EBM & Statistiek III

Robby.DePauw@Ugent.be

2021-11-08

Contents

Over deze cursus

De cursus EBM & Statistiek III bouwt verder op de leerstof uit EBM & Statistiek I en EBM & Statistiek II. De situering, vereiste begincompetanties 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 lesgever

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.

Uiprintversie 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

Het onderzoekproces bestaat uit verschillende belnagrijke stappen die nodig zijn om tot een duidelijk en rechtlijnige conclusie te komen. In de voorbije jaren hebben jullie de verschillende onderzoekstappen bestudeerd en zijn jullie geëindig met één van de laatste stappen van het onderzoeksproces, de **inductieve statistiek**. Enkele kernbegrippen hierbij waren *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. De aanvaarde methode is dat men voorlopig aanneemt dat het effect niet bestaat, en nagaat of deze veronderstelling stand kan houden in het licht van de verzamelde. De veronderstelling dat ene effect (verschil, associatie, ...) niet bestaan is dan ook vaak het uitgangspunt binnen H_0. Op basis van de verzamelde data proberen we H_0 te verwerpen in het voordeel van H_a. De beslissing om H_0 te verwerpen gebeurt uiteindelijk op basis van de p-waarde. Als cut-off punt wordt hier vaak 5% (of 0.05) genomen. We noemen deze ciut-off waarde ook wel het significantieniveau of α Waarom specifiek deze cut-off gebruikt wordt, wordt uitgelegd in het volgende deel van de cursus.

Aangenomen wordt dat geen effect geassocieerd wordt met H_0, wat niet betekend dat er geen a:ternatieve mogelijk zijn. We kunnen bijvoorbeeld ook stellen dat we onder H_O veronderstellen dat een effect een specifiek getal moet zijn.

Fouten bij hypothesetoetsen

Op basis van een hypothesetoetse 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:

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
```

Null hypothesis is...	True	False
Rejected	**Type I error** False positive Probability = α	**Correct decision** True positive Probability = 1−β
Not rejected	**Correct decision** True negative Probability = 1−α	**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.

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 evidentietabellen 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.

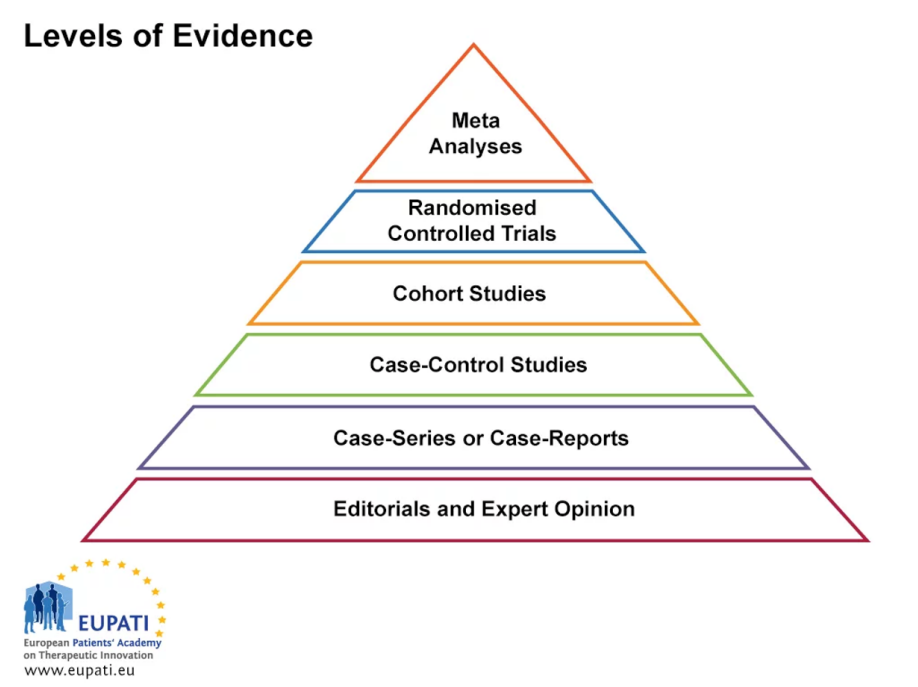


Figure 2.1: Evidentiepiramide.

Volgend voorbeeld verduidelijkt het probleem van vervuilig 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 variëreren 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

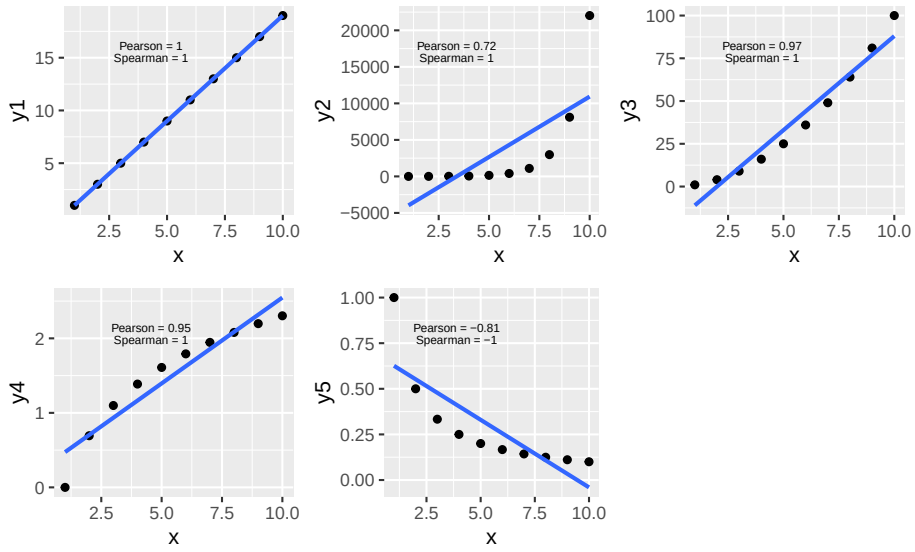


Figure 2.2: Different types of associations and their respective correlations

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 Peason correlatie (ρ)

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

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

Studie naar tevredenheid

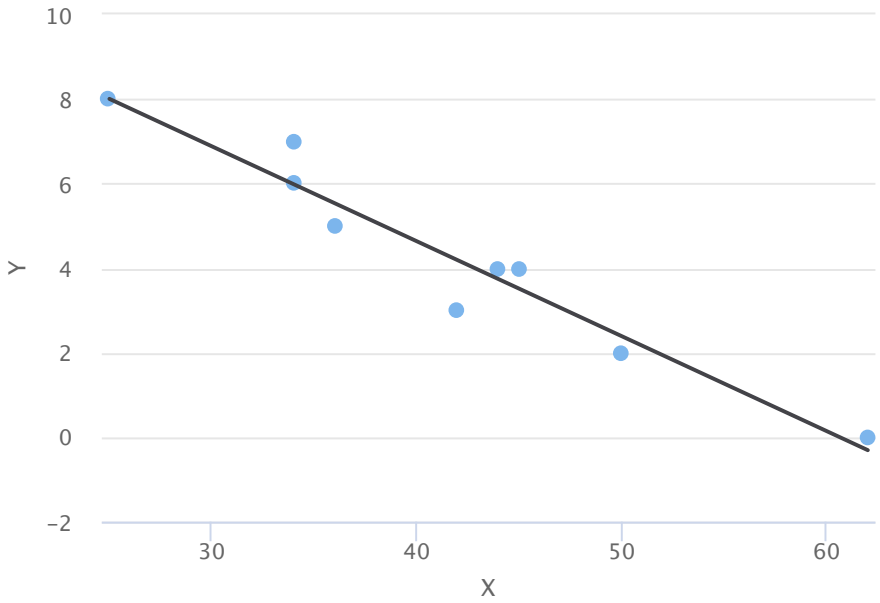
Gegeven is de volgende dataset `tevreden` met informatie over een bevraging bij kinesitherapeuten over hun tevredenheid van honoraria.

- 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`, waarbin $n = 9$ kinesitherapeueten 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

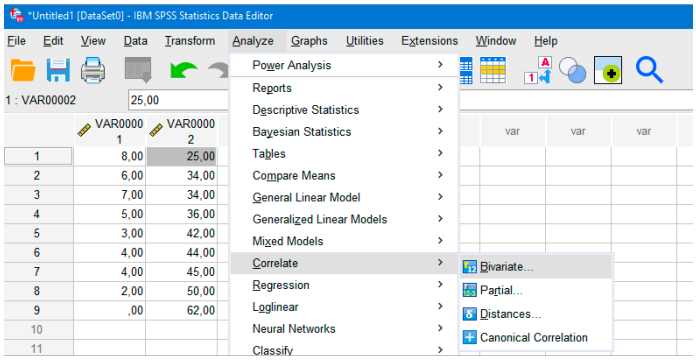


Figure 2.4: SPSS

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 $r_s = 1$.

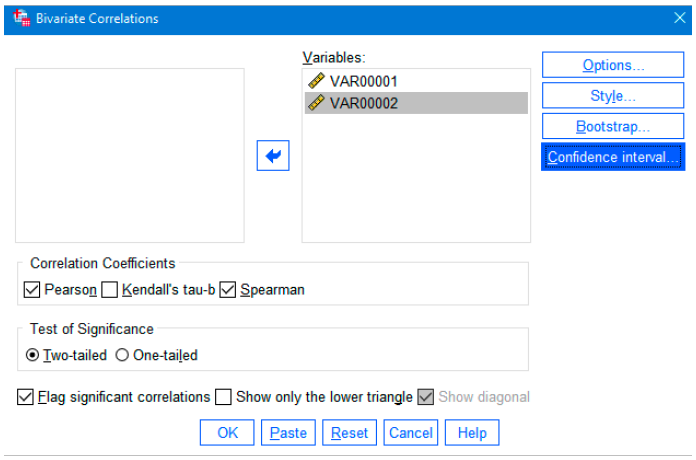


Figure 2.5: SPSS

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

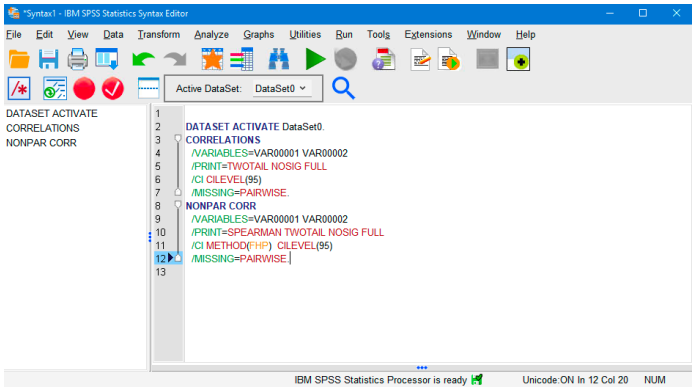


Figure 2.6: SPSS

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

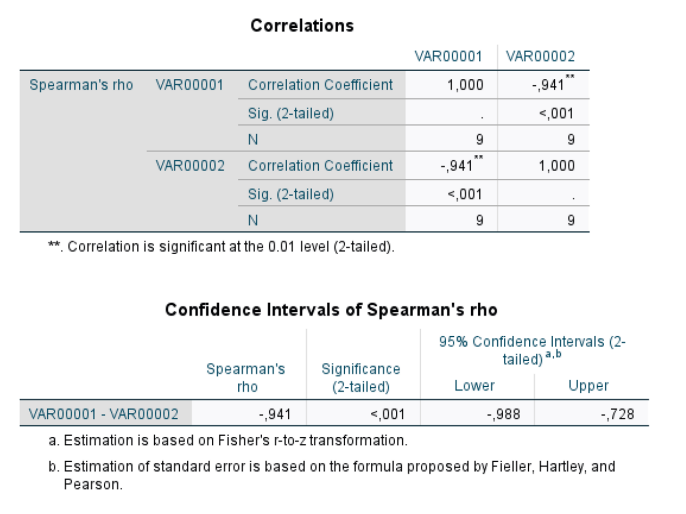
**. Correlation is significant at the 0.01 level (2-tailed).

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	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	
			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 lestel 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 geïntereseerd in het verband tussen leeftijd en het anatal 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.

vallen	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

Exercise 2.3. Een correlatie kunnen we steeds weergeven aan de hand van een lijn.

- Waar
- Onwaar

Alle volgende oefeningen hebben te maken met 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 afgenomen en de tijd (in s) wordt opgenomen. Verder wordt er bij elke deelnemer ook bevraagd hoevaak 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 2.4. 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 2.5. Welke grafiek past het best voor een visualizatie van deze gegevens?

Exercise 2.6. Welke correlatie zouden jullie selecteren voor het berekenen van een correlatie en waarom?

Exercise 2.7. 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, β = -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 verstorende 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.

$Y = \beta_0 + \beta_1 X_1$

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

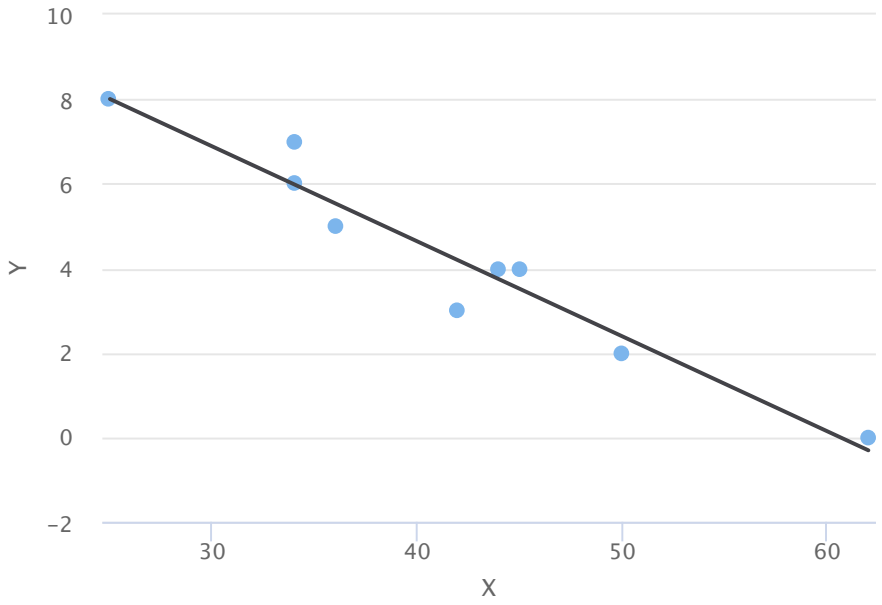
We beschikken in deze specifieke steekproef over een totaal van 9 observaties ($n = 9$).

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



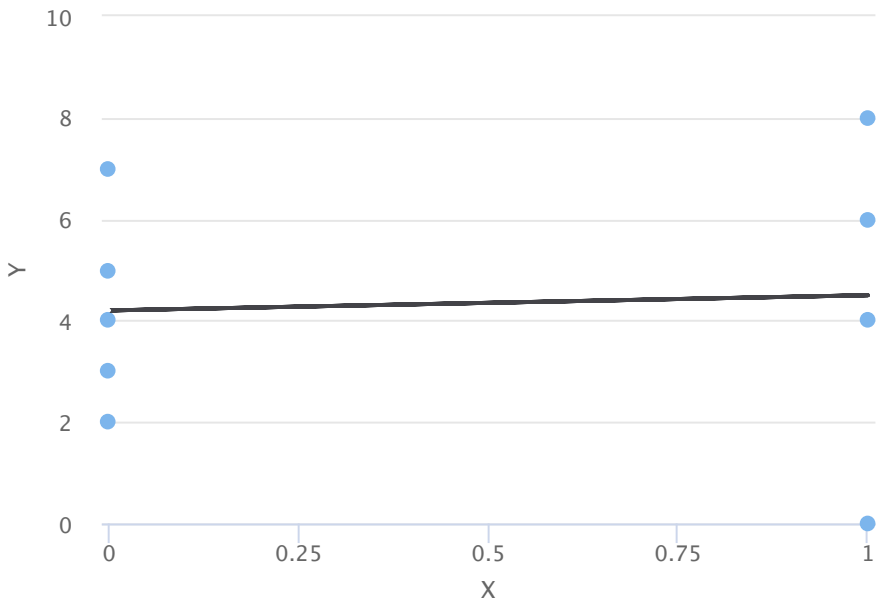
Lineaire regressie geeft net als ρ het lineaire verband weer (in tegenstelling tot r_s, waarbij het monotone verband geschat wordt). Op basis van X en Y uit de `tevreden`dataset werd volgend regressiemodel geschat:

$Y = 13.6 + (-0.2)X$

Interpretatie: Wanneer de leeftijd X toeneemt met één jaar, dan daalt $(-)$ de tevredenheidscore 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.

z	y
M	4.5
V	4.2



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

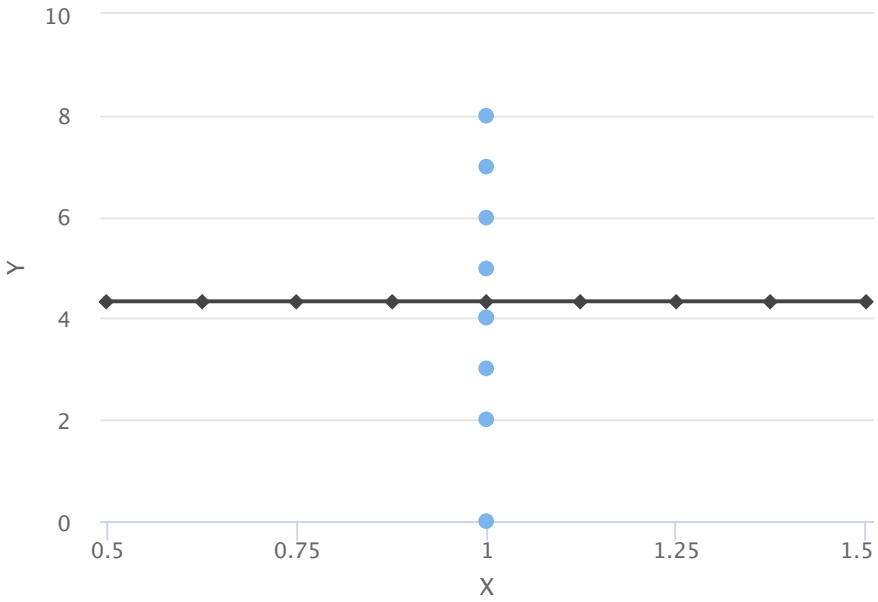
$$Y = 4.5 + (-0.3)Z$$

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 gegveen door vrouwen).

Eigenschappen van een lineair regressiemodel

Schatten van lineaire regressie

Eén van de belangrijste 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 $\sum(Y_i - \hat{Y})^2$ word ofwel het verschiul tussen de blauwe bollen en de regressielijn. Tot slot hebben we de verklaarde variantie van het model, welke uitdruk in welke mate het model meer verklaard van de variantie van Y t.o.v. $\bar{Y}$. Dit wordt weergegeven door $\sum(\hat{Y}_i - \bar{Y})^2$.

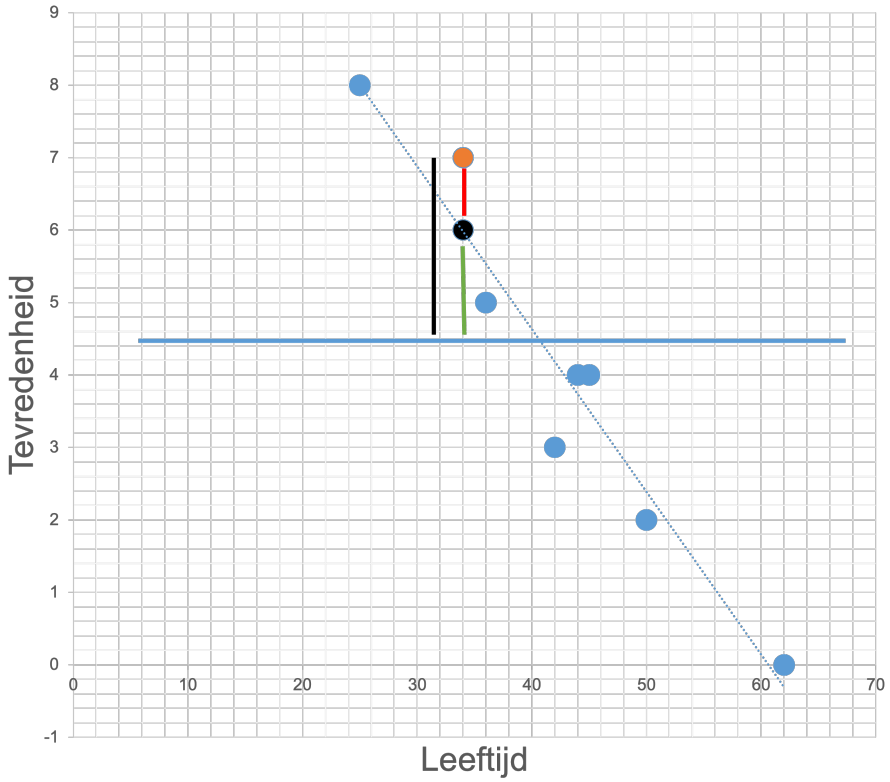


Figure 3.1: Grafische weergave van variantiebronnen

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 -
Residu/Fout	$\sum (Y_i - \hat{Y})^2$	$n-m-1$	$\frac{\sum (Y_i - \hat{
Totaal	$\sum (Y_i - \bar{Y})^2$	$n-1$	$\frac{\sum (Y_i - \bar{

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 varieerd 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 onafhelijke 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 signifcant 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 waaorden

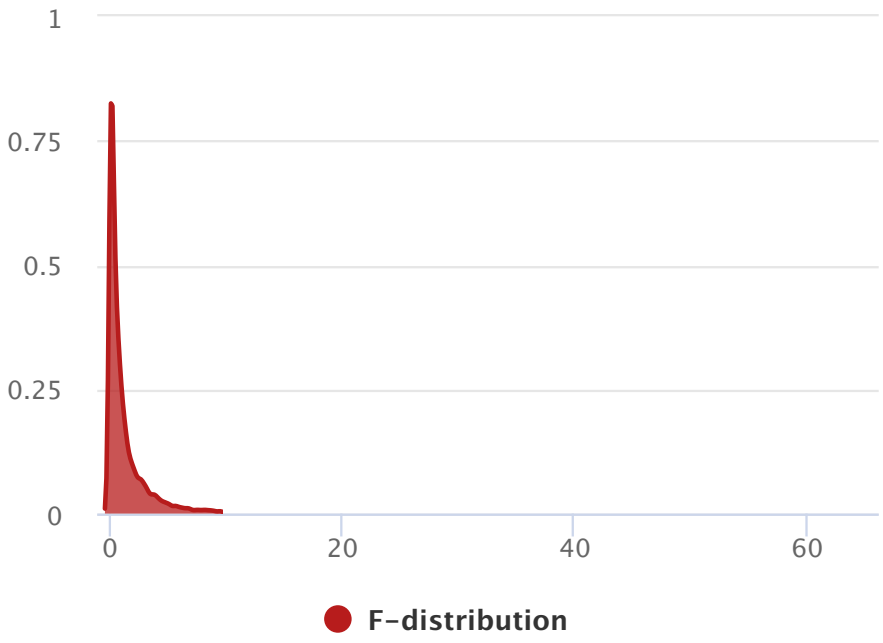


Figure 3.2: Grafische weergave van variantiebronnen

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: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon$, waarbij Y de afhankelijke continue variabele is, X_i de onafhenkelijke 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
- Onafhakelijke 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, **onafhakelijke 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 bevraagd 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 cetagoriën, 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 categoriëen 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 ingeschat. 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

- $Y = \beta_0 + \beta_1(actieveroker) + \beta_2(vroegereroker)$
- $Y = 64 - 2 \times (actieveroker) + 0.5 \times (vroegereroker)$

Voor een actieve roker komen we dan volgende schatting uit: $\hat{Y} = 64 - 2 \times (actieveroker = 1) + 0.5 \times (vroegereroker = 0)$ ofwel $\hat{Y} = 64 - 2 \times 1 = 62kg$. 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	0		
	Much improve	5.21	1.80 to 8.61	0.003
	Slightly improve	12.43	8.20 to 16.67	<0.001
	No change/worse	−0.94	−8.53 to 6.66	0.809

Figure 3.3: Dummy coding in wetenschappelijk artikel

Multicollineariteit

Wanneer we gebruiken maken van meervoudige lineaire regressie, kan er multicollienariteit optreden. Hierbij zijn de verschillende onafhankelijke variabelen onderling gecorreleerd. Multicollienariteit heeft invloed op schatten van de regressiecoëffi¨cnten, 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/m2)	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/m^2) 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: $Y = 195.5\check{}3.4X$ OF: $6MWT = 195.5\check{}3.4 \times BMI$ Kan je voorspellen wat de uitkomst op de 6MWT zou zijn voor een person 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

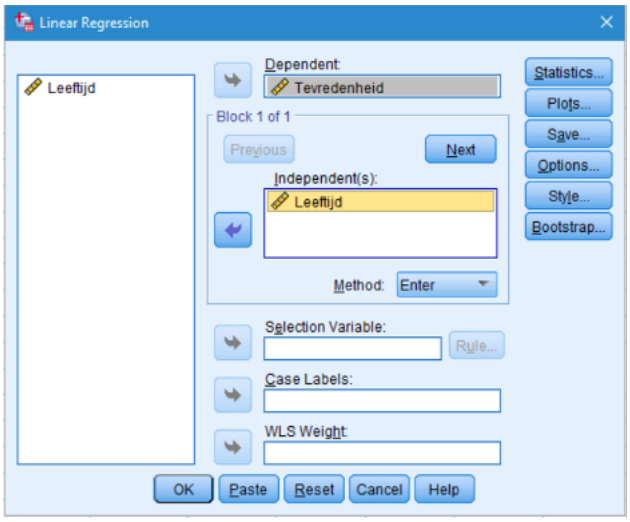


Figure 3.5: Model invullen

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:

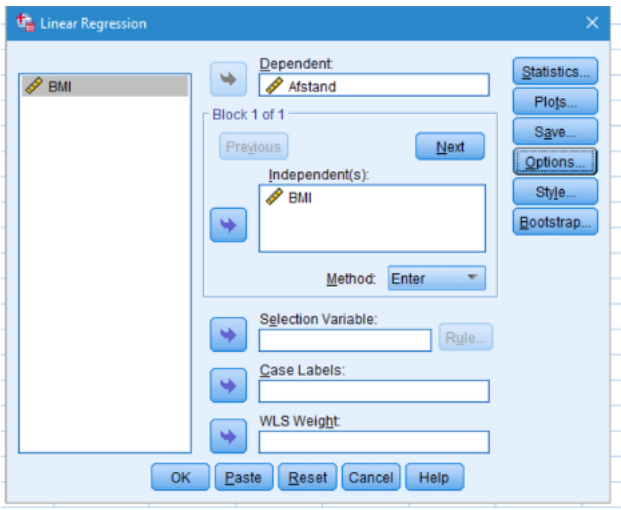


Figure 3.7: Model invullen obesaitas

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 moete 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 correlatiecoeffiënt 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 extensiekracht 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 lieanire regressie is steeds continu.