twee componenten:

- De afhankelijke uitkomstvariabele (Y), welke een categorische variabele dient te zijn.
- De onafhankelijke variabele(n) (X), welke zowel continu als categorisch van aard kunnen zijn.

In de regressievergelijking wordt Y niet rechtstreeks gemodelleerd, maar de kans op het optreden van $Y = 1$, welke uitgedrukt wordt als $p = P(Y = 1)$.

$$\text{logit}(p) = \beta_0 + \beta_1 X_1$$

Wanneer er slechts één onafhankelijke variabele X_1 spreken we van een enkelvoudig binair logistisch regressiemodel. Indien er meerdere onafhankelijke variabelen $X_1, X_2, X_3, \dots, X_i$ spreken we van een meervoudig binair logistisch regressiemodel.

Een regressiemodel bestaat uit twee delen, een afhankelijk deel (de uitkomst of wat er geschat moet worden) en een onafhankelijk deel (de verklarende variabelen).

De logit-functie komt overeen met de log-odds: $\text{logit}(p) = \log(\text{odds}) = \log(\frac{p}{1-p})$.

Studie naar slaagkansen

Gegeven is de volgende dataset **geslaagd** met informatie over een bevraging bij studenten over hun gestuurde uren en slaagkansen op een examen.

- Y : geslaagd (1 = ja, 0 = nee)
- X : gestudeerd (uren)

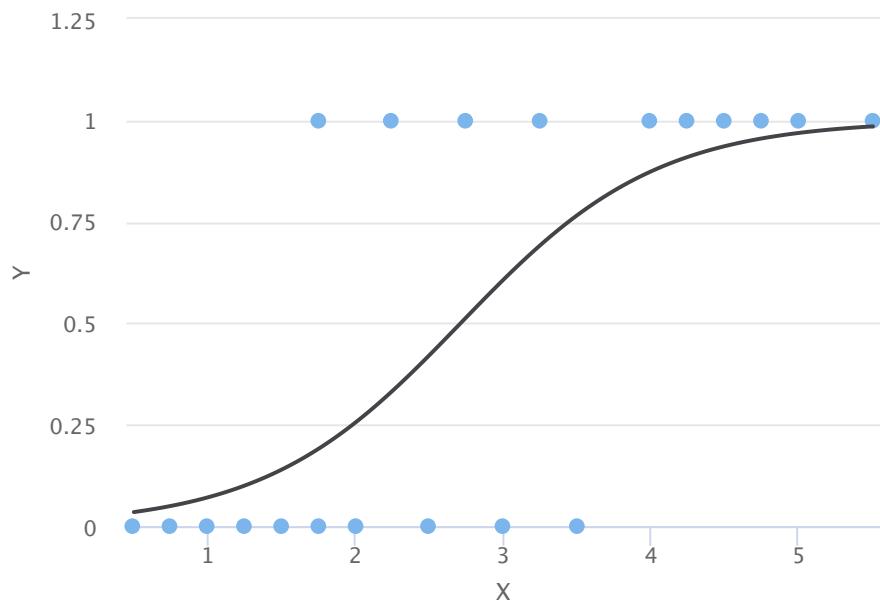
We beschikken in deze specifieke steekproef over een totaal van 20 observaties ($n = 20$).

de dataset ziet er als volgt uit:

Als we de relatie visueel willen weergeven tussen het aantal gestudeerde uren en het al dan niet slagen op een examen, krijgen we volgende plot:

Table 4.5: Steekproef naar studeergedrag.

X	Y
0.50	0
0.75	0
1.00	0
1.25	0
1.50	0
1.75	0
1.75	1
2.00	0
2.25	1
2.50	0
2.75	1
3.00	0
3.25	1
3.50	0
4.00	1
4.25	1
4.50	1
4.75	1
5.00	1
5.50	1



Op basis van de `geslaagd` dataset krijgen we volgende schatting

$$\text{logit}(p) = -4.0777 + 1.5046X \quad \text{log}(\text{odds}(p)) = -4.0777 + 1.5046X$$

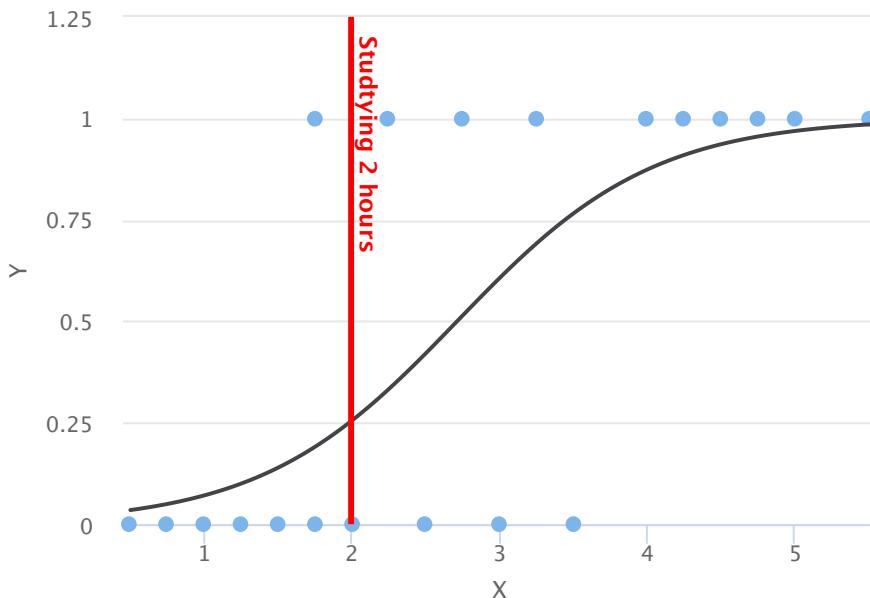
$$\text{of } \text{odds}(p) = \exp(-4.0777 + 1.5046X).$$

Wanneer we de regressievergelijking invullen krijgen we dus een inschatting van $\text{odds}(p)$.

Interpretatie: Er is geen rechtstreekse interpretatie van de β -coëfficiënten. Wanneer we de exponentiële functie toepassen op de β -coëfficiënten krijgen we wel een bekende interpretatie, namelijk de OR. De $OR_{gestudeerd} = \exp(1.5046) = 4.5$ kunnen we interpreteren als een toename van de odds met 4.5 keer wanneer studenten één uur extra studeerden.

Interpretatie: Wanneer we een inschatting willen maken van de odds op slagen voor iemand die 2 uur gestudeerd heeft, vullen we de volledige regressievergelijking in. Deze wordt $\text{odds}(p) = \exp(-4.0777 + 1.5046 \times 2) = 0.344$, waarbij de odds op slagen dus 0.344 bedraagt. De odds kunnen opnieuw omgezet worden naar een kans p door gebruik te maken van volgende formule: $p = \frac{\text{odds}(p)}{\text{odds}(p)+1} = \frac{0.344}{0.344+1} = 25.6\%$. De kans op een geslaagd resultaat bedraagt dus 25.6%.

Deze inschatting kunnen we ook als volgt visueel voorstellen:



Eigenschappen van een logistisch regressiemodel

Het schatten van een binair logistische regressievergelijking gebeurt niet op basis van de SSE, SSR of SST zoals bij lineaire regressie. Er wordt gebruik gemaakt van likelihood criteria, die ook aanwezig zullen zijn in de SPSS output bij het opstellen van een regressiemodel. Om een betekenis te geven aan de β -coëfficiënten van een binair logistisch regressiemodel dienden we $\exp(1.5046) = 4.5$, welke we interpreteerden als een toename van de odds op slagen met 4.5 wanneer een student één uur langer gestuurd heeft. Wanneer we echter geïnteresseerd zijn in een toename per 2 uur studeren, kunnen we de OR niet zomaar vermenigvuldigen met 2. We moeten hiervoor de β -coëfficiënt vermenigvuldigen met 2 en deze omzetten naar een OR. In het voorbeeld van onze studie naar het slagen dienen we dus volgende berekening uit te voeren: $\exp(1,5046 \cdot 2) = 20.3$. Deze kunnen we interpreteren als volgt: wanneer een student twee uur langer gestudeerd heeft, dan zal diens odds op slagen toenemen met 20.3.

We hoeven niet per se steeds de OR om te zetten naar een β -coëfficiënt en vice versa. We kunnen ook steeds de OR verheffen tot een bepaalde macht. In het voorbeeld van het aantal uren studeren komt dit op het volgende neer: om de interpretatie van één uur studeren te veranderen in een interpretatie voor twee uur studeren voeren we volgende bewerking uit: $4.5^2 = 20.3$. Inderdaad, exact wat we daarnet hebben uitgerekend.

Table 4.6: Steekproef naar studeergedrag en uitkomsten op basis van model.

x	y	y_p	y_fit
0.50	0	0.0347103	0
0.75	0	0.0497729	0
1.00	0	0.0708920	0
1.25	0	0.1000286	0
1.50	0	0.1393445	0
1.75	0	0.1908365	0
1.75	1	0.1908365	0
2.00	0	0.2557032	0
2.25	1	0.3335302	0
2.50	0	0.4216265	0
2.75	1	0.5150109	1
3.00	0	0.6073586	1
3.25	1	0.6926173	1
3.50	0	0.7664808	1
4.00	1	0.8744475	1
4.25	1	0.9102776	1
4.50	1	0.9366237	1
4.75	1	0.9556107	1
5.00	1	0.9690971	1
5.50	1	0.9851944	1

4.0.1 Performantie van het regressiemodel

Voor elke observatie kunnen we een schatting maken van de odds en dus ook de kans. Op basis van deze kans kunnen we beslissen of we denken dat de gebeurtenis al dan niet zich voor zal doen. We beslissen dus op basis van P of $Y = 1$ of $Y = 0$. (vb. Wanneer $P < 0.5$, dan is $Y = 0$ en anders $Y = 1$). In onderstaande tabel kunnen jullie de kans op slagen voor elke x terugvinden gedefinieerd als y_p . Op basis van deze y_p ($P(Y = 1)$) beslissen we of $Y = 1$ of $Y = 0$, wat gedefinieerd is als y_{fit} .

Op basis van deze uitkomst kunnen we een 2x2 kruistabel maken, waarbij we kijken in welke mate de observaties overeenstemmen met de predicties.

Op basis van deze tabel kunnen we een inschatting maken van de accuraatheid, sensitiviteit en specificiteit. In de kolommen vind je de predicties terug en in de rijen de observaties. De accuraatheid geeft weer hoeveel van alle observaties we correct kunnen voorspellen. In deze studies komt dit neer op $\frac{16}{20} = 80$. De sensitiviteit geeft aan hoeveel van de positieve observaties ($Y = 1$) we juist hebben kunnen inschatten. In dit geval is dit $\frac{8}{10} = 80$. De sensitiviteit geeft aan

Table 4.7: Geobserveerde vs voorspelde y.

	0	1
0	8	2
1	2	8

Table 4.8: Overzicht van data binnen de sportclub

Testbatterij	Blessure
pos	0
pos	1
neg	0
neg	0
neg	0
pos	1
neg	1
neg	0
neg	0
pos	0

hoeveel van de negatieve observaties ($Y = 0$) we juist hebben kunnen inschatten. In dit geval is dit $\frac{8}{10} = 80$.

De observaties zijn steeds de referentie. In bovenstaande tabel worden deze weergeven in de rijen (zoals in SPSS).

SPSS en oefeningen

In dit deel kunnen jullie voor het eerst kennis maken met binair logistische regressie binnen het statistisch softwareprogramma SPSS. Probeer op een gestructureerde manier alle stappen in dit document te volgen en hierna ook de vragen te beantwoorden.

Als kinesitherapeut binnen een sportclub wil je een preventieprogramma opstarten. Je gaat van slag met het aanmaken van een testbatterij waarin kracht, lenigheid en psychologisch welzijn gemeten wordt. Wanneer de testbatterij positief scoort, dan is er een verhoogd risico op een blessure. In je studie neem je bij het begin van het seizoen bij iedereen deze testbatterij af en je volgt alle spelers op gedurende het volledige seizoen.

Exercise 4.1. Bereken voor de volledige groep sporters wat het risico en odds op het krijgen van een blessure is. Bereken hierna ook de RR en OR voor een

blessure bij een positieve testbatterij ten opzicht van een negatieve testbatterij.

Exercise 4.2. Het verband tussen het oplopen van een blessure en de testbatterij wordt weergegeven aan de hand van volgende formule: $odds(p) = \exp(-1.609 + 1.609 \times X)$ OF: $odds_{blessure} = \exp(-1.609 + 1.609 \times Test_+)$ Kan je voorspellen wat de kans op een blessure is voor iemand met een negatieve test op basis van het model?

Exercise 4.3. Hoe kan je β_{test_+} omzetten naar een OR?

Open SPSS en voeg de data in die in de les gebruikt werd voor het beschrijven van het verband tussen de tevredenheid van kinesitherapeuten over hun honoraria en hun leeftijd. Uiteindelijk zou je volgende dataset moeten hebben ingegeven:

Tevredenheid	Leeftijd	Tevreden
8	25	tevreden
6	34	tevreden
7	34	tevreden
5	36	tevreden
3	42	niet tevreden
4	44	niet tevreden
4	45	niet tevreden
2	50	niet tevreden
0	62	niet tevreden
6	42	tevreden

Figure 4.1: Data voor tevredenheid

Ga naar `analyze > regression > binary logistic` en vul daar volgende gegevens aan:

Let er op dat we tevredenheid onder dependent (afhankelijk) plaatsen en leeftijd onder het “block”. Klik hierna op `Options...` en selecteer onder de statistieken “CI for exp(B)” and “Hosmer-Lemeshowgoodness-of-fit”. Klik hierna op `Paste` en voer de Syntax uit.

Wanneer je de input `paste` in de syntax en uit voert, zal je merken dat SPSS een intieme codering uitvoert zoals aangegeven in figuur 4.4. De uitkomst **niet tevreden** van de variabele van de binaire variabele **tevreden** werd gehercodeerd naar 0, terwijl de uitkomst **tevreden** werd gehercodeerd naar 1.

Exercise 4.4. Kan je in de output het regressiemodel terugvinden?

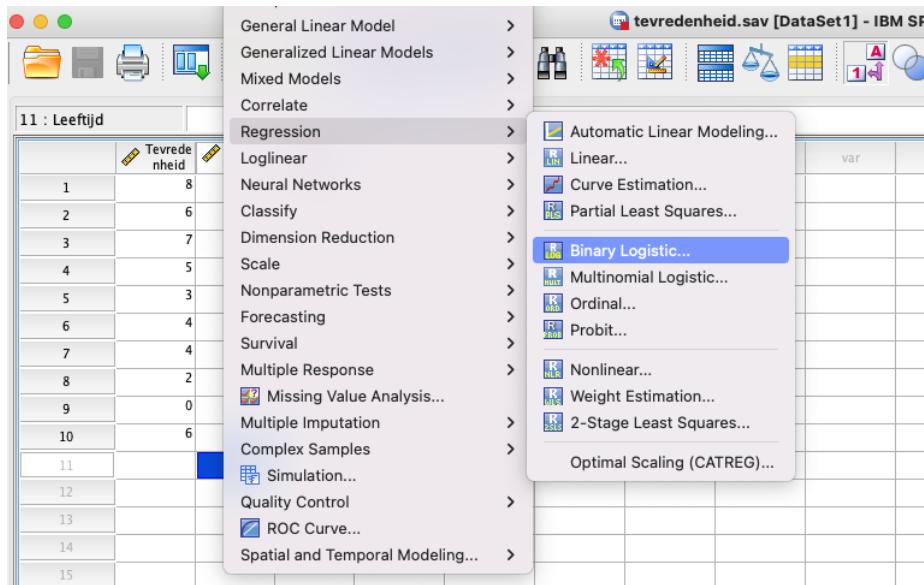


Figure 4.2: Model invullen

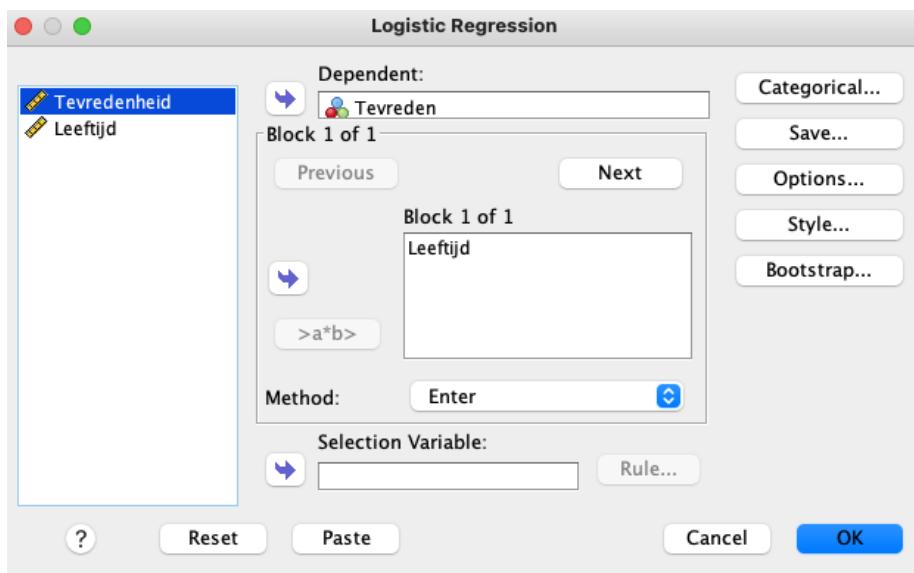


Figure 4.3: Model invullen

Dependent Variable Encoding	
Original Value	Internal Value
niet tevreden	0
tevreden	1

Figure 4.4: Hercodering binnen SPSS

Exercise 4.5. Is er een significant verband tussen de leeftijd en de score voor tevredenheid? Waaruit leid je dit af?

Exercise 4.6. Wat valt er op aan het model? Denk je dat het hier om een valide model gaat? Of zijn er eventuele problemen die we moeten aanpakken? Indien er problemen zijn, kan je dan opmerken waaraan deze problemen kunnen liggen?

Chapter 5

Meervoudige regressie

Inleiding

Voorgaande hoofdstukken waren vooral gefocust op een regressievergelijking, waarbij de afhankelijke uitkomst (Y) wordt verklaard aan de hand van één onafhankelijke factor (X). Uiteraard weten we dat er in werkelijkheid vaak meer dan één factor gerelateerd zal zijn aan de uitkomst.

Voorbeeld: Wanneer we een idee willen hebben over het functioneren van een patiënt 3 jaar na het plaatsen van een totale heupprothese. Hiervoor kunnen we gebruik maken van een WOMAC score. Bij enklevoudige regressie kunnen we focussen op het verband tussen deze WOMAC-score en bijvoorbeeld de leeftijd. Bij meervoudige regressie willen we echter ook rekening houden met andere factoren, zoals het geslacht, maar ook misschien het BMI of het al dan niet ervaren van comorbiditeiten. We bouwen hierbij nog steeds één model, dat al deze factoren omvat.

Meervoudige regressie (zowel lineair als binair) wordt dus vooral gebruikt in de volgende context: * Meenemen van covariaten * Nagaan van interacties * ...

Een **nadeel** van meervoudige regressie is dat het meenemen van meer variabelen vaak gepaard gaat met de nood aan een grotere steekproef.

Visualisatie

Een manier om een regressiemodel te visualiseren is de Directed acyclic graph (DAG), hierbij worden factoren aan elkaar gerelateerd door pijlen. Een voorbeeld hiervan is weergegeven in figuur 5.1. Hierbij merken we op dat er een pijl wordt getrokken van de exposure (blootstelling) naar de outcome (uitkomst). De DAG wil hiermee visualiseren dat er een theoretische invloed is van een bepaalde

blootstelling (bv. de leeftijd) op een bepaalde uitkomst (bv. spierkracht). Een belangrijke **opmerkingen** hierbij is dat de pijl een theoretisch concept is, waarbij er veronderstelt wordt dat er een invloed is van één factor op de andere. Uiterraard weten we dat zo een oorzaak-gevolg relatie alleen duidelijk kan worden in een gerandomiseerde studie. Bij observationele studie zonder randomisatie, stelt de pijl eerder een ‘associatie’ voor.



Figure 5.1: Voorbeeld van DAG

Wanneer we het principe van de DAG gaan toepassen op basis van de studieresultaten die werden weergegeven in figuur 5.2, krijgen we een DAG zoals weergegeven in figuur 5.3. Het is belangrijk om te vermelden dat we ons ook bij een meervoudig regressiemodel, focussen op één van de opgenomen factoren (in dit geval de leeftijd). Het meervoudige regressiemodel is weergegeven in de rechterkolom van figuur 5.2. Dit is aangegeven in de titel van de tabel door het woord **multivariate**. We merken dat er drie factoren werden opgenomen, namelijk leeftijd, het al dan niet hebben van lage rugpijn en kwaliteit van leven (SF-36).

Table 6 Univariate and multivariate logistic regression analysis of preoperative predictors and postoperative characteristics for the patients who scored worst (lowest quartile) in WOMAC function at 3.6 year follow up compared with the patients who scored better								
Predictor	No	OR	95% CI	p Value	Multi-variate step (n=160)	OR	95% CI	p Value
<i>Preoperative characteristics</i>								
Age	189	1.09	1.04 to 1.15	0.001	1	*1.09	1.03 to 1.15	0.002
Sex	189	1.35	0.70 to 2.58	0.37				
BMI	170	1.11	1.01 to 1.22	0.03				
Comorbid conditions >2	189	1.48	0.62 to 3.51	0.38				
SF-36 BP	186	0.97	0.58 to 0.99	0.007	2	*0.97	0.94 to 0.995	0.018
SF-36 PF	183	0.98	0.96 to 0.997	0.03				
SF-36 MH	176	0.99	0.97 to 1.004	0.16				
<i>Postoperative characteristics</i>								
Comorbid conditions >2	189	2.06	0.93 to 4.55	0.08				
Complication	189	3.55	1.03 to 12.18	0.04				
Worst side pain	189	1.29	0.48 to 3.11	0.61				
Pain index hip	189	2.50	0.80 to 7.82	0.12				
Pain contralateral hip	189	1.95	0.75 to 5.10	0.17				
Low back pain	189	3.31	1.59 to 6.89	0.001	1	3.31	1.59 to 6.89	0.001
Living alone	189	1.50	0.77 to 2.90	0.23				

*Per one year or scale unit increase.
BP, bodily pain; PF, physical function; MH, mental health.

Figure 5.2: Studie naar WOMAC

Binnen de DAG zoals weergegeven in figuur 5.3, werd er gefocust op de relatie tussen de leeftijd en WOMAC score . Verder werd deze relatie binnen het meervoudige regressiemodel gecorrigeerd voor het hebben van lage rugpijn en kwaliteit van leven . De relatie tussen leeftijd en WOMAC wordt m.a.w. gecorrigeerd voor andere covariaten (soms ook benoemd als confounding of vervuilende variabelen). We kunnen deze oefening herhalen voor elk van de relaties binnen het meervoudige regressiemodel.

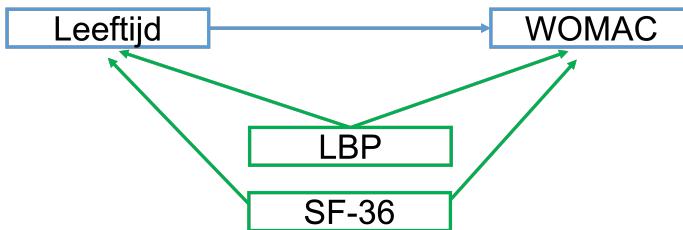


Figure 5.3: DAG voor WOMAC-studie

Terminologie

Vooraleer gebruik te maken van zo'n DAGs is het belangrijk om op de hoogte te zijn van de terminologie. In de eerste plaats is het belangrijk een onderscheid te maken tussen een **collider** en een **confounder**. Een **collider** ontstaat wanneer in werkelijkheid 2 factoren de onderliggende oorzaak zijn van een andere factor zoals weergegeven in figuur 5.4. Wanneer we in ons model zouden corrigeren voor een **collider**, dan ontstaat er net bias.

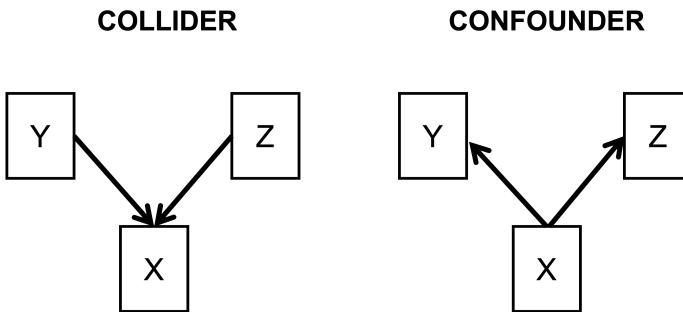


Figure 5.4: Voorbeeld van DAG

Een concreet voorbeeld van zo'n **collider** bias is de obesitas paradox. In deze paradox lijkt het alsof obesitas zorgt voor een verlaagde mortaliteit bij patiënten met een cardiovasculaire aandoening. Wanneer we dus obesitas samen met een cardiovasculaire aandoening in één model, ontstaat er collider bias. Het is namelijk zo dat binnen de groep van patiënten met zo'n aandoening, dat deze patiënten - wanneer ze opgenomen worden in het ziekenhuis - langer lijken te overleven. Wanneer we echter kijken binnen de volledige populatie, is het duidelijk dat obesitas leidt tot een verhoogde sterfte. Mensen met obesitas hebben misschien wel meer reserves, waardoor hun overlevingstijd langer wordt wanneer we enkel kijken naar de mensen die leiden aan andere aandoeningen. **Belangrijk** hierbij is te weten dat het opnemen van covariaten dus niet steeds leidt tot het corrigeren van een relatie.

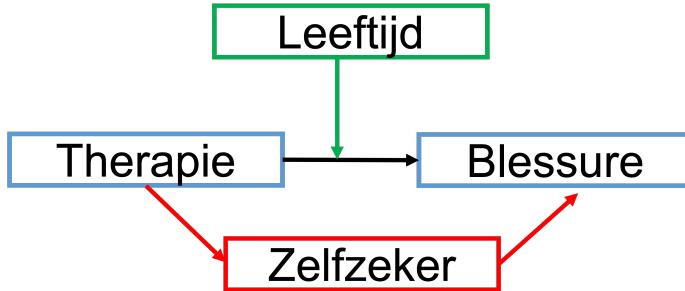


Figure 5.5: Terminologie van een DAG

Binnen de DAG die weergegeven is in figuur @ref(fig: dag5) worden de verschillende relatie weergegeven:

- Directe effect: deze geeft het directe effect weer van het geven van een therapie op het oplopen van blessures.
- Moderatie: een modererende factor is een factor die een invloed heeft op de relatie tussen twee factoren.
- Mediatie: een mediërende factor is een factor die onrechtstreeks een invloed weergeeft tussen twee factoren. Deze worden ook indirecte effecten genoemd.

In het voorbeeld, wordt het directe effect weergegeven door een direct effect van de therapie op het al dan niet oplopen van blessures. De mediator in dit voorbeeld is zelfzeker zijn. De therapie kan immers ervoor zorgen dat de sporter zijn zelfzekerheid veranderd (toeneemt of afneemt) wat ook een effect kan hebben op het al dan niet oplopen van blessures. Tot slot is er de leeftijd als moderator. Het is immers zo dat mensen met een oudere leeftijd misschien minder effect hebben van de therapie in vergelijking met jongeren. De leeftijd heeft dus een invloed op het effect van de therapie. Deze modererende effecten worden in een model vaak geschat op basis van een interactie-effect.

Toepassingen DAG

In een gerandomiseerde studie vonden de onderzoekers een positief effect van kinesitherapie op kwaliteit van leven bij patiënten na een knie-prothese. Binnen deze studie werd er gevonden dat de effectiviteit veel hoger lag bij jongere patiënten ten opzichte van oudere patiënten.

Hoe kunnen we de rol van leeftijd definiëren? *Leeftijd* is in dit geval een moderator.

In een gerandomiseerde studie vonden de onderzoekers een positief effect van kinesitherapie op kwaliteit van leven bij patiënten na een knie-prothese. Binnen deze studie werd er gevonden dat de effectiviteit veel hoger lag bij jongere patiënten ten opzichte van oudere patiënten.

Moeten we corrigeren voor geslacht in deze studie wanneer we weten dat geslacht een mogelijke confounder is? *Eigenlijk niet*, aangezien de studie gerandomiseerd is. Randomisatie wil zeggen dat de therapie enkel gebaseerd is op toeval en er geen factoren zijn die van invloed kunnen zijn op de therapie.

In een niet-gerandomiseerde studie vonden de onderzoekers een positief effect van kinesitherapie op kwaliteit van leven bij patiënten na een knie-prothese. Binnen deze studie werd er gevonden dat de effectiviteit veel hoger lag bij jongere patiënten ten opzichte van oudere patiënten. Belangrijke confounders zijn geslacht en mate kan kinesiofobie.

Hoe zou een DAG eruit zien? *Zie figuur 5.6*

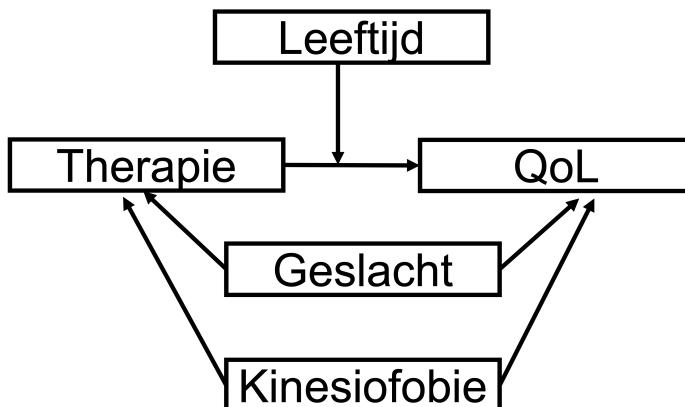


Figure 5.6: DAG

Meervoudige regressie

Een meervoudig lineair regressiemodel wordt weergegeven aan de hand van $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \epsilon$. Binnen deze regressievergelijking kunnen we confounder opnemen en kunnen we ook interactie-effecten (moderatie) opnemen, maar geen mediatie.

Confounder

Een voorbeeld van de interpretatie van een regressiemodel wanneer we confounder opnemen is als volgt:

Meervoudige regressie: opnemen van meerdere factoren » correctie van het effect voor de andere factoren.

$$\text{Tevredenheid} = \beta_0 + \beta_1 \times \text{leeftijd} + \beta_2 \times \text{geslacht}(man)$$

- β_1 = inschatting van het effect van leeftijd op tevredenheid gecorrigeerd voor geslacht.
- β_2 = inschatting van het effect van geslacht op tevredenheid gecorrigeerd voor leeftijd.

Moderatie

Bij een moderatie-analyse wordt er gebruik gemaakt van interacties. Een interactie wordt als volgt weergegeven in een regressievergelijking:

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_1 \times X_2 + \epsilon$$

β_3 geeft in deze formule de interactie weer tussen X_1 en X_2 . De interpretatie van β_3 is echter moeilijk op basis van bovenstaande formule. Wanneer we deze formule herschrijven als

$$Y = \beta_0 + \beta_2 \times X_2 + (\beta_1 + \beta_3 \times X_1) \times X_2 + \epsilon,$$

kunnen we het effect van deze β_3 beter evalueren. De associatie tussen Y en X_2 wordt nu weergegeven door $(\beta_1 + \beta_3 \times X_1)$, waarbij de associatie:

- $\beta_1 + \beta_3$ is wanneer $X_1 = 1$
- β_1 is wanneer $X_1 = 0$

Stel dat $Y = \text{tevredenheid}$, $X_1 = \text{leeftijd}$ en $X_2 = \text{geslacht}$. Wanneer er bij mannen een toename zou zijn in tevredenheid bij een stijgende leeftijd en bij vrouwen een afname in tevredenheid bij een stijgende leeftijd, dan is er een interactie-effect, waarbij β_3 significant zal zijn.

Een visuele weergave van een interactie-effect is terug te vinden in figuur 5.7. Zwart geeft de associatie weer voor vrouwen, terwijl lichtgrijs de associatie weergeeft voor mannen.

We kunnen interactie of moderatie-effecten gebruiken voor alle combinaties van variabelen:

- categorisch x categorisch
- categorisch x continu
- continu x continu

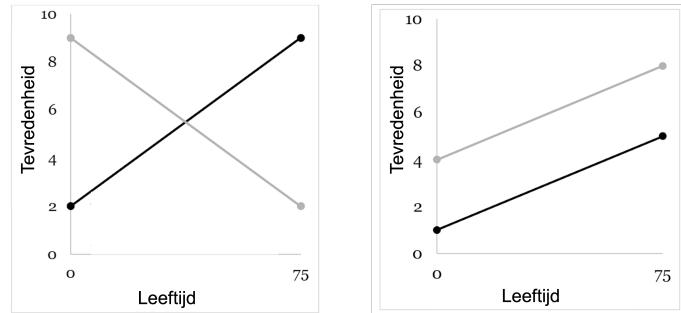


Figure 5.7: DAG

Selectie van het meest passende regressiemodel

Lineaire regressie: * R^2 (determinatiecoëfficiënt) moet zo hoog mogelijk zijn * Geweten confounders * Significantie van de variabelen in het model

Logistische regressie: * Naglekerke's R^2 (determinatiecoëfficiënt) moet zo hoog mogelijk zijn * Geweten confounders * Significantie van de variabelen in het model * Zo hoog mogelijk: Acc, Se of Sp.

Chapter 6

Voorbeeldexamen

Om dit voorbeeld examen te kunnen maken dien je te beschikken over een SPSS (.sav) dataset. Om deze dataset te downloaden kan je op volgende knop duwen:

Inleiding

Dit voorbeeldexamen geeft een reële indruk van de verwachtingen op het examen met betrekking tot jullie kennis en vaardigheden binnen SPSS. Aan de hand van deze vragen, zouden jullie in staat moeten zijn om alle vragen binnen dit voorbeeld examen op te lossen. Zoals jullie zien, zal het examen bestaan uit drie delen: (1) een praktisch deel waar jullie aan de slag kunnen gaan met SPSS en (2) een praktisch deel met fragmenten uit een artikel dat jullie dienen te interpreteren en (3) een laatste deel met enkele korte theoretische vragen over de geziene stof. Als alles goed gaat, zouden jullie in staat moeten zijn om dit voorbeeld examen binnen het uur op te lossen.

DEEL I

Neem de dataset SPSS_DATA_REVAKI.sav (die je hierboven kan downloaden) en open deze in SPSS. Probeer hierna alle vragen op een correcte manier te beantwoorden. Alle info over deze dataset vinden jullie binnen het variable-frame van SPSS.

Maak een lineair regressiemodel waarbij de afhankelijke variabelen de CRP-waarden in het bloed is en de onafhankelijke variabelen diabetes (REF = geen diabetes), leeftijd (jaren) en geslacht (REF = man). Beantwoordt hierna volgende vragen. Voor alle duidelijkheid, de referentiecategorie (REF) is de categorie die geassocieerd moet worden met waarde 0.

Exercise 6.1. Selecteer de meest gepaste R^2 van het model en schrijf deze neer.

Exercise 6.2. Hoeveel observaties zitten er in het model?

Exercise 6.3. Kan je tot twee cijfers na de komma weergeven wat de associatie is tussen CRP en leeftijd (per 10 jaar)?

Exercise 6.4. Noteer hieronder de variabelen die significant zijn (van sterkste associatie naar zwakste associatie in absolute eenheden).

Maak een logistisch regressiemodel waarbij de afhankelijke variabele de status van de patiënt is (waarbij we als uitkomst kijken naar transferred patiënten). Als onafhankelijke variabelen kunnen jullie gebruik maken van leeftijd, geslacht (REF = vrouw), obesitas, hoge bloed druk en cardiovasculaire aandoeningen (hierbij is bij alle comorbiditeiten “niet aanwezig” de REF). Beantwoord hierna volgende vragen.

Exercise 6.5. Wat is de OR voor cardiovasculair risico?

Exercise 6.6. Hoeveel % is extra correct geklassificeerd als we het model met variabelen beschouwen t.o.v. het model zonder variabelen?

Exercise 6.7. Wat is de p-waarde voor het model?

Exercise 6.8. Wat is het 95% betrouwbaarheidsinterval rond de OR voor geslacht?

DEEL II

Bekijk aandachtig de knipsel uit het artikel Bodin, J., Ha, C., Le Manac'h, A. P., Sérazin, C., Descathia, A., Leclerc, A., ... & Roquelaure, Y. (2012). Risk factors for incidence of rotator cuff syndrome in a large working population. Scandinavian journal of work, environment & health, 436-446. Beantwoordt op basis van deze knipsels de vragen zo correct mogelijk. Er is steeds maar één juist antwoord.

Exercise 6.9. Op basis van bovenstaande tabel zien we (onderaan) dat de Hosmer-Lemeshow test niet significant is. Wat kunnen we hieruit afleiden?

De model-fit is adequaat en zowel model 1 als model 2 zijn voldoende representatief. De model-fit is inadequaat en de data en het model vertonen sterke discrepanties. *De residuen zijn normaal verdeeld en alle voorwaarden van het model zijn voldaan.* De residuen zijn niet normaal verdeeld en niet alle voorwaarden van het model zijn voldaan.

Exercise 6.10. Multivariate model in de titel wil zeggen dat...

Er rekening werd gehouden met andere mogelijke verstorende (confounding) variabelen. De OR apart zijn berekend voor elk model en er dus mogelijks confounding kan zijn van de resultaten. *Er rekening werd gehouden met meerdere afhankelijke variabelen in het model.* Geen van bovenstaande.

Table 4. Multivariate model of risk factors for incident rotator cuff syndrome (RCS) in the male working population.^a

	Men (N=825; 51 RCS)									
	Model 1				Model 2					
	N	%	OR	95% CI	P-value	N	%	OR	95% CI	P-value
Age										
<40	444	3.2	1		0.001	444	3.2	1		0.001
40-44	161	6.8	2.3	1.0-5.2		161	6.8	2.3	1.0-5.2	
45-49	132	12.9	4.6	2.2-9.8		132	12.9	4.7	2.2-10.0	
≥50	88	10.2	3.6	1.5-8.8		88	10.2	3.7	1.5-9.0	
High perceived physical exertion ^b										
No	662	5.3	1		0.163					
Yes	163	9.8	1.6	0.8-3.2						
Repeated and sustained posture with the arms above shoulder level (≥2h/day)										
No	732	5.3	1		0.043					
Yes	93	12.9	2.2	1.0-4.7						
High perceived physical exertion ^b and repeated and sustained posture with the arms above shoulder level (≥2h/day)										
No factor						613	4.9	1		
One factor						168	8.3	2.0	1.0-3.8	
Both factors						44	15.9	3.3	1.3-8.4	
Low coworker support										
No	681	5.3	1		0.033					0.035
Yes	144	10.4	2.0	1.1-3.9		681	5.3	1		
						144	10.4	2.0	1.1-3.9	

^a Hosmer-Lemeshow goodness-of-fit test: P=0.540 in Model 1 and P=0.901 in Model 2.^b RPE Borg scale ≥15.

Figure 6.1: Bodin et al. (2012).

Exercise 6.11. Waarom is bij de leeftijdscategorie van <40 een OR van 1 weergegeven en geen 95% CI berekend?

Deze categorie is de referentiecategorie waarmee we de andere categorieën dienen te vergelijken. Dit wil zeggen dat de categorie < 40 jaar niet significant geassocieerd is met verhoogd of verlaagd risico. *In deze categorie zaten er onvoldoende observaties om een uitspraak te kunnen doen.* Geen van bovenstaande.

Exercise 6.12. In bovenstaand model 2 is er voor de factor “High perceived physical exertionb and repeated and sustained posture with the arms above shoulder level (more or equal then 2h/day)”.

Een significant verhoogd risico wanneer “both factors” aanwezig zijn t.o.v. “one factor”. Zijn de odds voor rotator cuff syndrome (RCS) 3.3 keer hoger in de groep waarbij “both factors” positief zijn t.o.v. “no factor”. *Er geen significant risico gevonden voor deze factor aangezien de p-waarde niet kleiner is dan 0.01.* Geen van bovenstaande.

DEEL III

Probeer onderstaande theoretische vragen zo correct mogelijk te beantwoorden. Er is steeds maar één juist antwoord.

Exercise 6.13. Om een zo correct mogelijk model te bouwen met een lineaire regressie-analyse is het belangrijk dat:

Er een lineair verband bestaat tussen de onafhankelijke en afhankelijke uitkomstvariabelen. De residuen niet normaal verdeeld zijn en een gelijke variantie hebben.

De VIF-waarde kleiner is dan 1, zodat multicollineariteit zoveel mogelijk uitgesloten kan worden. Geen van bovenstaande.

Exercise 6.14. Vul op een zo correct mogelijke manier volgende tabel aan. Geef aan welk cijfer je verwacht in de verschillende vakjes (tot 3 cijfers na de komma). Het gaan om een lineair regressiemodel met 4 ONafhankelijke variabelen.

	SS (Sum of Squares)	df	MS (Mean square)	F	Sig
Regression	A	B		C	< .001
Residual	150				
Total	550	19			

Figure 6.2: Invulexamen

Vul hierna de waarden in voor A, B en C.