FACULTEIT PSYCHOLOGIE EN
PEDAGOGISCHE WETENSCHAPPEN

METHODEN IN DE PSYCHOLOGIE

Beatrijs Moerkerke

Vakgroep Data-analyse

Academiejaar 2020-2021

UNIVERSITEIT
GENT

Methoden in de psychologie - Academiejaar 2020-2021

Hoofdstukken in cursus

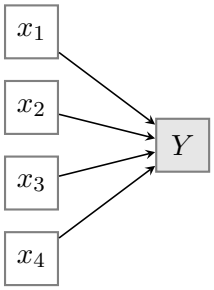
1. Statistische en praktische significantie
2. Inleiding: matrixalgebra
3. Het Algemeen Lineair Model
4. Lineaire regressie, Variantie- en covariantie-analyse
5. Het multivariaat lineair model
6. Het Veralgemeend Lineair Model
7. Logistische regressie
8. Meta-analyse

Overzicht inhoud

1. **Statistische en praktische significantie**
   p-waarde, effectgroottes en statistische power
2. **Het algemeen lineair model: lineaire regressie, variantie- en covariantie-analyse, moderatie & mediatie**

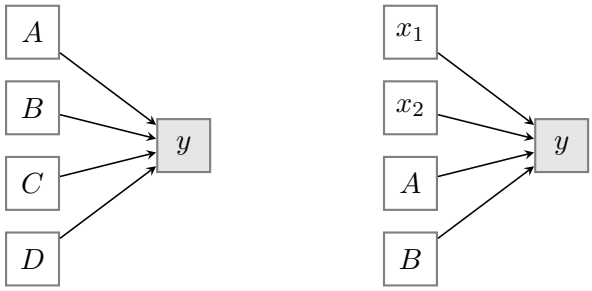
   modelleren van 1 continue uitkomst in functie van predictoren / onafhankelijke variabelen

   Univariaat regressiemodel



   Zowel Y als x zijn numerieke variabelen van minstens intervalniveau.

   Anova en Ancova



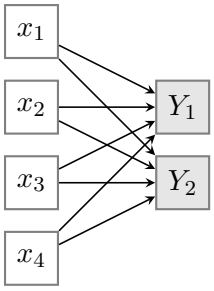
   - A, B, C en D zijn *factoren*: categorische predictoren van nominaal niveau

- x_1, x_2 zijn *covariaten*: numerieke predictoren van minstens intervalniveau

3. **Het multivariaat lineair model**

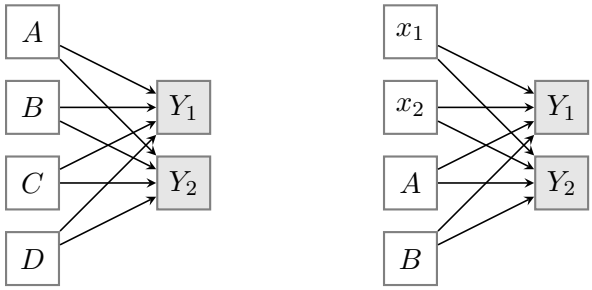
   modelleren van meerdere continue uitkomsten in functie van predictoren / onafhankelijke variabelen

   Multivariate regressie



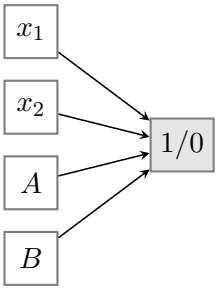
   - $\mathbf{Y} = [Y_1, Y_2]$: twee afhankelijke variabelen die conceptueel samenhoren
   - multivariate toetsen: houden rekening met de correlatiestructuur tussen de afhankelijke variabelen

   Manova en Mancova



4. **Het veralgemeend lineair model: logistische regressie**

   modelleren van 1 binaire uitkomst in functie van predictoren / onafhankelijke variabelen



5. **Meta-analyse**
   integreren van onderzoeksbevindingen uit verschillende studies

Statistische en praktische significantie

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

Dit hoofdstuk is deels gebaseerd op hoofdstuk 6 en 7 van Thompson (2006) en hoofdstuk 8 van Kline (2008).

1 Inleiding

Onderzoekers moeten vaak oordelen over de belangrijkheid of 'significantie' van hun resultaten van een empirische studie.

Er kan hierbij een onderscheid gemaakt worden tussen statistische en praktische significantie.

- **Statistische significantie**

  Een onderzoeker zet zijn/haar onderzoeksvragen om in **studiehypothesen** die in functie van 1 of meerdere **populatieparameters** uitgedrukt kunnen worden.

  Bij statistische toetsen hoopt men op basis van de steekproef een **nulhypothese (H_0)** overtuigend te kunnen **verwerpen** en zo op een alternatieve hypothese (d.i. de studiehypothese, H_1) te kunnen overstappen.

  Omdat het makkelijker is om uitspraken empirisch te weerleggen dan om ze te bewijzen, ligt de nadruk op het verwerpen van H_0.

  Voorbeelden

  - mannen roken meer dan vrouwen (H_1)
  - er is een verschil in eindresultaat tussen Leermethode A en Leermethode B (H_1)
  - het gemiddeld IQ in Vlaanderen is 110 (H_0)

  Een veelgebruikte maat hierbij is de p-waarde, dit is de kans om een resultaat te bekomen dat minstens even extreem is als het geobserveerde als de nulhypothese H_0 waar is.

  Kleine p-waarde $\rightarrow$ sterke aanwijzing tegen de nulhypothese.

- **Praktische significantie**

  p-waarden zijn berekend onder de veronderstelling dat de nulhypothese waar is. Ze meten dus enkel hoeveel bewijskracht er aanwezig is tegen de nulhypothese.

  De p-waarde incorporeert geen informatie over de waarde van de resultaten waar de onderzoeker belang aan hecht (bvb. grootte van een effect) en geeft dus geen uitsluitsel over het praktische belang van de resultaten.

  Bij praktische significantie ligt de nadruk op de grootte van het effect van een behandeling of op de mate waarin variabelen met elkaar geassocieerd zijn.

  Voorbeelden

  - Hoe sterk is roken met geslacht geassocieerd?
  - Hoe groot is het verschil in eindresultaat tussen leermethode A en leermethode B?

  Het is mogelijk dat heel kleine effecten die heel nauwkeurig gemeten zijn, een kleine p-waarde opleveren en dus statistisch significant bevonden worden maar in de praktijk weinig relevant zijn.

  De resultaten van twee verschillende studies waarvoor de p-waarden identiek zijn, zijn niet noodzakelijk praktisch even significant.

Bij klinische significantie ligt de nadruk op het schatten van de grootte van een behandelingseffect. Het onderzoek richt zich hier voornamelijk op een uitkomst waarvoor erkende diagnostische testen bestaan, bvb. de graad van depressie.

Voorbeeld

Wanneer klinisch depressieve personen at random toegekend worden aan een nieuwe therapie of een controlegroep, welk percentage van de patiënten is niet langer depressief op het einde van de interventie? Hoe verhoudt dit percentage zich tot het percentage in de controlegroep?

De ervaring leert dat naast andere zaken die het interpreteren van onderzoeksresultaten bemoeilijken (zoals vertekening, confounding), het verkeerd interpreteren van statistische resultaten ook vaak voorkomt. Dit is niet enkel een probleem voor de interpretatie van de eigen onderzoeksresultaten (bvb. voor scripties) maar ook bij het evalueren van resultaten uit de vakliteratuur.

Onthoud dat het een alerte en getrainde geest vergt om verantwoorde conclusies te kunnen trekken op basis van aangeleverd feitenmateriaal!

2 Statistische significantie

In deze sectie gaan we (op een eerder informele manier) dieper in op het toetsen van hypothesen zonder in detail op berekeningen in te gaan. Deze zijn aan bod gekomen in de cursus Statistiek II.

Merk op dat in dit hoofdstuk de Bayesiaanse benadering voor toetsen van hypothesen niet besproken wordt. Typisch voor deze benadering is dat men bepaalt hoe waarschijnlijk de nulhypothese is, gegeven de geobserveerde data.

2.1 Toetsen van hypothesen

Bij statistische toetsen is de nulhypothese bijna altijd uitgedrukt als de hypothese van 'geen effect'.

Voorbeeld

H_0: het gemiddelde studieresultaat bij Leermethode A (μ_A) is gelijk aan het gemiddelde studieresultaat bij Leermethode B (μ_B) of $H_0 : \mu_A = \mu_B$.

Toetsen van hypothesen is het proces waarbij we op basis van statistische significantie beslissen om de nulhypothese al dan niet te verwerpen. In het Engels spreekt men van 'null hypothesis statistical significance testing' (NHSST) of kortweg 'Null Hypothesis Significance Testing' (NHST).

Belangrijk: de nul- en alternatieve hypothese zijn altijd uitgedrukt in termen van populatieparameters. We wensen immers meer te weten te komen over onderzoeksstellingen m.b.t. deze parameters en we baseren ons op schatters in een steekproef om tot een besluit te komen.

2.1.1 De p-waarde

Ronald A. Fisher (1890-1962) wordt beschouwd als één van de grondleggers van de moderne statistische besluitvorming én van de statistische genetica. Hij was werkzaam op het gebied van de statistiek, evolutietheorie en erfelijkheidsleer. In zijn publicatie *Statistical Methods for research workers* (1925) introduceerde hij o.a. het idee van significantieniveaus als een middel om het verschil te onderzoeken tussen de geobserveerde data en wat men verwacht te zien onder de nulhypothese. Dit is wat we nu kennen als de p-waarde.

Fisher stelde de p-waarde voor als een maat om de discrepantie tussen de data en de nulhypothese te kwantificeren. Zijn redenering was dat hypotheses nooit bewezen kunnen worden maar enkel weerlegd (gefalsifieerd). M.a.w. de nulhypothese kan nooit bewezen en dus nooit aanvaard worden, maar kan wel verworpen worden. Stel dat we geïnteresseerd zijn in het testen van de volgende hypotheses: $H_0 : \Delta = 0$ met bvb. $\Delta = \mu_A - \mu_B$. H_0 drukt dan uit dat er geen verschil is in het gemiddelde resultaat tussen methode A en methode B.
De alternatieve hypothese stelt dat er wel een verschil is in gemiddelde resultaat, bvb. $H_1 : \mu_A > \mu_B$ of kortweg $H_1 : \Delta > 0$.

$H_1 : \Delta \neq 0$ en $H_1 : \Delta < 0$ zijn ook mogelijke alternatieve hypotheses. Om de concepten m.b.t. het toetsen van hypothesen uit te leggen, beschouwen we hier enkel $H_1 : \Delta > 0$.

Om H_0 tegenover H_1 te toetsen kunnen we gebruik maken van een toetsingsgrootheid of teststatistiek T (notatie: afkorting, heeft niets te maken met de t-verdeling).

Δ wordt geschat op basis van een steekproef, we noteren deze schatter met $\hat{\Delta}$. We noteren de standaardfout als $SE(\hat{\Delta})$.

Een toetsingsgrootheid T houdt niet enkel rekening met de grootte van het geobserveerde effect, $\hat{\Delta}$, maar ook met de precisie van de schatter. Om bovenstaande hypothesen te toetsen ziet T er typisch als volgt uit:

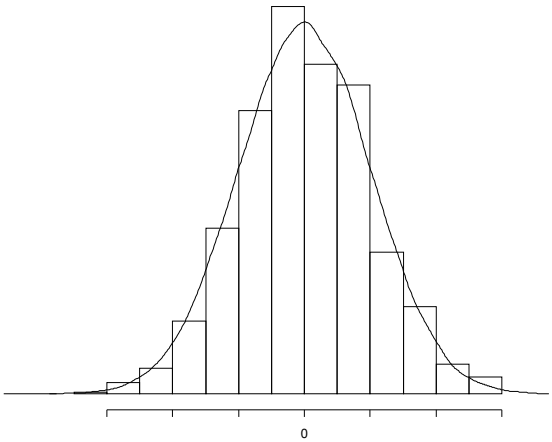
$$T = \frac{\hat{\Delta}}{SE(\hat{\Delta})}.$$

Het geobserveerde effect wordt gestandaardiseerd onder H_0: $\frac{\hat{\Delta}-0}{SE(\hat{\Delta})}$.

In het geval dat we 2 gemiddeldes vergelijken, is de waarde 'verschil in gemiddeldes/standaardfout(verschil)' gekend als een z-statistiek (variantie(s) gekend). Deze drukt de afwijking van 0 uit (in aantal standaardfouten).

T is een toevalsveranderlijke (denk aan de steekproevenverdeling van $\hat{\Delta}$ - notatie: grote letter). De geobserveerde toetsingsgrootheid in een welbepaalde steekproef noteren we als volgt: $t = d^*/se(\hat{\Delta})$.

Om de significantie van het geobserveerde resultaat na te gaan, gaan we kijken naar de verdeling van de toetsingsgrootheid onder de nulhypothese. Deze verdeling noemen we de **nuldistributie**. Concreet bestaat de nuldistributie uit de verzameling van alle mogelijke waarden die T aanneemt als H_0 waar is, i.e. als $\Delta = 0$.



- Conceptueel: toetsingsgrootheid = gestandaardiseerde steekproevenverdeling

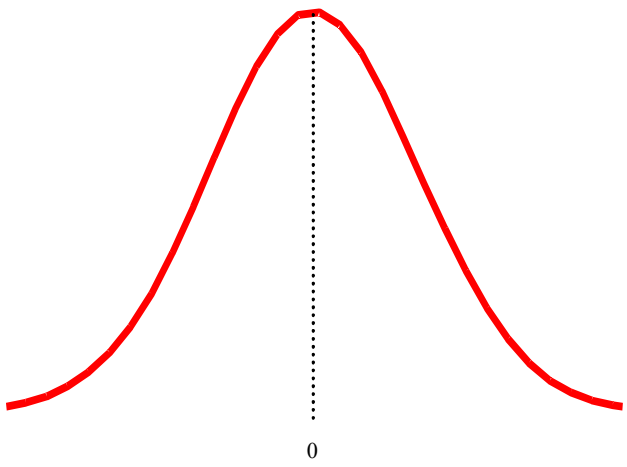
- Nuldistributie van toetsingsgrootheid = de kansverdeling van T onder H_0.

In de praktijk is het onmogelijk om herhaaldelijk steekproeven te trekken onder de nulhypothese om de nuldistributie van T te bepalen.

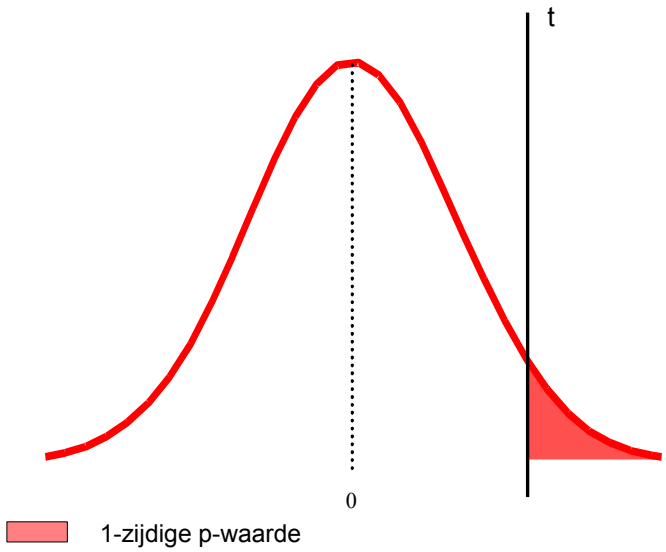
Voor veel toetsingsgrootheden kan de nuldistributie benaderd worden door een theoretische verdeling.

- De toets voor 1 gemiddelde of het vergelijken van 2 gemiddeldes: normale verdeling (variantie(s) gekend). Denk aan de centrale limietstelling!
- De toets voor 1 gemiddelde of het vergelijken van 2 gemiddeldes: t-verdeling (variantie(s) geschat op basis van steekproef).
- De toets voor het vergelijken van varianties d.m.v. een verhouding: F-verdeling.

Nuldistributie voor T:



De p-waarde voor bovenstaande toets is aangeduid op de onderstaande figuur:



De p-waarde is de kans om, indien H_0 waar is, een resultaat te bekomen dat minstens even extreem is als het geobserveerde resultaat.

Wat betekent dit concreet?

Hoe kleiner de p-waarde, hoe sterker de bewijskracht tegen de nulhypothese. De p-waarde meet de bewijskracht tegen H_0 in de richting van H_1.

Fisher stelde als voorbeeld de grens van 5% voor om te besluiten dat er voldoende bewijskracht is tegen de nulhypothese, dit diende niet als een 'gouden standaard'.

2.1.2 Foutmaten

Jerzy Neyman (1894-1981) en Egon Pearson (1895-1980) hadden een andere visie op het toetsen van hypothesen: hun standpunt bestond erin om procedures te ontwikkelen die toelaten om het aantal vergissingen of foute beslissingen (over vele experimenten) te beperken. Aan iedere statistische toets waarbij we binaire beslissingen nemen zijn foutenmarges verbonden.

Als men vertrekt van een nulhypothese H_0 die men (op basis van de gegevens) niet of wel verwerpt, leidt dit tot vier mogelijke situaties:

	Nulhypothese is	
	juist	fout
H_0 verwerpen	foute beslissing Type I fout α	juiste beslissing $1-\beta$
H_0 niet verwerpen	juiste beslissing $1-\alpha$	foute beslissing Type II fout β

Hierbij is:

$$\alpha \quad = \quad P(\text{Verwerp } H_0 | H_0 \text{ is juist})$$

$$\begin{aligned} & \Rightarrow \quad 1-\alpha = P(\text{Verwerp } H_0 \text{ niet} | H_0 \text{ is juist}) \\ \beta \quad & = \quad P(\text{Verwerp } H_0 \text{ niet} | H_0 \text{ is fout}) \\ & \Rightarrow \quad 1-\beta = P(\text{Verwerp } H_0 | H_0 \text{ is fout}) \end{aligned}$$

De twee types fouten zijn **Type I** en **Type II** fouten.

- Een Type I fout komt voor wanneer men H_0 verwerpt terwijl H_0 waar is. α is de kans op een Type I fout.

  Gevolg van het maken van een Type I fout: het verder onderzoeken van een 'vals alarm'.

- Een Type II fout komt voor wanneer men H_0 niet verwerpt terwijl H_0 fout is. β is de kans op een Type II fout.

  Gevolg van het maken van een Type II fout: een potentieel belangrijk effect wordt niet gedetecteerd.

  $1-\beta$ is kans dat een foute nulhypothese verworpen wordt. Deze kans wordt ook de **kracht** of het **onderscheidingsvermogen** van een toets genoemd (Engelse term: power).

Afhankelijk van het soort onderzoek, en de aard van de hypothese hecht de onderzoeker zeer veel, of zeer weinig belang aan hetzij Type I fouten, hetzij Type II fouten.

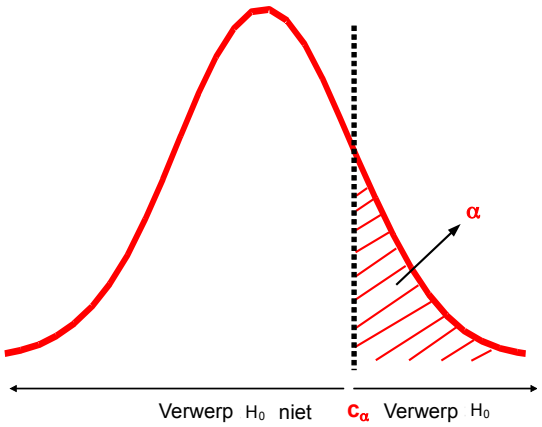
In de praktijk weten we uiteraard niet of we een fout maken bij het al dan niet verwerpen van H_0. We kunnen enkel de kans op een Type I en de kans op een Type II fout berekenen.

Bij het toetsen van een hypothese is het onmogelijk om zowel een Type I als een Type II fout te maken:

- Eenmaal de nulhypothese verworpen wordt, is het niet mogelijk dat er een Type II fout gemaakt is.
- Wanneer de nulhypothese niet verworpen wordt, is een Type I fout onmogelijk.

2.1.3 Beslissingscriteria

Om een beslissingcriterium te definiëren, legt men typisch een bovengrens op aan de kans op het maken van een Type I fout. Deze kans α is het significantieniveau.



Concreet: we verwerpen H_0 indien $T \geq c_\alpha$. Dit betekent dat H_0 verworpen wordt op het $\alpha \times 100\%$ significantieniveau. Dit is equivalent met het verwerpen van H_0 indien de p-waarde kleiner is dan of gelijk aan α. Op die manier wordt de kans op een Type I fout gecontroleerd.

Daaruit volgt de kans op het maken van een Type II fout.

α wordt vaak gelijk aan 5% gekozen in navolging van Fishers afkappunt van 5% voor de p-waarde, dit afkappunt beschouwde Fisher echter niet als een foutmaat. (Fisher: een kleine p-waarde toont enkel dat er zich ofwel een zeldzame gebeurtenis heeft voorgedaan ofwel dat de nulhypothese (theorie) fout is, er worden geen Type I en Type II fouten beschouwd).

Deze keuze is echter arbitrair: er is geen enkele reden waarom andere waarden niet mogelijk zijn.

De onderzoeker moet zelf beslissen welke waarde voor α (en dus welke betrouwbaarheid) de meest geschikte is voor elke concrete situatie.

Belangrijk is wel dat α vastgelegd wordt vóór het verzamelen van de data en vóór het berekenen van de toetsingsgrootheid.

1. Onderzoeksvraag $\rightarrow$ hypotheses
2. Verzamelen data
3. Toetsingsgrootheid
4. p-waarde
5. α bepaalt het uiteindelijke beslissingscriterium.

Wanneer we H_0 verwerpen, spreken we van een statistisch significant resultaat.

Merk op dat het wenselijk is om te zeggen 'H_0 kan niet verworpen worden' i.p.v. 'H_0 wordt aanvaard'. De reden hiervoor is dat het niet kunnen verwerpen van H_0 niet impliceert dat H_0 correct is, we hebben enkel niet voldoende bewijs om de nulhypothese te verwerpen.

2.2 Statistische power

2.2.1 Steekproefgrootte en precisie

De precisie waarmee een schatter gemeten is, is omgekeerd evenredig met de standaardfout. Hoe kleiner de standaardfout, hoe groter de precisie.

De standaardfout is steeds een functie van de steekproefgrootte n. Denk bvb. aan de standaardfout van het steekproefgemiddelde $\bar{X}$: $SE(\bar{X}) = \sigma/\sqrt{n}$ waarbij σ de standaarddeviatie is van de scores voor X.

In het algemeen geldt:
hoe groter de steekproefgrootte, hoe kleiner de standaardfout van de steekproevenverdeling van een statistiek.

Een manier om de precisie van een schatter te verhogen, is een grotere steekproef trekken.

Toetsingsgrootheden nemen de standaardfout mee in rekening. Dit betekent: hoe preciezer de schatter, hoe sneller H_0 verworpen wordt en dus hoe kleiner de kans op een Type II fout (voor een vaste α).

2.2.2 Power

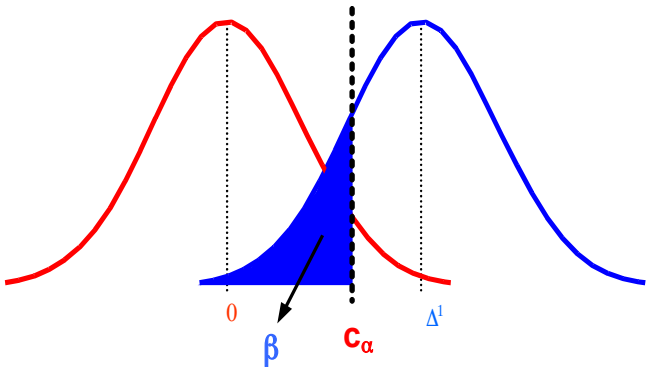
De power of het onderscheidingsvermogen van een toets is het complement van de kans op een Type II fout, i.e. power $= 1 - \beta$. De kans op een Type II fout is in de praktijk typisch een stuk groter dan de kans op

een Type I fout. Vaak wordt een power van (minstens) 80% aangeraden. Men moet er zich bewust van zijn dat er dan bij het niet kunnen verwerpen van H_0 nog steeds een aanzienlijke kans op een Type II fout is.

Veronderstel nog steeds dat we $H_0 : \Delta = 0$ tegenover $H_1 : \Delta > 0$ toetsen.

Bij het berekenen van de kans op een Type II fout, moeten we een welbepaald alternatief specifiëren (zie ook Neyman-Pearson).

Stel dat het onderliggende effect gelijk is aan Δ^1 met $\Delta^1 \neq 0$, wat is dan de kans op een Type II fout? Dit wordt geïllustreerd op de onderstaande figuur:



Power van een toets: P(Verwerp H_0 | $\Delta = \Delta^1$)
=1 - β

Kritiek op de benadering van Neyman-Pearson:

- de nood om een welbepaald alternatief te specifiëren,
- de complexiteit van de power omdat deze afhangt van parameters die vaak ongekend zijn (het alternatief en σ; verdeling van T onder H_1).

Het gebruik van statistiek wordt vaak (onterecht) gedomineerd door het categoriseren van resultaten als 'statistisch significant' of 'statistisch niet-significant' zonder veel aandacht te besteden aan de Type II fout.

Powerberekeningen en het berekenen van de steekproefgrootte zijn echter van groot belang bij empirisch onderzoek. Het specifiëren van een alternatief reflecteert waar de onderzoeker belang aan hecht (een minimaal effect dat aanwezig dient te zijn opdat de resultaten de moeite waard zouden zijn) of wat de onderzoeker verwacht (op basis van voorgaande studies). Het is dus logisch en noodzakelijk dat dit ook beschouwd wordt bij het opzetten en evalueren van een studie.

De effectgrootte (Engelse term: effect size) is het verschil tussen de vooropgestelde waarden voor H_0 en H_1. In ons voorbeeld is dit Δ^1.

De effectgrootte is sterk afhankelijk van de schaal die gehanteerd wordt. De **gestandaardiseerde** effectgrootte is onafhankelijk van de schaal en wordt als volgt gedefinieerd:

$$\delta = \frac{\text{effectgrootte}}{\sigma}$$

met σ de standaarddeviatie van de scores in de populatie.

Wanneer 3 van de 4 onderstaande elementen gekend zijn, kan het 4de element berekend worden:

- de steekproefgrootte n
- het significantieniveau α
- de kans op een Type II fout β
- de gestandaardiseerde effectgrootte δ

Er geldt:

- Hoe groter n, hoe kleiner β (bij constante α en δ)
- Hoe groter δ, hoe kleiner β (bij constante α en n). Dit betekent dat de power van een toets groter wordt voor een grotere effectgrootte en/of een kleinere standaarddeviatie σ.
- Hoe groter α, hoe kleiner β (bij constante n en δ). Indien α toeneemt, wordt een toets immers minder streng en bijgevolg wordt H_0 sneller verworpen.

Zie o.a. Statistiek II voor powerberekeningen in R.

2.2.3 Underpowered studies

In 2015 werd in Science een studie gepubliceerd door een grote groep onderzoekers over reproduceerbaarheid in psychologisch onderzoek (Estimating the reproducibility of psychological science, Open Science Collaboration 2015). In deze studie stelde men vast dat hoewel 97% van de onderzochte artikels statistisch significante resultaten publiceerde, men slechts 37% van deze effecten als statistisch significant bevond in een poging tot replicatie. Reproduceerbaarheid binnen psychologisch onderzoek kreeg sindsdien veel aandacht maar het is belangrijk te beseffen dat deze 'reproducibility crisis' zich niet beperkt tot dit vakgebiek maar een probleem vormt binnen wetenschap in het algemeen (biologie, scheikunde, economie, sociale wetenschappen, ...).

Eén van de zaken die mee aan de basis kan liggen van problemen met reproduceerbaarheid, is het gebruik van studies met een lage power. Bij studies die onvoldoende statistische power hebben om een welbepaald effect te detecteren, spreekt men van *underpowered studies*. Te weinig power betekent niet enkel dat er een kleine kans is om het vooropgestelde effect te detecteren maar ook dat de kans afneemt dat een statistisch significant resultaat een echt effect reflecteert. Die laatste kans is de *positieve predictieve waarde* van een test.

Als we een set van studies hebben met een statistisch significant resultaat m.b.t. een welbepaalde hypothese, dan is de positieve predictieve waarde van de test de proportie van deze studies waarbij onderliggend een echt effect aanwezig is. Als alle overige factoren constant blijven, dan neemt de positieve predictieve waarde af wanneer de statistische power lager wordt.

Dit wordt toegelicht door Ioannidis (2005) in het artikel *'Why Most Published Research Findings Are False'*.

De positieve predictieve waarde PPW van een toets kunnen we als volgt schrijven:

$$PPW = P(H_0 \text{ is fout} | H_0 \text{ wordt verworpen}) = P(H_0 \text{ is fout} | \text{statistisch significant resultaat})$$

Hierbij gebruiken we *conditionele kansen*. De conditionele kans op $Y = y$ gegeven $X = x$, genoteerd als $P(Y = y|X = x)$, drukt de kans uit dat $Y = y$ als men al weet dat $X = x$. Er geldt (regel van Bayes):

$$P(Y = y|X = x) = \frac{P((Y = y) \text{ en } (X = x))}{P(X = x)}.$$

Laat $1-\beta$ de power van een statistische toets voorstellen en α het significantieniveau. Volgens de regel van Bayes kunnen we de PPW als volgt herschrijven:

$$\begin{aligned} PPW &= \frac{P(H_0 \text{ is fout en statistisch significant resultaat})}{P(\text{statistisch significant resultaat})} \\ &= \frac{P(\text{statistisch significant resultaat}|H_0 \text{ is fout}) \times P(H_0 \text{ is fout})}{P(\text{statistisch significant resultaat})} \\ &= \frac{(1-\beta) \times P(H_0 \text{ is fout})}{P(\text{statistisch significant resultaat})} \end{aligned}$$

De noemer kunnen we herschrijven als volgt:

$$\begin{aligned} &P(\text{statistisch significant resultaat}|H_0 \text{ is waar}) \times P(H_0 \text{ is waar}) \\ &+P(\text{statistisch significant resultaat}|H_0 \text{ is fout}) \times P(H_0 \text{ is fout}) \end{aligned}$$

en dus vinden we dat

$$PPW = \frac{(1-\beta) \times P(H_0 \text{ is fout})}{\alpha \times P(H_0 \text{ is waar}) + (1-\beta) \times P(H_0 \text{ is fout})}.$$

Daarin zien we dat de PPW afhangt van α, β en de prevalentie π van een effect, namelijk $\pi = P(H_0 \text{ is fout}) = 1 - P(H_0 \text{ is waar})$. De prevalentie is meestal ongekend.

Neem als voorbeeld een effect dat met een grote kans aanwezig is, bvb. $\pi = 0.90$. Wanneer $(1-\beta) = 0.8$ (vaak aanbevolen als een minimum) en $\alpha = 0.05$, dan krijgen we de volgende PPW:

```
> prev<-0.90
> alpha<-0.05
> beta<-0.20
> ppw<-(1-beta)*prev/(alpha*(1-prev)+(1-beta)*prev)
> ppw
[1] 0.9931034
```

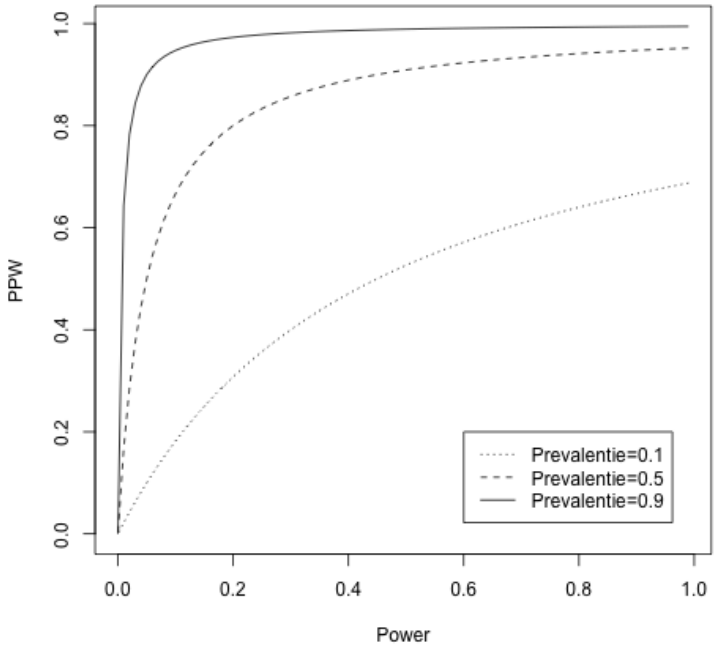
99% betekent dat we vrij veel vertrouwen kunnen hebben in het feit dat wanneer een statistisch significant resultaat bekomen wordt, er een echt effect aanwezig is. Wanneer de prevalentie van een effect een stuk kleiner wordt (bvb. $\pi = 0.10$), dan daalt deze PPW echter al serieus:

```
> prev<-0.10
> alpha<-0.05
> beta<-0.20
> ppw<-(1-beta)*prev/(alpha*(1-prev)+(1-beta)*prev)
> ppw
[1] 0.64
```

Wanneer bijkomend de power nog lager wordt, wordt dit problematisch:

```
> prev<-0.10
> alpha<-0.05
> beta<-0.40
> ppw<-(1-beta)*prev/(alpha*(1-prev)+(1-beta)*prev)
> ppw
[1] 0.5714286
```

Onderstaande figuur toont de relatie tussen power en de positieve predictieve waarde voor verschillende prevalenties. Hierbij wordt $\alpha = 0.05$ verondersteld.



2.3 Statistische significantie in de praktijk

Fishers grootste kritiek op Neyman-Pearson was dat het rapporteren van de kans op een Type I en een Type II fout, de variatie van de bewijskracht in de data niet reflecteert (Berger, 2003). M.a.w. het rapporteren van α ongeacht de grootte van de geobserveerde toetsingsgrootheid t vond hij onwetenschappelijk.

Anderzijds bekritiseerde Neyman de p-waarden omwille van het feit dat ze niet volgens het frequentistisch principe geïnterpreteerd kunnen worden. In tegenstelling tot α, kan een p-waarde niet gezien worden als een gemiddelde fout over een lange reeks experimenten.

Het feit dat Fisher enerzijds en Neyman en Pearson anderzijds niet op dezelfde golflengte zaten m.b.t. het toetsen van hypothesen, heeft gezorgd voor verwarring en verkeerd gebruik van toetsingsmethoden in de wetenschappelijke gemeenschap (Berger, 2003). In de praktijk worden bijna altijd p-waarden gebruikt maar vaak gaat men dan de p-waarde interpreteren als een foutmaat, namelijk als de kans op een Type I fout (frequentistisch). Dit leidt tot het overschatten van de bewijskracht tegen H_0.

α wordt op voorhand gespecifieerd en enkel het feit dat de p-waarde kleiner is dan of gelijk aan α is van belang bij Neyman-Pearson (die strikt genomen geen p-waarden beschouwden) en niet de grootte van de p-waarde.

Samengevat: α en β zijn foutmaten, maar de p-waarde is een conditionele kans die niet geïnterpreteerd kan worden als een foutmaat.

De p-waarde is de overschrijdingskans die we bekomen op basis van de geobserveerde data *op voorwaarde dat de nulhypothese waar is*. De p-waarde is dus ook niet de kans dat de nulhypothese waar is.

Verkeerde interpretaties en misverstanden over de p-waarde en foutmaten vormen nog vaak een probleem bij wetenschappelijk onderzoek. Het artikel van Regina Nuzzo (2014,

`http://www.nature.com/news/scientific-method-statistical-errors-1.14700`) over 'Statistical Errors' geeft een kort overzicht van deze problematiek.

Een kleine p-waarde wijst bovendien niet noodzakelijk op een betrouwbaar of repliceerbaar effect. De reproducibility crisis heeft ook duidelijk gemaakt dat in het verleden vaak te weinig aandacht geschonken werd aan replicatiestudies. De focus lag te veel op statistisch significante resultaten en het vernieuwende karakter van studies.

Als reactie op deze problemen heeft het tijdschrift **Basic and Applied Social Psychology** in 2015 beslist om het gebruik van NHST en p-waarden volledig te bannen. Het is echter niet zozeer het gebruik van deze technieken dat problematisch is, wel het verkeerd gebruik en het gelijkstellen van statistische significantie aan belangrijkheid. Een algemene noot van kritiek op NHST is inderdaad de sterkte van de conclusies die er op gebaseerd worden.

In 2016 heeft de American Statistical Association een statement rond p-waarden gepubliceerd. Daarin worden de meest courante misvattingen belicht. De hoofdboodschap is verder dat wetenschappelijke conclusies en beleidsbeslissingen nooit enkel gebaseerd mogen zijn op het feit of een p-waarde al dan niet kleiner is dan een specifieke afkapwaarde (zoals 0.05).

Onderstaande tabel uit Kline (2009, pag. 129) vat goed samen op welke niveaus het kan mislopen bij de interpretatie van statistische toetsen.

Table 5.3. Other Common Misinterpretations of Statistical Significance

Type	Description
H_0 rejected	
Magnitude	Statistical significance -> large effect size
Meaningfulness	Statistical significance -> theory behind alternative hypothesis is correct
Causality	Statistical significance -> causal mechanism is known
Quality	Statistical significance -> good design quality
Success	Statistical significance -> study is successful
H_0 not rejected	
Failure	Lack of statistical significance -> study is a failure
Zero	Lack of statistical significance -> zero population effect size
Equivalence	Failure to reject H_0: $\mu_1 - \mu_2 = 0$ -> two populations are equivalent
Other	
Reifiction	H_0 rejected in first study but not in second -> no evidene for replication
Sanctification	$p < .05$ is very meaningful, but $p = .06$ is not
Objectivity	Statistical tests are objective; all other methods are subjective

Note. ->, implies.

Nog enkele bijkomende opmerkingen:

- In het wetenschappelijk tijdschrift *Psychological Science* stelt Killeen (2005) een alternatief voor voor NHST, namelijk het berekenen van de kans op replicatie bij een welbepaald effect. In plaats van een p-waarde wordt dan p_{rep} gerapporteerd.

  p_{rep}: wat is de gemiddelde kans dat een replicatie een resultaat geeft in dezelfde richting?

  Deze aanpak is veelbelovend maar is in zijn huidige vorm niet vrij van kritiek: Iverson en Wagenmakers (2009) tonen aan dat de schatter voor replicatie die voorgesteld wordt niet optimaal is. Bovendien is p_{rep} volgens Iverson, Wagenmakers en Lee (2010) enkel een bovengrens voor de kans op replicatie.

- Andere auteurs (zoals Matthews, 2011) tonen aan hoe de Bayesiaanse aanpak voor toetsen van hypothesen nadelen en problemen vermijdt die aanwezig zijn bij NHST (Fisher, Neyman-Pearson). De Bayesiaanse aanpak benadrukt de keuze tussen 2 hypothesen en kwantificeert bewijskracht in de data.

2.4 Questionable research practices

De misvattingen rond NHST, het verkeerd interpreteren van resultaten en de overmatige focus op statistische significantie in de wetenschappelijke literatuur, hebben geleid tot enkele praktijken bij het uitvoeren van empirische studies die men benoemt met 'questionable research practices'. Dit zijn aanpakken die wetenschappelijk niet correct zijn maar vaak worden deze technieken niet te kwader trouw gebruikt.

Merk op dat niet enkel NHST een aanleiding geweest is voor het ontstaan en gebruik van 'questionable research practices'; we bespreken hier enkele gevallen die wel passen binnen deze context maar het probleem is breder dan louter een probleem bij statistische besluitvorming.

2.4.1 *p*-hacking

p-hacking is een brede term die verwijst naar het analyseren van data op verschillende manieren om ervoor te zorgen dat resultaten statististisch significant worden. Hoewel data misschien wel op verschillende manieren geanalyseerd kunnen worden, worden vaak enkel de resultaten gerapporteerd voor de analyses die de 'gewenste' resultaten opleveren. Het is dit proces dat problematisch is.

Enkele voorbeelden:

- Selectief verwijderen van outliers
- Observaties verwijderen
- Data verzamelen tot $p < 0.05$, dan stoppen of extra data verzamelen als resultaten niet statistisch significant zijn

  Onthou dat H_0 altijd verworpen zal kunnen worden als de steekproefgrootte maar groot genoeg wordt. Dit betekent dat bij grote steekproeven zelfs de kleinste afwijking van H_0 gedetecteerd zal worden.
- Verschillende condities creëren binnen een experimentele opzet maar enkel deze rapporteren waarvoor resultaten statistisch significant zijn
- Verschillende variabelen analyseren maar enkel rapporteren over statistisch significante resultaten

p-hacking verklaart het fenomeen waarbij men vaststelde dat er in de literatuur vaker dan verwacht *p*-waarden gerapporteerd worden die dicht in de buurt van 0.05 liggen (en net kleiner zijn).

2.4.2 HARKing

HARKing is 'Hypothesizing After the Results are Known'. Hier zal een onderzoeker eerst de resultaten bekijken en nadien hypotheses opstellen die overeenkomen met de resultaten. Op die manier wordt van een mogelijks toevallige post hoc vaststelling een a priori predictie gemaakt.

Dit is uiteraard een schadelijk proces voor het wetenschappelijk onderzoek en maakt het onmogelijk om hypotheses te falsifiëren aangezien altijd gezocht wordt naar hypotheses die overeenstemmen met de data.

Voorbeeld

In een onderzoek naar een aandoening waarbij de hoofdhypotheses stellen dat er een effect is van omgevingsfactoren, merkt men post hoc ook een verband tussen slaapgewoontes en de aandoening. In het geval van HARKing, gaan de onderzoekers deze verbanden omzetten in een hypothese en de studie beschrijven alsof dit vooraf voorspeld werd. Op die manier kan de "hypothese" bevestigd worden.

De juiste oplossing is om aan te geven dat naast wat onderzocht werd voor de hoofdhypotheses ook enkele verkennende, post-hoc analyses gebeurd zijn die mogelijks interessante pistes voor nieuw onderzoek bieden.

Dit kan hard gemaakt worden via **pre-registratie**: dit is een systeem waarbij onderzoekers voor de start van een studie hun hypotheses en werkwijze registreren (zie bvb. open science framework `https://osf.io`). Deze kunnen nadien niet meer gewijzigd worden. Op die manier is het duidelijk wat de resultaten zijn voor de hypotheses en welke analyses post-hoc gebeurd zijn en dus eerder verkennend van aard.

Sommige wetenschappelijke tijdschriften bieden peer-review aan voor het design van een studie (hoe men de studie gaat uitvoeren), het opstellen van hypotheses, de uit te voeren analyses etc. Wanneer dit goedgekeurd is, kunnen de onderzoekers aan de slag om de studie uit te voeren en data te verzamelen met de garantie dat het artikel gepubliceerd wordt, ongeacht de uitkomst. Op die manier zijn resultaten in de literatuur een betere neerslag van de onderliggende werkelijkheid, o.a. omwille van het feit dat het wetenschappelijke proces met zekerheid gerespecteerd wordt en dat bovendien niet enkel statistisch significante resultaten worden gepubliceerd. Als een echt effect aanwezig is, verwacht men in een aantal studies vals negatieve resultaten, dit zijn de Type II fouten. Analoog voor Type I fouten wanneer geen effect aanwezig is.

3 Praktische significantie

Kritiek op het gebruik van NHST betekent niet dat het een waardeloos proces is. Kritiek is deels te wijten aan het feit dat resultaten van NHST vaak verkeerd geïnterpreteerd worden en er op die manier verkeerde en/of te sterke conclusies getrokken worden. NHST is zinvol als we er de gepaste conclusies op baseren. Bovendien mogen we de resultaten van een studie niet beperken tot enkel de resultaten van NHST .

De American Psychological Association (APA) stelde in 1996 een Task Force on Statistical Inference samen waarvan de taak o.a. was om enkele controversiële kwesties toe te lichten omtrent de toepassingen van statistiek, inclusief toetsen voor significantie en mogelijke alternatieven hiervoor. De aanbevelingen van de Task Force werden 3 jaar later gepubliceerd (Wilkinson, 1999). Meer informatie kun je terugvinden op `http://www.apa.org/science/leadership/bsa/statistical/`.

Deze aanbevelingen werden opgenomen in de APA publication manual die gebruikt wordt door meer dan 1000 tijdschriften.

Twee vormen voor de bewijskracht van praktische significantie worden benadrukt: effectgroottes en betrouwbaarheidsintervallen.

3.1 APA Task Force on Statistical Inference

Hieronder worden kort enkele (!) puntjes aangehaald uit de aanbevelingen van de Task Force zoals beschreven in Wilkinson (1999). De lijst hieronder kan in geen geval beschouwd worden als een volledige samenvatting of vervanging van de APA richtlijnen. Meer details kunnen in het artikel zelf of in de APA publication manual teruggevonden worden.

- **Design**: specifieer duidelijk welk type studie je uitvoert (case studies, gecontroleerde experimenten, quasi-experimenten, statistische simulaties, surveys, observationele studies of meta-analyses).
- **Populatie**: de interpretatie van de resultaten van elke studie hangt af van de karakteristieken van de populatie die onderzocht wordt. Definieer duidelijk deze populatie.
- **Steekproef**: beschrijf het steekproefplan en benadruk criteria voor inclusie of exclusie. Indien je een gemakssteekproef gebruikt, maak dit duidelijk. Het gebruik van een gemakssteekproef diskwalificeert een studie niet voor publicatie, maar doen alsof het een lukrake steekproef (toevalsteekproef) voorstelt is niet objectief.
- **Randomisatie (willekeurige toewijzing)**: wanneer je causale besluiten wenst te trekken, is het toekennen van de steekproefeenheden aan de verschillende niveaus van de onafhankelijke variabele cruciaal. Randomisatie laat de sterkst mogelijke causale besluiten zonder bijkomende assumpties toe. Wanneer randomisatie gepland is, voorzie dan voldoende informatie om aan te tonen dat het proces van het toewijzen effectief random was.
- **Niet-willekeurige toewijzing**: randomisatie is niet altijd mogelijk. In dat geval moeten de effecten van variabelen die de relatie tussen de uitkomst en de causale variabele verstoren (confounders) minimaliseren. Deze moeten bepaald worden door de onderzoeker, adequaat gemeten worden en er moet voor gecorrigeerd worden bij het design of in de analyse. In het laatste geval moeten assumpties expliciet vermeld worden en in de mate van het mogelijke getest en gerechtvaardigd worden. Er moet ook gespeculeerd worden over eventuele ongemeten confounders en hoe deze kunnen leiden tot incorrecte besluiten.

  Men raadt ook aan om in deze context de term ‘contrastgroep’ i.p.v. ‘controlegroep’ te gebruiken.
- **Metingen - variabelen**: definieer expliciet de variabelen in de studie, toon aan hoe ze verband houden met het doel van de studie en leg uit hoe ze gemeten worden.
- **Metingen - instrumenten**: in het geval een vragenlijst of test gebruikt wordt, vat de psychometrische eigenschappen samen m.b.t. de manier waarop het instrument gebruikt wordt in de populatie. Psychometrische eigenschappen omvatten validiteit, betrouwbaarheid en andere kwaliteiten die conclusies beïnvloeden.
- **Metingen - procedure**: beschrijf geanticipeerde bronnen van studie-uitval zoals non-compliance (= niet therapiegetrouw), dropout, sterfgevallen, ... Geef aan hoe studie-uitval de generalisatie van de resultaten beïnvloedt.

  Beschrijf de omstandigheden waarin de metingen gebeuren. Beschrijf de specifieke methodes om om te gaan met onderzoekers-bias (vooral als je de data zelf verzameld hebt).
- **Power en steekproefgrootte**: geef informatie over de steekproefgrootte en het proces dat geleid heeft tot beslissingen i.v.m. de steekproefgrootte. Geef informatie over de effectgroottes, assumpties m.b.t. steekproeftrekking en metingen alsook over de analytische procedures die gebruikt werden bij de power berekeningen.

  Aangezien dergelijke berekeningen slechts zinvol zijn vóór het verzamelen en analyseren van de data, is het belangrijk hoe de keuze van een effectgrootte bepaald werd door voorafgaand onderzoek of voorafgaande theorieën. Dit is belangrijk om te vermijden dat men de onderzoekers ervan verdenkt gebruik te maken van de effectgrootte in de data of erger, dat de onderzoekers een effectgrootte bepalen om hun gebruikte steekproefgrootte te verantwoorden.

  Eenmaal de resultaten geanalyseerd zijn, vervangen betrouwbaarheidsintervallen de powerberekeningen.
- **Effectgroottes**: geef altijd de effectgroottes voor de hoofduitkomsten.

Wanneer de meeteenheden zinvol zijn op een praktisch niveau (bvb. aantal sigaretten per dag), geeft de APA de voorkeur aan niet-gestandaardiseerde effectgroottes.

Het is zinvol om deze effectgroottes in een praktische en theoretische context te plaatsen.

- **Intervalschattingen**: Intervalschattingen moeten gegeven worden voor elke effectgrootte van de hoofduitkomsten. Intervallen voor correlaties of andere coëfficiënten van associatie of variatie moeten gegeven worden indien mogelijk.

In de APA manual wordt aangegeven dat het ontbreken van de effectgrootte een gebrek is bij een studie. Veel wetenschappelijke tijdschriften eisen expliciet het rapporteren van effectgroottes.

3.2 Effectgroottes

Effectgroottes hebben als doel de grootte van een effect te kwantificeren, los van de steekproefgrootte. p-waarden kunnen niet gebruikt worden om belangrijkheid van resultaten te vergelijken. Dat is de reden waarom men effectgroottes gebruikt in meta-analyses (het combineren van resultaten van verschillende studies, zie verder).

Vanaf nu verwijzen we met de term 'effectgrootte' enkel naar gestandaardiseerde effectgroottes. Waarom is standaardiseren belangrijk?

- In de geneeskunde rapporteren onderzoekers vaak niet-gestandaardiseerde effectgroottes, zoals het verschil tussen de gemiddelde cholesterol bij een behandelingsgroep en bij een controlegroep. Metingen gebeuren in natuurlijke en universele meeteenheden.
- In de gedragswetenschappen zijn er vaak geen universele meeteenheden om bepaalde constructen te meten. Om resultaten vergelijkbaar te maken tussen studies, is het dus noodzakelijk om de effecten te standaardiseren zodat de schaalafhankelijkheid verdwijnt.
- Door te standaardiseren, kunnen ook de effectgroottes van studies die verschillende variabelen meten, vergeleken worden (bvb. verschillende variabelen om depressie te meten).

3.2.1 Families van parametrische effectgroottes

We kunnen de volgende families van effectgroottes (op groepsniveau) voor continue uitkomsten onderscheiden:

- d-familie (groepsverschillen): gestandaardiseerde gemiddelde verschillen
- r-familie (samenhang): associatiematen

Effectgroottes kunnen ook op case-niveau bepaald worden, bvb. door het berekenen van de proportie van scores van verschillende groepen boven of onder bepaalde referentiepunten. We laten dit hier buiten beschouwing.

- **Gestandaardiseerde gemiddelde verschillen**

  Een effectgrootte op populatieniveau gebaseerd op het verschil in gemiddeldes voor 2 populaties (bvb. experimentele en controlegroep) is:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

waarbij σ de standaarddeviatie van 1 van beide populaties is of een gepoolde standaarddeviatie over 2 groepen.

In de praktijk maken we gebruik van schatters i.p.v. populatieparameters:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S}$$

waarbij S een schatter is voor de standaarddeviatie in de populatie; S kan hierbij verschillende vormen aannemen.

Een d-statistiek drukt een gemiddelde verschil uit als de proportie van de standaarddeviatie van de variabele waarover dit verschil gemeten is.

Voorbeeld

Als $d = 0.6$, dan is $\bar{X}_1$ 0.6 standaarddeviaties (S) hoger dan $\bar{X}_2$.

- **Associatiematen**

  Associatiematen geven aan in welke mate (afhankelijke en onafhankelijke) variabelen covariëren. Hieronder enkele veel gebruikte maten:

  - r: Pearson produkt-moment correlatie
    Indien $Cor(X, Y) = r$ dan is r^2 de proportie gedeelde variantie tussen X en Y.
  - Lineaire regressie: R^2 of determinatiecoëfficiënt (R: meervoudige correlatiecoëfficiënt)
    geeft het % weer van de variatie in de afhankelijke variabele (uitkomst) die verklaard wordt door de onafhankelijke variabelen (predictoren) in het model.

    $$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Totaal}}}$$

    SSR: regressie kwadratensom; SST: totale kwadratensom

3.2.2 Schatten van effectgroottes voor het vergelijken van 2 onafhankelijke groepen

- d-statistieken

  Laat $\bar{X}_1$ en $\bar{X}_2$ de steekproefgemiddeldes voor de 2 populaties voorstellen, gebaseerd op respectievelijk n_1 en n_2 observaties. S_1^2 en S_2^2 stellen de steekproefvarianties voor:

  $$S_1^2 = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}{n_1 - 1}$$

  met X_{1i} $(i = 1, \ldots, n_1)$ de individuele scores in populatie 1. S_2^2 wordt op een analoge manier berekend.

  Cohen's d

  $$d = \frac{\bar{X}_1 - \bar{X}_2}{S}. \tag{1}$$

  Cohen (1988) definieerde S niet expliciet. Hij argumenteerde wel dat de precisie waarmee σ geschat wordt, toeneemt indien beide groepen mee in rekening gebracht worden. S is dus een schatter die de varianties van beide groepen poolt.

Hedges's g

$$g = \frac{\bar{X}_1 - \bar{X}_2}{S_P}$$

met

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Merk op dat men hierbij de veronderstelling maakt van gelijke populatievarianties binnen beide groepen. Wanneer g op deze manier gedefinieerd wordt, is deze een vertekende schatter voor δ. De schatter wordt echter onvertekend door vermenigvuldiging met een factor (vooral belangrijk voor kleine steekproefgroottes), een benadering hiervoor is:

$$\frac{\bar{X}_1 - \bar{X}_2}{S_p} \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right).$$

Hedges en Olkin (1985) noteren Hedges's g ook als d. In wat volgt, gebruiken we d om te verwijzen naar de d-familie van effectgroottes; er zal dan telkens aangegeven worden over welke d het gaat.

Glass's Δ

Merk op: Δ is hier wel degelijk een statistiek en geen populatieparameter, niet te verwarren met Δ zoals voorheen.

Veronderstel dat populatie 2 de controlegroep is, dan is

$$\text{Glass' } \Delta = \frac{\bar{X}_1 - \bar{X}_2}{S_2}.$$

Enkel de standaarddeviatie van de controlegroep wordt gebruikt bij het standaardiseren.

Wanneer empirische resultaten reeds aangetoond hebben dat de behandeling niet enkel het gemiddelde maar ook de variantie beïnvloedt, kan men kiezen om met deze effectgrootte te werken.

Bedenk bovendien dat als verschillende behandelingen met de controlegroep vergeleken dienen te worden, de effectgrootte op die manier niet afhangt van eventueel verschillende varianties.

- r-statistieken

  De punt-biseriële correlatie r_{pb} is de Pearson correlatiecoëfficiënt tussen de uitkomst (waarvoor we een gemiddelde kunnen berekenen) en de groepsindicator (gecodeerd met 0 en 1). Dit betekent dat r bvb. gebruikt kan worden om de sterkte van een (experimenteel) effect te meten.

  Wanneer men resultaten van verschillende studies gaat bundelen, is het mogelijk dat sommige studies met d werken en andere met r. Bijgevolg moeten beide effectgroottes naar dezelfde schaal gebracht worden. Vaak wordt d omgezet naar r.

  Voorbeeld

  We beschouwen een fictieve dataset waarbij het de bedoeling is om het effect van een behandeling op een score te bepalen.

  - `Trt`: controlegroep (0) en behandelingsgroep (1)
  - `Score`

  Het aantal observaties in de controlegroep is 5 en in de behandelingsgroep 10.

We voeren nu een t-toets uit om na te gaan of er een verschil is tussen de gemiddelde score in beide groepen (de nulhypothese stelt dat er geen verschil is). Indien we veronderstellen dat de varianties van de scores in beide groepen gelijk zijn, dan ziet de toetsingsgrootheid er als volgt uit:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

met $\bar{X}_k$ het steekproefgemiddelde van de score in groep k ($k = 1, 2$), $n_2 = 10$ en $n_1 = 5$.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

met S_k^2 de steekproefvariantie van de score in groep k ($k = 1, 2$). Merk op dat d hier staat voor Hedges's g.

In het softwarepakket R krijgen we het volgende resultaat:

```
> t.test(score[trt==1],score[trt==0],var.equal=TRUE)

Two Sample t-test

data:  score[trt == 1] and score[trt == 0]
t = 1.6554, df = 13, p-value = 0.1218
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.2920579  2.2067767
sample estimates:
mean of x mean of y
6.086629  5.129270
```

Indien we nu de correlatie berekenen tussen `score` en `trt` en op basis daarvan toetsen of de correlatiecoëfficiënt statistisch significant verschillend is van 0, geeft R ons het volgende resultaat:

```
> cor.test(score,trt)

Pearson's product-moment correlation

data:  score and trt
t = 1.6554, df = 13, p-value = 0.1218
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1208496  0.7658210
sample estimates:
cor
0.4172434
```

We bekomen dezelfde resultaten. We stellen vast dat

$$r = \sqrt{\frac{t^2}{t^2 + (n-2)}} = \sqrt{\frac{t^2}{t^2 + 13}} = 0.42. \tag{2}$$

Dit illustreert het verband tussen d en r in dit voorbeeld.

Cohen (1988) stelde de volgende formule (benadering) voor om d om te zetten naar r wanneer beide groepen ongeveer even groot zijn:

$$r = \frac{d}{\sqrt{d^2 + 4}}.$$

Bijgevolg is

$$d = \frac{2r}{\sqrt{1-r^2}}.$$

Aaron et al. (1998) stellen voor om de exacte transformatie te gebruiken i.p.v. een benadering omdat deze slecht presteert bij kleine steekproeven en in het geval beide groepen niet even groot zijn.

Via (2) is het inderdaad relatief eenvoudig om af te leiden dat

$$r = \frac{d}{\sqrt{d^2 + \frac{n(n-2)}{n_1 n_2}}} \tag{3}$$

met $n = n_1 + n_2$. Wanneer $n_1 = n_2$ en dus $n_1 = n_2 = n/2$, herleidt deze formule zich tot

$$r = \frac{d}{\sqrt{d^2 + \frac{4(n-2)}{n}}}.$$

Bij grotere steekproeven geldt $(n-2)/n \approx 1$ waardoor we opnieuw bij de benadering van Cohen terecht komen.

In het bovenstaande voorbeeld is $d = 0.91$:

```
> spooled<-sqrt((9*var(score[trt==1])+4*var(score[trt==0]))/13)
> meand<-mean(score[trt==1])-mean(score[trt==0])
> d<-meand/spooled
> d
[1] 0.9066841
```

Wanneer we op basis van (3) r bepalen, vinden we hetzelfde resultaat als hierboven:

```
> noemer<-sqrt(d^2+(13*15)/(5*10))
> d/noemer
[1] 0.4172434
```

3.2.3 Schatten van effectgroottes voor het vergelijken van 2 afhankelijke groepen

- Een d-statistiek bij een afhankelijk contrast (hierbij horen ook designs met matching!) wordt vaak een gemiddelde verandering genoemd.

Voorbeeld

Meting tijdstip 1	X_{11}	X_{12}	$\ldots$	X_{1n}
Meting tijdstip 2	X_{21}	X_{22}	$\ldots$	X_{2n}
Verschil D	$D_1 = X_{11} - X_{21}$	$D_2 = X_{12} - X_{22}$	$\ldots$	$D_n = X_{1n} - X_{2n}$

Mogelijkheden om het gemiddelde verschil $\bar{X}_1 - \bar{X}_2 = \bar{D}$ te standaardiseren:

- gebruik standaaddeviaties van originele scores: S_1, S_2 (Glass's Δ of Hedges's g).
- gebruik standaarddeviatie S_D van individuele verschillen D

Opmerking

Bij een gepaarde t-toets is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_D/\sqrt{n}} = \frac{\bar{D}}{S_D/\sqrt{n}}$$

Om over te gaan naar Hedges's g is de formule dan:

$$g = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2+S_2^2}{2}}} = t\sqrt{\frac{2S_D^2}{n(S_1^2+S_2^2)}}$$

- r_{pb} is een associatiemaat voor onafhankelijke contrasten; voor afhankelijke contrasten bestaan verschillende andere maten; we gaan hier niet dieper op in.

3.2.4 Associatiematen voor meer complexe designs

- d-statistieken en bivariate correlaties vergelijken slechts 2 gemiddelden/groepen per keer.
- Vaak zijn er meer dan 2 condities.
- In een regressiecontext is het mogelijk om effectgroottes te berekenen voor elk individueel effect (contrasten, hoofdeffecten, interactie-effecten of combinaties van effecten).

  We komen hier op terug in het hoofdstuk over lineaire regressie.

3.2.5 Interpretatie

Effectgroottes zijn nuttig omdat ze als maat gebruikt kunnen worden voor de belangrijkheid van effecten.

Cohen (1988) stelde enkele vuistregels voor om de 'grootte' van een effect kwalitatief te evalueren:

$d = 0.25$	$\rightarrow$	klein effect
$d = 0.50$	$\rightarrow$	medium effect
$d = 0.80$	$\rightarrow$	groot effect

of in termen van r:

$r = 0.10$	$\rightarrow$	klein effect, effect verklaart 1% van de totale variantie
$r = 0.30$	$\rightarrow$	medium effect, effect verklaart 9% van de totale variantie
$r = 0.50$	$\rightarrow$	groot effect, effect verklaart 25% van de totale variantie

Maak echter niet de fout om dit klakkeloos over te nemen. In dat geval wordt opnieuw dezelfde fout gemaakt als bij het gebruik van p-waarden (blindelings richtlijnen volgen). Deze richtlijnen zijn niet gebaseerd op empirische bevindingen!

Grootte $\neq$ belangrijkheid: de belangrijkheid van een effect hangt af van de context. Het vinden van een behandeling met een klein effect op erg dodelijke ziektes zal waarschijnlijk belangrijker bevonden worden dan het vinden van een behandeling met een groot effect op niet-dodelijke ziektes waarvoor al behandelingen beschikbaar zijn.

Tabel 6.3 uit Kline (2009, pag. 174):

Table 6.3. How to Fool Yourself with Effect Size Estimation

1. Ignore effect size at the case level (i.e., estimate it at the group level only).
2. Apply T-shirt size categories to effect sizes without first consulting the empirical literature.
3. Believe that a T-shirt effect size of "large" indicates an important result and that a "small" effect does not.
4. Ignore other considerations in your research area for judging substantive significance, including theory and results of meta-analytic studies.
5. Estimate effect size magnitude only for statistically significant results.
6. Believe that effect size estimation is a substitute for replication.
7. Fail to report confidence intervals for effect sizes, when it is feasible to do so. That is, forget that effect sizes are subject to sampling error, too.
8. Forget that effect size reflects characteristics of your study, such as the design, variability of participants, and how variables are measured.
9. Blindly substitute effect size magnitude for statistical significance as the criterion for judging scientific merit.

Note. These points are based on those presented in Kline (2004, p. 136).

De effectgroottes van één enkele studie dienen vergeleken te worden met de effecten in gerelateerde studies.

3.2.6 Winner's curse

De winner's curse verwijst naar het fenomeen dat wanneer effectgroottes enkel geschat worden voor statistisch significante resultaten, effectgroottes (serieus) overschat worden wanneer studies underpowered zijn.

Dit is te begrijpen als volgt: bij een lage steekproefgrootte n, zal een grotere geboserveerde effectgrootte nodig zijn om de grens van statistische significantie te bereiken dan wanneer n (en dus ook de power) groot is.

Veronderstel dat er een onderliggend effect is. Enkel wanneer studies voldoende power hebben (en dus in de meeste gevallen de nulhypothese zullen verwerpen), zullen geschatte effectgroottes voor statistisch significante resultaten convergeren naar de echte, onderliggende effectgrootte.

Dit toont zowel aan dat underpowered studies problematisch zijn als dat het belangrijk is om effectgroottes te schatten los van statistische significantie.

3.3 Effectgroottes voor binaire uitkomsten

Tot hiertoe hadden we het over het vergelijken van uitkomsten van intervalniveau tussen 2 of meerdere groepen. Studies kunnen ook ontworpen zijn om categorische uitkomsten over verschillende groepen te vergelijken. De niveaus van een categorische variabele kunnen geordend (*ordinaal niveau*) of ongeordend (*nominaal niveau*) zijn.

In dit stuk gaan we dieper in op het vergelijken van een binaire (nominale) uitkomst (Y) tussen 2 onafhankelijke groepen (X).

3.3.1 Designs

We onderscheiden verschillende onderzoeksdesigns: retrospectieve, prospectieve en cross-sectionele designs.

- Retrospectieve designs

  Voorbeeld

  709 patiënten (*cases*) met longkanker worden bevraagd naar hun rookgedrag (X); daarnaast worden ook 709 personen zonder longkanker bevraagd (*controles*). De frequenties per categorie kunnen in een (2×2)-tabel weergeven worden:

	Longkanker		
Roker	Cases	Controls	Totaal
Ja	688	650	1 338
Nee	21	59	80
Totaal	709	709	1418

  - Deelnemers worden geselecteerd op basis van hun status voor de uitkomst Y (voorbeeld: longkanker of geen longkanker); nadien worden de deelnemers volgens groep opgesplitst (X). Er wordt in het 'verleden' gekeken naar hun blootstelling (voorbeeld: roken of niet roken).
  - Dergelijke studies worden ook *case-controle* studies genoemd.
  - Dit design wordt vaak gebruikt bij gebeurtenissen die eerder zeldzaam zijn in een populatie zodat er zeker voldoende cases in de studie opgenomen zijn.
  - Binnen dergelijk design kunnen we $P(X = x|Y = y)$ schatten maar in het algemeen is het niet mogelijk om $P(Y = y|X = x)$ te schatten aangezien er op uitkomst geselecteerd is.
    $P(Y = y|X = x)$ kan wel geschat worden indien $P(Y = y)$ gekend is.

    In bovenstaand voorbeeld betekent dit dat we de condititonele distributie op roken (ja/nee) kunnen vergelijken tussen de groep met longkanker en geen longkanker MAAR we kunnen de condititonele distributie van longkanker (ja/nee) niet vergelijken tussen de rokers en niet-rokers tenzij we de prevalentie van longkanker in de populatie kennen.

    Indien de prevalentie van longkanker (P(longkanker)) gekend is, kunnen we via de regel van Bayes de conditionele kans op longkanker gegeven roken bepalen:

$$P(\text{longkanker}|\text{roken}) = \frac{P(\text{roken}|\text{longkanker})P(\text{longkanker})}{P(\text{roken}|\text{longkanker})P(\text{longkanker}) + P(\text{roken}|\text{geen longkanker})P(\text{geen longkanker})}.$$

    Het design laat immers wel toe om P(roken|longkanker) en P(roken|geen longkanker) te schatten.

    De kans op longkanker bij niet-rokers kan dan op een analoge manier bepaald worden.

- Prospectieve designs

  Voorbeeld

  Een studie selecteert deelnemers uit een populatie van tieners en bepaalt 60 jaar later de prevalentie van longkanker voor de rokers en niet-rokers.

  Er zijn 2 types van prospectieve designs:

  1. Klinische, gerandomiseerde studies: verdeel de deelnemers at random tot de groep rokers of de groep niet-rokers (in de praktijk is dit voor dergelijk voorbeeld uiteraard niet mogelijk).

2. Cohort studies: laat de deelnemers zelf bepalen om al dan niet te roken en observeer in de toekomst welke deelnemers longkanker krijgen.

- Cross-sectioneel design

  In dit design wordt een (grote) steekproef uit de populatie genomen. Van iedere deelnemer wordt bepaald tot welke categorie (cel) hij/zij behoort.

We gebruiken onderstaand voorbeeld om mogelijke associatiematen tussen 2 binaire variabelen X en Y te definiëren.

Voorbeeld

Bij het vergelijken van 2 behandelingen voor depressie kan men kijken naar het verschil tussen het al dan niet hervallen na een bepaalde tijd. Type behandeling (X) vormt de groepen; *hervallen* is de uitkomst (Y) die bestaat uit 2 niveaus (ja/nee).

In het algemeen ziet een frequentietabel (kruistabel) er als volgt uit:

	Hervallen	Niet hervallen	Totaal
Behandeling 1	A	B	$n_1 = A + B$
Behandeling 2	C	D	$n_2 = C + D$
Totaal	$m_1 = A + C$	$m_0 = B + D$	$N = A + B + C + D$

Totale steekproefgrootte: $N = A + B + C + D$; aantal patiënten met behandeling 1 die hervallen zijn: A; totaal aantal patiënten die hervallen zijn: $m_1 = A + C$, etc.

Noteer:

π_i : kans op hervallen (= aanwezigheid van een gebeurtenis) in behandeling i ; $i = 1, 2$

P_i : schatter voor π_i op basis van steekproef

Voor bovenstaande tabel (indien niet geselecteerd wordt op status van al dan niet hervallen):
$p_1 = A/(A + B) = A/n_1$, $p_2 = C/(C + D) = C/n_2$.

VRAAG: is er een verschil tussen behandeling 1 en behandeling 2 m.b.t. hervallen?

3.3.2 Risicoverschil RV

- $\widehat{RV} = P_1 - P_2$ is het risicoverschil in de steekproef; dit is een schatter voor het risicoverschil in de populatie $RV = \pi_1 - \pi_2$
- Makkelijk te interpreteren
- Wanneer de variabelen X en Y onafhankelijk zijn, is $RV = \pi_1 - \pi_2 = 0$.
- Let wel: het bereik hangt af van de waarden van π_1 en π_2 dus een risicoverschil is niet noodzakelijk vergelijkbaar tussen steekproeven van populaties met andere risico's.
- Een verschil is mogelijks belangrijker wanneer beide proporties dicht bij 0 of 1 liggen dan wanneer beide proporties dicht bij 0.5 liggen, bvb. $(p_1 - p_2) = .09$ als

$$(.10 - 0.01) = .09 \quad \text{of} \quad (.50 - .41) = .09$$

  In het eerste geval is p_1 10 keer groter dan p_2 terwijl p_1 in het tweede geval enkel 1.2 keer groter is dan p_2.

- (Benaderende) betrouwbaarheidsintervallen voor $\pi_1 - \pi_2$ maken typisch gebruik van de volgende benadering voor de standaardfout van $P_1 - P_2$ (grote steekproeven):

$$SE_{RV} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

- Indien we de volgende frequentietabel observeren

	Hervallen	Niet hervallen	Totaal
Behandeling 1	28	656	684
Behandeling 2	18	658	676
Totaal	46	1314	1360

schatten we dat het risicoverschil tussen beide behandelingen gelijk is aan $(p_1 - p_2) = (.041 - .027) = .014$.

3.3.3 Relatief riscio RR

- $RR = \pi_1/\pi_2$: verhouding van kans op hervallen binnen 2 behandelingsarmen
- $\widehat{RR} = P_1/P_2$
- In ons voorbeeld vinden we dat $(p_1/p_2) = (.0409/.0266) = 1.54$
- Makkelijk te interpreteren

  Voorbeeld

  $RR = 1.50$: risico op hervallen is 1.5 keer groter in behandelingsarm 1

  $RR = 0.70$: risico op hervallen is 70% van de kans op hervallen in behandelingsarm 2
- Wanneer de variabelen X en Y onafhankelijk zijn, is $RR = \pi_1/\pi_2 = 1$.
- Let wel: bereik hangt af van noemer
- Het interval om een lagere kans op hervallen aan te geven in behandelingsarm 1 loopt van 0 tot 1 terwijl het interval dat een groter risico aangeeft van 1 tot $+\infty$ loopt.
- Het RR wordt vaak geanalyseerd op de log-schaal

  Log-transformatie van x: $\ln(x) = \log_e(x)$ of kortweg $\log(x)$
- $\log(RR)$ is symmetrisch rond 0:
  - $\log(1) = 0$
  - $\log(4) = 1.39$ en $\log(1/4) = -1.39$
- Een benadering voor de standaardfout van $\log(\widehat{RR})$:

$$SE_{\log(RR)} = \sqrt{\frac{1-P_1}{n_1 P_1} + \frac{1-P_2}{n_2 P_2}}$$

3.3.4 Odds ratio OR

- $P_i/(1-P_i)$: Odds om te hervallen in behandelingsgroep i ($i=1,2$); schatter voor $\pi_i/(1-\pi_i)$

  Voorbeeld

  $p_1 = 0.80$ geeft een odds in behandelingsgroep 1 van 0.8/0.2 of 4 tegen 1 om te hervallen.

  Merk op:

  - De odds neemt waarden aan tussen nul en oneindig.
  - De odds is gelijk aan 1 als en slechts als de kans zelf gelijk is aan 1/2.
  - De odds neemt toe als de kans toeneemt.

- De odds ratio (op basis van de steekproef) om te hervallen in behandelingsgroep 1 tegenover behandelingsgroep 2 is

$$\widehat{OR} = \frac{\frac{P_1}{(1-P_1)}}{\frac{P_2}{(1-P_2)}}$$

  Voorbeeld

  Indien OR=1.5 betekent dit dat de odds op hervallen in de eerste behandelingsgroep 1.5 keer groter is dan in behandelingsgroep 2.

  Tabel:

$$\frac{p_1}{1-p_1} = \frac{A}{B} \quad \frac{p_2}{1-p_2} = \frac{C}{D} \Rightarrow \widehat{OR} = \frac{AD}{BC}$$

  In bovenstaand voorbeeld schatten we bijgevolg dat de odds ratio gelijk is aan $\frac{28\times658}{18\times656} = 1.56$.

- Indien X en Y onafhankelijk zijn, is $OR = 1$.

- De odds ratio is moeilijker te interpreteren (minder intuïtief) maar bezit goede statistische eigenschappen.

- De odds ratio kan op een geldige manier geschat worden in retrospectieve, prospectieve en cross-sectionele designs. Zoals we eerder zagen, is het bij retrospectieve designs niet mogelijk om het risicoverschil en het relatief risico te schatten aangezien in dergelijke studies de risico's op hervallen niet geschat kunnen worden. Het is daarentegen wel zo dat de odds ratio op hervallen voor behandeling 1 tegenover behandeling 2 hetzelfde is als de odds ratio op behandeling (1 versus 2) voor hervallen tegenover niet hervallen. Het is deze eigenschap die het mogelijk maakt om de odds ratio te schatten in elk type van design.

- Wanneer de kans op hervallen onder beide behandelingen klein is ($< 5\%$), dan is de odds een goede benadering voor het risico. Dit is omdat A/B dan dicht bij $A/(A+B)$ ligt en C/D dicht bij $C/(C+D)$. In dat geval is ook de odds ratio een goede benadering voor het relatief risico. Dit is een bijzonder nuttige observatie omdat de odds ratio bepaalde wiskundige eigenschappen heeft die ze aantrekkelijker maakt dan een relatief risico in statistische modellen (denk bvb. aan logistische regressie, dit komt verder in deze cursus aan bod); en het relatief risico niet in alle designs geschat kan worden.

- Net als bij het relatief risico loopt het interval voor een kleinere odds in behandelingsgroep 1 van 0 tot 1 terwijl het interval voor een grotere odds van 1 tot $+\infty$ loopt.

- De odds ratio wordt ook vaak geanalyseerd op de log-schaal.

- De *log odds ratio* is symmetrisch rond 0.

- De benaderende standaardfout van $\widehat{OR}$ op de log-schaal is:

$$SE_{\log(OR)} = \sqrt{\frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}}$$

- De odds ratio kan omgezet worden in een gestandaardiseerd gemiddeld verschil logit(d).
  Hier: logit=log-transformatie van OR.
  $\log(\widehat{OR})$ is bij benadering normaal verdeeld met standaarddeviatie $pi/\sqrt{3}$ (pi: het getal pi, 3.1415...).

$$\text{logit}(d) = \frac{\log(\widehat{OR})}{pi/\sqrt{3}}$$

  logit(d) is vergelijkbaar met een gestandaardiseerd gemiddeld verschil van een uitkomst van intervalniveau tussen 2 groepen.

  Voorbeeld

$$\begin{aligned} \widehat{OR} &= 2.25 \\ \text{logit}(d) &= \frac{\log(2.25)}{pi/\sqrt{3}} = 0.45 \end{aligned}$$

  Dit betekent dat een odds ratio van 2.25 vergelijkbaar is met een behandelingseffect van ongeveer een halve standaardafwijking op een continue uitkomst.

3.4 Betrouwbaarheidsintervallen

De APA publication manual argumenteert dat betrouwbaarheidsintervallen in het algemeen de beste strategie zijn om te rapporteren. Het gebruik van betrouwbaarheidsintervallen wordt dan ook sterk aangeraden.

Een betrouwbaarheidsinterval voor een parameter heeft typisch de volgende vorm:

$$\text{Schatter} \pm \text{ (kritische waarde) } \times \text{(Standaardfout van schatter)}$$

Een $[(1-\alpha)100]\%$ betrouwbaarheidsinterval omvat met kans $1-\alpha$ de ware parameter. Denk aan de frequentistische betekenis: indien we alle mogelijke lukrake steekproeven van grootte n zouden trekken en telkens een $[(1-\alpha)100]\%$ betrouwbaarheidsinterval voor de parameter zouden berekenen, omvat $(1-\alpha) \times 100\%$ van deze intervallen de ware parameter en $\alpha \times 100\%$ niet.

Een betrouwbaarheidsinterval geeft naast een puntschatting ook een idee van de nauwkeurigheid van een schatter.

Enkele misvattingen omtrent betrouwbaarheidsintervallen moeten vermeden worden (Thompson, 2006):

- Sommige onderzoekers interpreteren betrouwbaarheidsintervallen met een hoge betrouwbaarheid (kleine α) alsof de intervallen een resultaat impliceren dat 100% zeker is. Hoewel 95% dicht bij 100% ligt, is 95% simpelweg niet gelijk aan 100%.

  In werkelijkheid weten we niet of het bekomen interval de onbekende populatieparameter omvat.

- Vaak gaat men uit van de veronderstelling dat betrouwbaarheidsintervallen NHST's zijn in een andere vorm.

  Er bestaat inderdaad een relatie tussen het opstellen van een betrouwbaarheidsinterval rond $\hat{\theta}$, en het toetsen van een hypothese $H_0 : \theta = \theta_0$ die we als volgt kunnen formuleren:

Indien θ_0 zich niet in het betrouwbaarheidsinterval rond $\hat{\theta}$ bevindt dan verwerpen we H_0 op het $\alpha \times 100\%$ significantieniveau. Omgekeerd, indien θ_0 zich wel in het betrouwbaarheidsinterval rond $\hat{\theta}$ bevindt dan kunnen we H_0 niet verwerpen.

(Deze relatie geldt in principe enkel voor tweezijdig toetsen, i.e. $H_1 : \theta \neq \theta_0$. Het is echter mogelijk dezelfde relatie te laten gelden voor eenzijdige toetsen door het opstellen van *eenzijdige betrouwbaarheidsintervallen*.)

Conclusie: een $[(1-\alpha)100]\%$ betrouwbaarheidsinterval omvat alle waarden voor de parameter θ waarvoor H_0 niet verworpen kan worden op het $\alpha \times 100\%$ significantieniveau.

Dit betekent echter niet dat betrouwbaarheidsintervallen enkel zinvol zijn in een context waarin we toetsen, hun interpretatie reikt verder dan dat. De nadruk bij de interpretatie moet liggen op de range van plausibele waarden voor de parameter.

Om het duidelijk te stellen: een betrouwbaarheidsinterval kan geconstrueerd worden zonder het specifiëren van een nulhypothese terwijl dit voor NHST onmogelijk is.

Betrouwbaarheidsintervallen laten toe om op een zinvolle manier resultaten van verschillende studies te vergelijken. Ze kunnen ook grafisch voorgesteld worden (zie volgende sectie) en geven meer informatie dan een p-waarde.

3.5 Betrouwbaarheidsintervallen voor effectgroottes

Zowel puntschattingen voor effectgroottes als bijhorende betrouwbaarheidsintervallen zijn belangrijk om te rapporteren. Het berekenen van een exact betrouwbaarheidsinterval voor een **gestandaardiseerde** effectgrootte is technisch echter een stuk complexer dan het geval van niet-gestandaardiseerde effectgroottes.

De reden hiervoor is dat men niet langer gebruik kan maken van centrale verdelingsfuncties zoals de gekende t- of F- of χ^2- verdeling. Deze worden vervangen door hun niet-centrale tegenhangers, de kritische waarden bepalen is dan een iteratief proces. Er zijn dus geen formules beschikbaar. Een technische uiteenzetting laten we hier achterwege. Onthoud dat statistische software zoals R het toelaat om betrouwbaarheidsintervallen voor gestandaardiseerde effectgroottes te bepalen (zie bvb. R package `MBESS`).

Een andere optie is om benaderende betrouwbaarheidsintervallen voor effectgroottes op te stellen (hiervoor is dan geen speciale software noodzakelijk). De basisformule is dan als volgt:

$$\text{Geschatte effectgrootte} \pm \text{ (kritische waarde) } \times \text{(Benaderende standaardfout)}$$

met de kritische waarden afkomstig uit centrale verdelingen (zoals voorheen). De benaderende standaardfout wordt ook vaak de *asymptotische* standaardfout genoemd. Dit is de waarde die men verwacht te zien bij grote steekproeven.

De benaderende methode is geschikt voor g, Δ, RV, RR (log-schaal) en de OR (log-schaal) maar niet voor R^2 (en aanverwante associatiematen). Voor deze laatste blijft een niet-centrale intervalschatting noodzakelijk.

Voorbeeld

We illustreren hier het berekenen van een benaderend 95% betrouwbaarheidsinterval voor de OR op hervallen voor behandeling 1 tegenover behandeling 2. Onze tabel was als volgt:

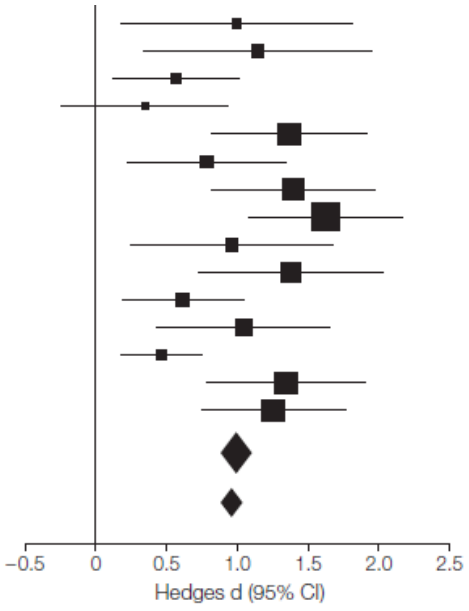
	Hervallen	Niet hervallen	Totaal
Behandeling 1	28	656	684
Behandeling 2	18	658	676
Totaal	46	1314	1360

met $n_1 = 684$, $n_2 = 676$; $p_1 = 28/684$ en $p_2 = 18/676$.

1. $\widehat{OR} = 1.56 \rightarrow \log(\widehat{OR}) = \log(1.56) = 0.44$.
2. $$SE_{\log(OR)} = \sqrt{\frac{1}{684 \times (28/684) \times (656/684)} + \frac{1}{676 \times (18/676) \times (658/676)}} = 0.31$$
3. Benaderend 95% betrouwbaarheidsinterval voor OR op de log-schaal (gebaseerd op normale benadering):
$$[0.44 - 1.96 \times 0.31, 0.44 + 1.96 \times 0.31] = [-0.17, 1.05]$$
4. Benaderend 95% betrouwbaarheidsinterval op originele schaal:
$$[\exp(-0.17), \exp(1.05)] = [0.84, 2.86]$$

Is de relatie tussen behandeling en hervallen statistisch significant op het 5% significantieniveau?

Onderstaande figuur is een figuur uit het artikel van Leichsenring en Rabung (2008) waar men een meta-analyse uitvoerde om de effecten van long-term psychodynamische psychotherapie te onderzoeken. De effectgroottes die beschouwd werden zijn Hedges's g (aangeduid met d). Er werden zowel observationele studies als randomised controlled trials bestudeerd (RCT's). De figuur toont de effectgroottes voor de observationele studies (positieve effecten betekenen dat psychotherapie voor verbetering zorgde), tesamen met een 95% betrouwbaarheidsinterval.



De grootte van de vierkanten reflecteert de steekproefgrootte. De ruiten stellen een gewogen gemiddelde voor van de effectgroottes over de studies heen (boven: enkel observationele studies, onder: ook met RCT's).

Dergelijke figuren vatten resultaten goed samen en geven ook duidelijk de precisie in de studies weer. Op dergelijke figuren is het ook mogelijk dat resultaten over studies consistenter zijn dan traditionele analyses aanvankelijk aangaven.

4 Besluit

p-waarden meten enkel de bewijskracht in de data tegen de nulhypothese.

Statistische significantie impliceert niet noodzakelijk praktische significantie (en vice versa). Bij NHST heeft de onderzoeker nergens moeten aangeven welke effectgrootte voor hem/haar belangrijk is en dus kunnen we niet verwachten dat de resultaten van dergelijke toetsen deze implicatie hebben. Thompson (2006) raadt daarom ook aan om te schrijven dat een resultaat 'statistisch signficant' is i.p.v. enkel 'significant'.

Een p-waarde is ook geen maat voor de repliceerbaarheid van een effect. Olejnik en Algena (2000) vatten het als volgt samen: statistische significantie gebaseerd op een statistische toets geeft informatie over de waarschijnlijkheid dat een geobserveerd effect enkel aan toeval te wijten is (omwille van steekproeffout). We weten echter niet wat de kans is om opnieuw een dergelijk resultaat te bekomen.

Effectgroottes helpen om de bewijskracht in de data te evalueren. Effectgrootte, steekproefgrootte, significantieniveau α en power zijn gerelateerd bij het toetsen van hypothesen: elk van deze waarden kan bekomen worden, als de andere gekend zijn. Bij het evalueren van de resultaten van meerdere studies (meta-analyses) worden effectgroottes gebruikt die, indien nodig, tot eenzelfde schaal herleid worden.

Denk er aan: de APA geeft de voorkeur aan niet-gestandaardiseerde effecten indien deze metingen betekenisvol zijn op een praktisch niveau. Ga dus niet automatisch d of r berekenen indien effectgroottes bepaald moeten worden.

Rapporteer niet enkel de p-waarde maar ook de effectgrootte en een betrouwbaarheidsinterval.

Het rapporteren van betrouwbaarheidsintervallen wordt sterk aangeraden. Deze laten de lezer toe om zelf te oordelen over de bewijskracht in de data. Het is makkelijker om een betrouwbaarheidsinterval te interpreteren zonder p-waarde dan een p-waarde zonder betrouwbaarheidsinterval.

5 Referenties

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). Equating r-based and d-based effect-size indices: Problems with a commonly recommended formula. *Paper presented at the annual meeting of the Florida Educational Research Association.* Orlando, FL.

Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science, 18*, 1-32.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (second ed.)*. New Jersey: Lawrence Erlbaum Associates.

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L.V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Orlando: Academic Press.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *Plos Medicine, 2*, 696-701, Article Number: e124. doi: 10.1371/journal.pmed.0020124

Iverson, G.J., Lee, M.D., & Wagenmakers E.J. (2009). p_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review, 16*, 425-429.

Iverson, G. J., Wagenmakers, E.J., & Lee, M. D. (2010). A modelaveraging approach to replication: The case of p_{rep}. *Psychological Methods, 15,* 172-181.

Killeen, P.R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345-353.

Kline, R.B. (2009). *Becoming a Behavioral Science Researcher: A guide to producing research that matters.* New York: Guildford Press.

Leichsenring, F., & Rabung, S. (2008). Effectiveness of Long-term Psychodynamic Psychotherapy. *Journal of the American Medical Association, 300*, 1551-1565.

Matthews, W.J. (2011). What might judgment and decision making research be like if we took a Bayesian approach to hypothesis testing? *Judgment and Decision Making, 6*, 843-856.

Nuzzo, R. (2014). Statistical errors. *Nature, 506*, 150-152.

Olejnik, S., & Algina, J. (2000). Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations. *Contemporary Educational Psychology, 25*, 241-286.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349.* doi: 10.1126/science.aac4716

Thompson, B. (2006). *Foundations of Behavioral Statistics: an insight-based approach.* New York: The Guilford Press.

Wilkinson, L. (1999). Statistical methods in psychology journals - Guidelines and explanations. *American Psychologist, 54*, 594-604.

Inleiding: matrixalgebra

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

In de cursus zullen af en toe matrices aan bod komen. De reden hiervoor is dat de berekeningen soms te complex worden en weinig inzicht bieden, bvb. wanneer we werken met meerdere variabelen. In dat geval kunnen bepaalde zaken op een meer overzichtelijke en compacte manier voorgesteld worden met matrices.

1 Wat is een matrix?

Een matrix is een rechthoekig getallenschema, dus getallen geordend in rijen en kolommen. Enkele voorbeelden:

$$\begin{bmatrix} 1 & 0 \\ 5 & 10 \end{bmatrix} \qquad \begin{bmatrix} 4 & 7 & 12 & 16 \\ 3 & 15 & 9 & 8 \end{bmatrix}$$

De eerste matrix is 2×2-dimensioneel en de tweede 2×4-dimensioneel. Het eerste cijfer verwijst naar het aantal rijen, het tweede naar het aantal kolommen.

De elementen van een matrix kunnen als volgt symbolisch worden aangeduid:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

We gebruiken de algemene notatie a_{ij} voor het element in de i-de rij en de j-de kolom.

We zullen matrices aanduiden met een grote, vette letter. Naar de hierboven gedefinieerde matrix kan vanaf nu gerefereerd worden met $\mathbf{A}$.

In het algemeen kan een matrix met ℓ rijen en k kolommen als volgt voorgesteld worden:

$$\mathbf{A} = [a_{ij}] \quad \text{met } i = 1, ..., \ell;\ j = 1, ..., k$$

2 De getransponeerde van een matrix

De getransponeerde van een matrix $\mathbf{A}$, genoteerd als $\mathbf{A}'$, wordt verkregen door de overeenkomstige rijen en kolommen van $\mathbf{A}$ om te wisselen. Bvb.:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 5 \\ 1 & 3 & 4 \end{bmatrix}$$

$$\mathbf{A}' = \begin{bmatrix} 2 & 1 \\ -1 & 3 \\ 5 & 4 \end{bmatrix}$$

De eerste rij van $\mathbf{A}$ wordt de eerste kolom van $\mathbf{A}'$ en de tweede rij van $\mathbf{A}$ wordt de tweede kolom van $\mathbf{A}'$. Of: de eerste kolom van $\mathbf{A}$ wordt de eerste rij van $\mathbf{A}'$ enzovoort.

Als $\mathbf{A}$ $\ell \times k$-dimensioneel is, is $\mathbf{A}'$ $k \times \ell$-dimensioneel.

Twee keer transponeren geeft weer de originele matrix: $(\mathbf{A}')' = \mathbf{A}$.

Algemeen hebben we:

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{\ell 1} & \dots & a_{\ell k} \end{bmatrix} = [a_{ij}] \quad \text{met } i = 1, \dots, \ell;\ j = 1, \dots, k$$

$$\mathbf{A}' = \begin{bmatrix} a_{11} & \dots & a_{\ell 1} \\ \vdots & & \vdots \\ a_{1k} & \dots & a_{k\ell} \end{bmatrix} = [a_{ji}] \quad \text{met } j = 1, \dots, k;\ i = 1, \dots, \ell$$

3 Speciale matrices

- **Vierkante matrices**
  Een vierkante matrix bevat evenveel rijen als kolommen: $\ell = k$. Bvb.:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad \begin{bmatrix} 4 & 7 \\ 3 & 9 \end{bmatrix}$$

- **Vectoren**
  Een kolomvector is een matrix met slechts 1 kolom ($k = 1$). Bvb.:

$$\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} \qquad \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

  Een rijvector is een matrix met slechts 1 rij ($\ell = 1$). Bvb.:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix} \qquad \begin{bmatrix} 4 & 7 \end{bmatrix}$$

  Als $\mathbf{A}$ een kolomvector is, is zijn getransponeerde $\mathbf{A}'$ een rijvector. Wanneer we spreken van een vector, bedoelen we een kolomvector. Een rijvector zal steeds aangeduid worden met het symbool $'$.

- **Scalairen**
  Dit is een matrix met slechts 1 rij en 1 kolom ($\ell = k = 1$) of kortweg een getal.
- **Symmetrische matrices**
  Een matrix is symmetrisch indien $\mathbf{A} = \mathbf{A}'$. Bvb.:

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{bmatrix} \qquad \mathbf{A}' = \begin{bmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{bmatrix}$$

- **Diagonaalmatrix**
  Bij een diagonaalmatrix zijn alle off-diagonaal elementen 0. Bvb.:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

- **Identiteitsmatrix**
  Bij een *identiteitsmatrix* zijn alle off-diagonaal elementen 0 en alle diagonaal elementen 1. Bvb., de 3×3 identiteitsmatrix:

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Matrix waarvan alle elementen gelijk zijn aan 1.

  Notatie: $\mathbf{1}$.
- Matrix waarvan alle elementen gelijk zijn aan 0 (nulmatrix).

  Notatie: $\mathbf{0}$

4 Gelijkheid van matrices

Twee matrices $\mathbf{A}$ en $\mathbf{B}$ zijn gelijk als ze van dezelfde dimensionaliteit zijn en indien alle overeenkomstige elementen identiek zijn. Bijvoordbeeld:

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 4 \\ 7 \\ 3 \end{bmatrix}$$

dan impliceert $\mathbf{A} = \mathbf{B}$ dat:

$$a_1 = 4 \qquad a_2 = 7 \qquad a_3 = 3$$

Algemeen:

$$\mathbf{A} = \mathbf{B} \Leftrightarrow a_{ij} = b_{ij} \text{ voor } i = 1, \ldots, \ell; j = 1, \ldots, k$$

5 Bewerkingen met matrices

5.1 Optellen van matrices

- Het optellen van matrices vereist dat $\mathbf{A}$ en $\mathbf{B}$ *conform* zijn (van dezelfde dimensie). De som van twee matrices is een matrix met dezelfde dimensie waarbij elk element de som is van de twee overeenkomstige elementen van de twee matrices.
- Bvb., als

$$\mathbf{A} = \begin{bmatrix} 1 & -3 \\ 5 & 2 \\ -2 & 4 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 6 & 3 \\ -2 & 1 \\ 7 & -5 \end{bmatrix}$$

  Dan is

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1+6 & -3+3 \\ 5-2 & 2+1 \\ -2+7 & 4-5 \end{bmatrix} = \begin{bmatrix} 7 & 0 \\ 3 & 3 \\ 5 & -1 \end{bmatrix}$$

- Algemeen, als

$$\mathbf{A} = [a_{ij}] \quad \mathbf{B} = [b_{ij}] \quad \text{met } i = 1, \ldots, \ell;\ j = 1, \ldots, k$$

  dan

$$\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$$

- Merk op dat $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$, zoals in gewone algebra.

5.2 Scalaire vermenigvuldiging

Een scalair is een gewoon getal, of een symbool dat een getal voorstelt. Bij het vermenigvuldigen van een matrix met een scalair wordt elk element van de matrix vermenigvuldigd met deze scalair.

- Bvb., neem de matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 2 & 7 \\ 9 & 3 \end{bmatrix}$$

- dan is $4\mathbf{A}$, waar 4 de scalair is gelijk aan:

$$4\mathbf{A} = 4 \begin{bmatrix} 2 & 7 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} 8 & 28 \\ 36 & 12 \end{bmatrix}$$

- op deze manier is $k\mathbf{A}$ gelijk aan:

$$k\mathbf{A} = k \begin{bmatrix} 2 & 7 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} 2k & 7k \\ 9k & 3k \end{bmatrix}$$

- Algemeen: $k\mathbf{A} = [ka_{ij}] = \mathbf{A}k$
- Merk op: $(k\mathbf{A})' = k\mathbf{A}'$

5.3 Matrixvermenigvuldiging

Vermenigvuldigen van 2 matrices is iets complexer. Neem de twee matrices:

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 4 & 1 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 4 & 6 \\ 5 & 8 \end{bmatrix}$$

Dan zal het product $\mathbf{AB}$ een 2×2 matrix zijn die bekomen wordt door eerst de kruisproducten van de rijen van $\mathbf{A}$ met de kolommen van $\mathbf{B}$ te nemen en deze kruisproducten te sommeren. Om het element te vinden in de eerste rij en de eerste kolom van het product $\mathbf{AB}$ werken we met de eerste rij van $\mathbf{A}$ en de eerste kolom van $\mathbf{B}$:

$$(2 \times 4) + (5 \times 5) = 8 + 25 = 33$$

Vervolgens nemen we de eerste rij van **A** en de tweede kolom van **B** om het element te vinden op de eerste rij en de tweede kolom van het product **AB**.

$$(2 \times 6) + (5 \times 8) = 12 + 40 = 52$$

Zo werken we door tot we de kruisproducten van elke rij van **A** met elke kolom van **B** hebben. Zo bekomen we het product **AB**:

$$\mathbf{AB} = \begin{bmatrix} 33 & 52 \\ 21 & 32 \end{bmatrix}$$

Een ander voorbeeld:

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 5 & 8 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 5 & 8 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 26 \\ 41 \end{bmatrix}$$

Wanneer we **AB** berekenen, zeggen we dat **A** navermenigvuldigd is met **B** of dat **B** voorvermenigvuldigd is met **A**. De reden voor deze precisering is dat de vermenigvuldigingsregels uit de gewone algebra niet van toepassing zijn in de matrixalgebra. In gewone algebra is $ab = ba$, in matrix algebra geldt over het algemeen dat $\mathbf{AB} \neq \mathbf{BA}$. Indien **AB** gedefinieerd is, is het product **BA** zelfs niet noodzakelijk gedefinieerd.

Het product **AB** is enkel gedefinieerd indien het aantal kolommen van **A** gelijk is aan het aantal rijen van **B**, zodat het berekenen van de kruisproducten mogelijk is.

De dimensionaliteit van het product **AB** is gelijk aan het aantal rijen in **A** en het aantal kolommen van **B**.

Algemeen, als $\mathbf{A}$ $n \times m$-dimensioneel is en $\mathbf{B}$ $m \times p$-dimensioneel is, dan zal het product $\mathbf{AB}$ een $n \times p$ dimensionele matrix zijn, waarvan het element op de i-de rij en j-de kolom gedefinieerd is als:

$$\sum_{k=1}^{m} a_{ik} b_{kj}$$

zodat

$$\mathbf{AB} = \left[\sum_{k=1}^{m} a_{ik} b_{kj} \right] \quad \text{met } i = 1, ..., n;\ j = 1, ..., p.$$

Noot: indien $\mathbf{AB} = \mathbf{AC}$, met $(\mathbf{A} \neq \mathbf{0})$ dan geldt niet automatisch $\mathbf{B} = \mathbf{C}$! Bvb.

$$\underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} 2 \\ 1 \end{bmatrix}}_{\mathbf{B}} = \underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} 1 \\ 1.5 \end{bmatrix}}_{\mathbf{C}} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$$

Identiteitsmatrices zijn het neutraal element in matrix vermenigvuldiging:

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

Een vierkante matrix is *orthogonaal* als $\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \mathbf{I}$

5.4 Lineaire afhankelijkheid en rang van een matrix

Neem als voorbeeld:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 2 & 10 & 6 \\ 3 & 4 & 15 & 1 \\ 3 & 1 & 15 & 2 \end{bmatrix}$$

In dit voorbeeld is de derde kolom een veelvoud van de eerste kolom:

$$\begin{bmatrix} 5 \\ 10 \\ 15 \\ 15 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

We zeggen dat de kolommen van $\mathbf{A}$ lineair afhankelijk zijn. Ze bevatten overbodige informatie, aangezien een van de kolommen kan bekomen worden door een lineaire combinatie te nemen van de andere kolommen.

De set van k kolomvectoren $\mathbf{k^*}_1, \ldots, \mathbf{k^*}_k$ in een $\ell \times k$ matrix zijn lineair afhankelijk als een vector kan uitgedrukt worden als een lineaire combinatie van de andere vectoren. Indien geen enkele van de vectoren zo kan uitgedrukt worden, is de set vectoren lineair onafhankelijk.

Een meer formele definitie:

- wanneer k scalairen $\lambda_1, \ldots, \lambda_k$, die niet alle 0 zijn, kunnen gevonden worden zodat:

$$\lambda_1 \mathbf{k^*}_1 + \lambda_2 \mathbf{k^*}_2 + \ldots + \lambda_c \mathbf{k^*}_k = \mathbf{0}$$

  dan zijn de k kolomvectoren *lineair afhankelijk*.

- Indien de enige set waarvoor de gelijkheid opgaat, gelijk is aan

$$\lambda_1 = 0, \ldots, \lambda_c = 0$$

  dan is de set van kolomvectoren *lineair onafhankelijk*.

In ons voorbeeld leidt $\lambda_1 = 5,\ \lambda_2 = 0,\ \lambda_3 = -1,\ \lambda_4 = 0$ tot:

$$5 \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix} + 0 \begin{bmatrix} 2 \\ 2 \\ 4 \\ 1 \end{bmatrix} - 1 \begin{bmatrix} 5 \\ 10 \\ 15 \\ 15 \end{bmatrix} + 0 \begin{bmatrix} 1 \\ 6 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Dus zijn de kolommen lineair afhankelijk. Sommige λ_j zijn hier gelijk aan 0, maar dit geeft niet: voor lineaire afhankelijkheid is het enkel nodig dat niet *alle* λ_j gelijk zijn aan 0.

Rang van een matrix

De rang van een $(\ell \times k)$-matrix is gelijk aan het aantal lineair onafhankelijke rijen/kolommen van de matrix.

Bijgevolg is de rang nooit groter dan $\min(\ell, k)$, het minimum van de twee waarden ℓ en k.

Wanneer de rang van een $(\ell \times k)$-matrix gelijk is aan $\min(\ell, k)$, dan is de matrix van volledige rang.

In het vorig voorbeeld is de rang van de matrix gelijk aan 3.

Wanneer een matrix het product is van twee matrices, kan de rang nooit groter zijn dan het minumum van de rang van de twee vermenigvuldigde matrices. Dus: als $\mathbf{C} = \mathbf{AB}$, dan is de rang van $\mathbf{C}$ niet groter dan min(rang $\mathbf{A}$, rang $\mathbf{B}$).

5.5 Inverse van een matrix

In gewone algebra is de inverse van een getal gelijk aan zijn omgekeerde. De inverse van 6 is dus 1/6. Wanneer we een getal vermenigvuldigen met zijn inverse krijgen we 1, want $6 \times (1/6) = 1$.

In matrixalgebra is de inverse van een matrix $\mathbf{A}$ een andere matrix, genoteerd als $\mathbf{A}^{-1}$ zodat:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

De identiteitsmatrix $\mathbf{I}$ speelt dus dezelfde rol als het getal 1 in gewone algebra.

De inverse van een matrix is enkel gedefinieerd voor *vierkante* matrices die van *volledige rang* (*niet-singulier*) zijn. In dat geval is de inverse uniek.

Voorbeeld: de inverse van de matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 7 \\ 3 & 2 & 5 \\ 5 & 2 & 8 \end{bmatrix}$$

is

$$\mathbf{A}^{-1} = \begin{bmatrix} -1/3 & 1 & -1/3 \\ -1/18 & 1.5 & -8/9 \\ 2/9 & -1 & 5/9 \end{bmatrix}$$

aangezien

$$\mathbf{A}^{-1}\mathbf{A} = \begin{bmatrix} -1/3 & 1 & -1/3 \\ -1/18 & 1.5 & -8/9 \\ 2/9 & -1 & 5/9 \end{bmatrix} \begin{bmatrix} 1 & 4 & 7 \\ 3 & 2 & 5 \\ 5 & 2 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Aangezien het berekenen van de inverse van de rang van een matrix vrij rekenintensief is en meestal door de computer wordt gedaan gaan we hier niet op in.

We kunnen 1 speciaal geval opmerken. De inverse van een diagonaalmatrix is opnieuw een diagonaalmatrix die bestaat uit de inversen van de elementen van de diagonaal:

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$$

6 Variantie-covariantiematrix

Laat $\mathbf{Z}$ een vector voorstellen van n kansvariabelen:

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}.$$

De verwachting van $\mathbf{Z}$ is dan:

$$E(\mathbf{Z}) = \begin{bmatrix} E(Z_1) \\ E(Z_2) \\ \vdots \\ E(Z_n) \end{bmatrix}.$$

De varianties van de n kansvariabelen en de covarianties tussen elke paar van de kansvariabelen worden voorgesteld in een *variantie-covariantiematrix*:

$$\mathrm{Var}(\mathbf{Z}) = \begin{bmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \dots & \mathrm{Cov}(Z_1, Z_n) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Var}(Z_2) & \dots & \mathrm{Cov}(Z_2, Z_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(Z_n, Z_1) & \mathrm{Cov}(Z_n, Z_2) & \dots & \mathrm{Var}(Z_n) \end{bmatrix}.$$

De variantie-covariantiematrix kan meer algemeen als volgt voorgesteld worden:

$$[\mathrm{Cov}(Z_i, Z_j)] \quad \text{met } i = 1, \ldots, n; j = 1, \ldots, n$$

waarbij $\mathrm{Cov}(Z_i, Z_i) = \mathrm{Var}(Z_i)$.

De variantie-covariantiematrix is $n \times n$-dimensioneel (vierkante matrix).

Merk verder op dat $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cov}(Z_j, Z_i)$, bijgevolg is een variantie-covariantiematrix een symmetrische matrix.

Het Algemeen Lineair Model

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

In dit stuk worden enkele basisbegrippen omtrent het lineair model samengevat (*zie ook Statistiek II*).

1 Lineaire modellen

Een lineair statistisch model of kortweg een **lineair model** is een statistisch model dat lineair is in zijn parameters (β). Bijvoorbeeld:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3^3 + \varepsilon$$

De parameters β komen niet voor in een niet-lineaire vorm
(bvb. β^4, $\log(\beta)$, ...).

2 Afhankelijke en onafhankelijke variabelen

2.1 Afhankelijke variabelen

De variabele(n) Y die men wenst te begrijpen, verklaren, voorspellen noemt men de **afhankelijke** variabele(n), of de uitkomst(en) of respons variabele(n).

Is er slechts één afhankelijke variabele, dan spreekt men van **univariate** modellen. Zijn er meerdere afhankelijke variabelen die men terzelfdertijd beschouwt, dan spreekt men van **multivariate** modellen.

In deze cursus beschouwen we zowel univariate als multivariate modellen.

2.2 Onafhankelijke variabelen

De variabelen X die we in ons model opnemen omdat we als onderzoeker vermoeden dat ze een 'effect' (invloed) kunnen hebben op de afhankelijke variabele, noemt men de **onafhankelijke** variabelen of **predictoren**.

Het is aan de onderzoeker om op grond van voorkennis of theorieën, vakliteratuur, ervaring, ... te bepalen welke variabelen in het model worden opgenomen.

3 Univariate lineaire modellen: een overzicht

3.1 Structuur component

Het structuur model van lineaire modellen met 1 afhankelijke variabele kan geschreven worden in de volgende vorm voor de $i$de observatie:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

met $i = 1, 2, 3, \ldots, n$. De X's corresponderen met hetzij de originele, hetzij gehercodeerde waarden van de onafhankelijke variabelen en worden als constanten beschouwd.

Er zijn twee soorten onbekende parameters:

1. de regressiegewichten of regressiecoëfficiënten: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$
2. de fouttermen ε_i

De afhankelijke variabele Y_i en de fouttermen ε_i zijn **kansvariabelen** met een bepaalde verwachting en variantie.

3.2 Stochastische component

De kernassumptie van het stochastisch model is dat de verwachting van de fouttermen gelijk is aan nul:

$$\mathrm{E}(\varepsilon_i) = 0 \quad \text{voor } i = 1, 2, 3, \ldots, n$$

Deze assumptie impliceert dat:

$$\mathrm{E}(Y_i | x_{i1}, x_{i2}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

Dit vat de essentie samen van lineaire modellen: men probeert de *verwachte waarde* van Y te schrijven als een lineaire combinatie van p variabelen.

Enkele notationele afspraken:

- $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \sigma_{ij}^2$

3.3 Gauss-Markov model

De bovenstaande modellen bevatten $p+1$ parameters voor de regressiecoëfficiënten, en $\frac{n(n+1)}{2}$ parameters voor de varianties σ_i^2 en de covarianties σ_{ij}^2, terwijl er slechts n observaties zijn. Het is onmogelijk om op basis van de data de waarden voor alle onbekende parameters te schatten.

Een Gauss-Markov model is een lineair model waarbij het aantal onbekende parameters gereduceerd wordt via bijkomende assumpties. Een eerste assumptie veronderstelt dat de variantie van de fouttermen voor elke observatie gelijk is:

$$\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2 \quad \text{voor alle } i.$$

De tweede assumptie stelt dat de fouttermen onderling niet gecorreleerd zijn:

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$$

voor alle mogelijke koppels (i, j). De overblijvende onbekende parameters zijn nu:

1. de $p+1$ regressiegewichten of regressiecoëfficiënten: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$
2. één gemeenschappelijke variantie voor de fouttermen: σ_ε^2

In totaal zijn er nu $p+2$ onbekende parameters.

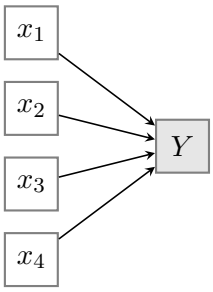
3.4 Bijzondere gevallen

3.4.1 Lineaire regressie

Hoewel alle lineaire modellen als 'regressie' modellen kunnen worden beschouwd, wordt de term **lineaire regressie** traditioneel gebruikt om te verwijzen naar de analyse van lineaire modellen waarbij alle onafhankelijke variabelen van *intervalniveau* zijn.

In het bijzonder geval van slechts één onafhankelijke variabele spreekt men van **enkelvoudige regressie** (Engels: 'simple regression').

In het geval van meerdere onafhankelijke variabelen, spreken we van **meervoudige regressie**.



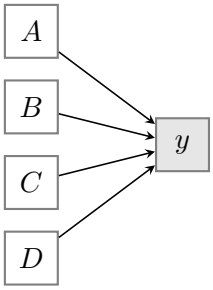
- x_1, x_2, x_3 en x_4 zijn numerieke predictoren van minstens intervalniveau

3.4.2 Variantie-analyse

Traditioneel wordt de term **variantie-analyse** gebruikt om te verwijzen naar de analyse van lineaire modellen waarbij alle onafhankelijke variabelen van *nominaal niveau* zijn. In het Engels spreekt men van 'analysis of variance' of kortweg ANOVA. De onafhankelijke variabelen worden vaak **factoren** genoemd.

In het bijzonder geval van slechts één onafhankelijke variabele spreekt men van **enkelvoudige variantie-analyse** of **eenwegsvariantie-analyse** (Engels: 'oneway anova'). Dit is equivalent met een t-toets voor onafhankelijke groepen, maar het aantal groepen is groter dan 2.

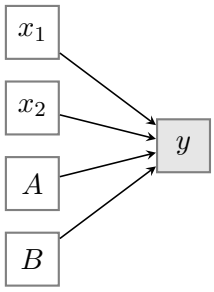
Indien men in de analyse rekening houdt met meerdere onafhankelijke variabelen spreekt men van **meervoudige variantie-analyse** of **meerwegsvariantie-analyse** (Engels: 'multiway anova'), of kortweg variantie-analyse.



- A, B, C en D zijn *factoren*: categorische predictoren van nominaal niveau

3.4.3 Covariantie-analyse

Indien de onafhankelijke variabelen zowel van intervalniveau als van nominaal niveau zijn, spreekt men traditioneel van **covariantie-analyse**.



- A, B, C en D zijn *factoren*: categorische predictoren van nominaal niveau
- x_1, x_2 zijn *covariaten*: numerieke predictoren van minstens intervalniveau

3.4.4 Het nulmodel

Het nulmodel is een lineair model zonder onafhankelijke variabelen. Het bevat enkel een constante:

$$Y_i = \beta_0 + \varepsilon_i.$$

Het nulmodel dient als een soort referentiepunt: het 'slechtste model mogelijk'.

Andere modellen (met p onafhankelijke variabelen) worden met dit nulmodel vergeleken om na te gaan of de predicties significant beter zijn dan de predicties van het nulmodel.

Lineaire regressie
Variantie- en covariantie-analyse

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

1 Het lineair regressiemodel

1.1 Lineaire regressie

Regressie is een statistische techniek om het verband tussen één (*univariaat*) of meerdere (*multivariaat*) **uitkomst(en)** en een set van **predictoren** te onderzoeken. Een regressiemodel is een hypothetisch statistisch model dat de relatie tussen de uitkomst en de predictoren beschrijft. De invloed van de predictoren op de uitkomst(en) wordt gemodelleerd.

In dit hoofdstuk beschouwen we **univariate regressie**.

De uitkomst is de te verklaren variabele of **afhankelijke** variabele (Y) en de predictoren zijn de **onafhankelijke** variabelen (X). Men spreekt van regressie van Y op X. Y is gemeten op minstens intervalniveau.

We werken met het softwarepakket R. Alle data en R-code voor de voorbeelden uit de cursus worden ter beschikking gesteld (zie verder).

Lineaire regressie kan gebruikt worden om:

- het effect van predictoren op de uitkomst of de samenhang tussen uitkomst en predictoren na te gaan.
- toekomstige observaties te voorspellen, gegeven de waarden voor de predictoren.
- een algemene beschrijving van de datastructuur weer te geven.

Lineaire regressie is al aan bod gekomen in Statistiek II. We bouwen er hier op verder. Sommige zaken worden herhaald om nadien verder uit te breiden. Niet alles wat in Statistiek II aan bod komt, wordt hier expliciet herhaald, maar wordt wel beschouwd als voorkennis. Er zullen bvb. zaken aan bod komen in de oefenlessen die niet meer aan bod komen in de theorie, maar die geziene stof uit Statistiek II zijn.

Een lineair regressiemodel met p predictoren voor n onafhankelijke observaties wordt als volgt voorgesteld:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n \\ Y_i &= \beta_0 + \sum_{\ell=1}^{p} \beta_\ell x_{i\ell} + \varepsilon_i \qquad i = 1, \ldots, n \end{aligned} \tag{1}$$

- $x_{i\ell}$: score voor de predictor X_ℓ voor observatie i ($\ell = 1, \ldots, p$)

- β_0, β_1, ..., β_p: regressiecoëfficiënten (populatieparameters)

Het stochastisch deel van model (1) bestaat uit 3 assumpties met betrekking tot de fouttermen:

1. $E(\varepsilon_i) = 0$ voor alle i
2. $\text{Var}(\varepsilon_i) = \sigma^2$ voor alle i, i.e. constante variantie van de fouttermen of *homoscedasticiteit*
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ voor alle $i \neq j$

Bijkomende veronderstellingen: de predictoren zijn onafhankelijk van ε_i, de predictoren zijn zonder fout gemeten.

Onbekende parameters: β_0, β_1, ..., β_p, σ^2. In totaal zijn dit $p+2$ parameters, $p+1$ regressiecoëfficiënten en σ^2.

$E(Y_i|x_{i1}, x_{i2}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$
$\Rightarrow$ Betekenis β_ℓ ($\ell = 1, \ldots, p$): als de ℓ-de predictor (X_ℓ) met 1 eenheid stijgt terwijl alle overige predictoren constant blijven dan neemt de verwachte waarde van Y toe met β_ℓ.

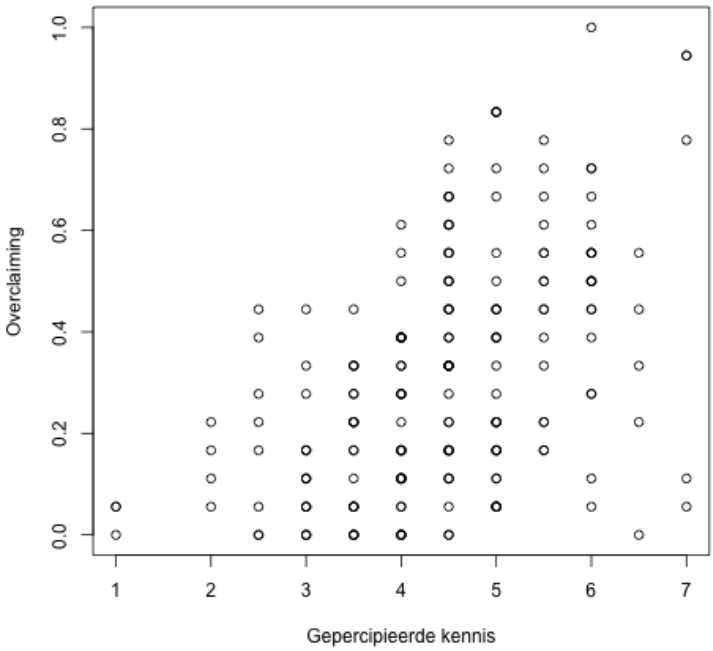
De term 'lineair' krijgt vaak een dubbele betekenis:

1. Primaire betekenis: het regressiemodel is lineair in de parameters.
   Model (1) is lineair in de parameters. De parameters komen niet voor in een niet-lineaire vorm zoals β_1^2, $\log(\beta_1)$, $\exp(\beta_2), \ldots$.
2. Vaak gebruikt men de term echter ook om naar de aard van de samenhang tussen de variabelen te verwijzen. In veel toepassingen gebruikt men immers een lineaire regressiefunctie, dit betekent dat het model ook lineair is in de predictor(en) X. Dit is niet noodzakelijk.
   $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ ($i = 1, \ldots, n$) is, in tegenstelling tot model (1), een statistisch model met een niet-lineair verband tussen de uitkomst en predictor, het verband is kwadratisch. Toch is dit model een lineair regressiemodel aangezien het lineair is in de parameters.

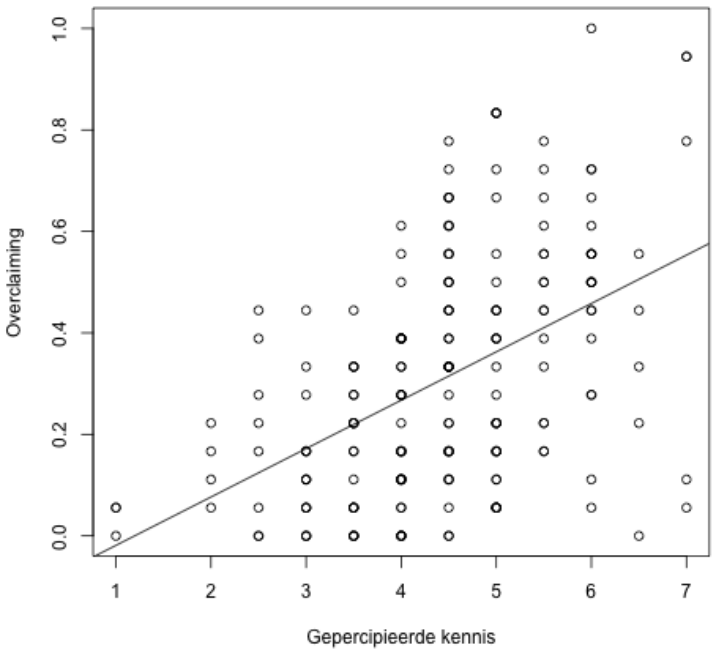
De term regressie impliceert vaak dat ook alle onafhankelijke variabelen van intervalniveau zijn. Dit is echter niet strikt noodzakelijk. We beschouwen ook regressie met enkel nominale onafhankelijke variabelen en regressie met onafhankelijke variabelen van zowel nominaal als intervalniveau.

Bij de start van de data-analyse is het nuttig om een *scatterplot* of *spreidingsdiagram* te maken. Hierbij wordt de afhankelijke variabele Y op de y-as gezet en de onafhankelijke

variabele X op de x-as. Veronderstel dat men geïnteresseerd is in de mate van overclaiming in functie van gepercipieerde kennis (zie sectie 1.2.1 voor meer uitleg over de data).



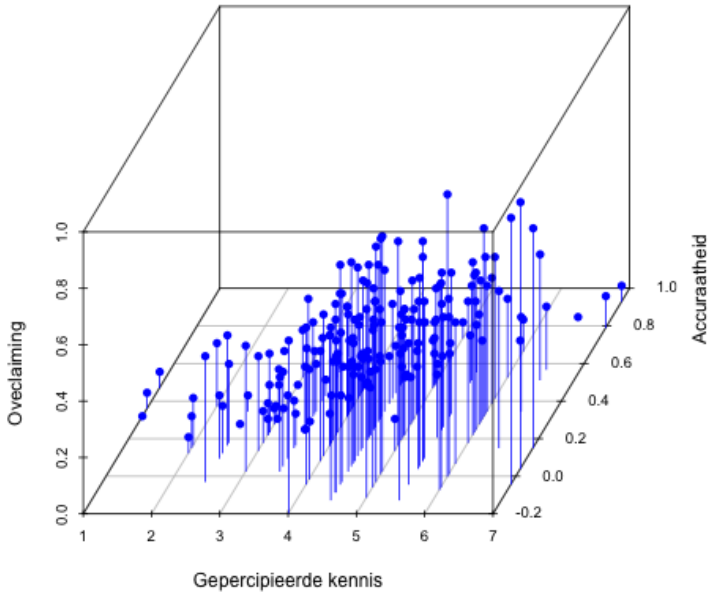
Met behulp van lineaire regressie kunnen we de best passende rechte door de puntenwolk bepalen, d.i. de rechte die zo goed mogelijk de trend van de gegevens benadert.



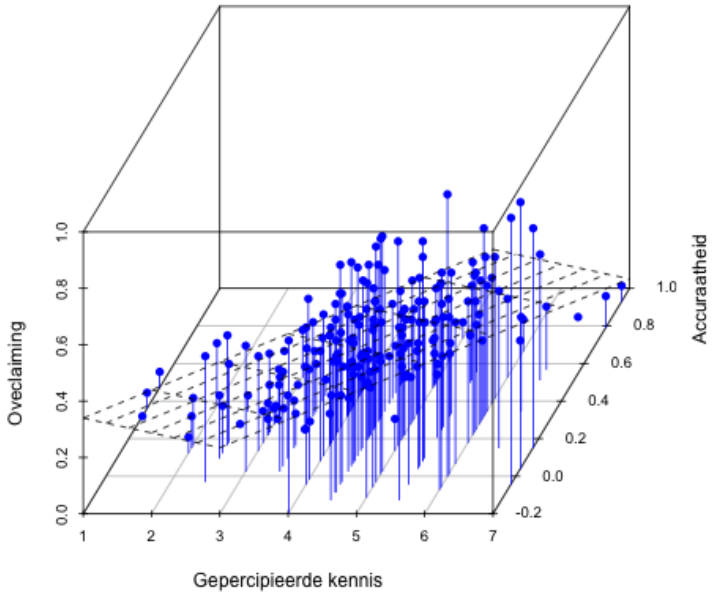
Algemeen kan elk lineair regressieprobleem waarbij p onafhankelijke variabelen of predictoren beschouwd worden, grafisch voorgesteld worden in $p + 1$ dimensies.

Wanneer we 2 predictoren beschouwen kunnen we analoog een 3-dimensioneel spreidingsdiagram maken en daardoor het best passende vlak tekenen.

Veronderstel dat we in bovenstaand voorbeeld de mate van overclaiming voorspellen in functie van gepercipieerde kennis en accuraatheid, dan ziet het 3-dimensioneel spreidingsdiagram er als volgt uit:



Met behulp van lineaire regressie kunnen we het best passende vlak door de puntenwolk bepalen:



In deze cursusnota's gaan we aan de slag met enkele voorbeelden die we gebruiken om een aantal technische aspecten m.b.t. lineaire regressie te bespreken. Niet alle zaken die aan bod kwamen in Statistiek I en Statistiek II worden herhaald. Zo zijn datavisualisaties, het univariaat verkennen van de dataset, het bekijken van onderlinge relaties tussen variabelen essentieel, alsook het nagaan van de assumpties die aan de basis liggen van de analyses. Aan deze zaken wordt de nodige aandacht besteed in de oefenlessen.

In de volgende sectie worden de voorbeelden die we doorheen de cursus gaan gebruiken, geïntroduceerd.

1.2 Voorbeelden

1.2.1 Overclaiming

We beschouwen hier data uit een studie uit de volgende paper:

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds:

Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26*, 1295-1303.

Bron: Open Stats Lab
https://sites.trinity.edu/osl/data-sets-and-activities/regression-activities

Zowel de code die hoort bij de analyses (`overclaiming.R`) als de data (`overclaiming.csv`) zijn terug te vinden op Ufora.

Achtergrond

Mensen kunnen hun eigen kennis overschatten, soms zelfs kennis van concepten, gebeurtenissen en mensen die niet bestaan en dus niet gekend kunnen zijn. Dit fenomeen heet *overclaiming*.

In deze studie wenst men na te gaan in welke mate zelf-gepercipieerde kennis de mate van overclaiming voorspelt.

Methode en design

202 participanten nemen deel aan de studie (85 vrouwen, 115 mannen, 2 personen van wie gender niet gekend is; gemiddelde leeftijd is 33.5 jaar met een standaarddeviatie gelijk aan 10.0). De steekproef werd getrokken uit een lijst van Amazon en het betreft enkel particpanten uit de Verenigde Staten.

De participanten vullen een vragenlijst in om hun algemene kennis rond persoonlijke financiën te scoren en een test om overclaiming te bepalen. Beide testen worden afgenomen in een countergebalanceerde volgorde over participanten.

Bij de overclaimingtaak worden 15 items gepresenteerd in een random volgorde, waarvan 12 items gaan over bestaande termen en 3 items over niet-bestaande termen uitgevonden door de onderzoekers.
We are interested in common knowledge about personal finance. You will see 15 terms related to personal finance. Please rate your knowledge about each term by choosing the appropriate number from 1 (never heard of it) to 7 (very knowledgeable).

Daarna wordt bij de participanten ook nog een test afgenomen rond financiële geletterdheid (FINRA Investor Education Foundation).

Data

De dataset bevat de volgende variabelen:

order_of_tasks De volgorde van de taken: `order_of_tasks=1` wanneer de participanten eerst de test rond gepercipieerde kennis afleggen en dan de test voor overclaiming; `order_of_tasks=2` wanneer de participanten eerst de test voor overclaiming afleggen en dan de test voor gepercipieerde zelfkennis

In de analyses brengen wij deze variabele niet mee in rekening.

self_perceived_knowledge Score voor gepercipieerde kennis, wordt als van intervalniveau verondersteld

De vragen voor algemene kennis rond persoonlijke financiën zijn als volgt:

> In general, how knowledgeable would you say you are about personal finance? (1 = not knowledgeable at all, 7 = extremely knowledgeable)
>
> How would you rate your general knowledge of personal finance compared to the average American? (1 = much less knowledgeable, 7 = much more knowledgeable)

Om de score te bekomen, wordt het gemiddelde genomen van het antwoord op beide vragen.

overclaiming_proportion Score voor overclaiming, wordt als van intervalniveau verondersteld

Overclaiming wordt gemeten door na te gaan voor welke proportie van onbestaande termen een participant kennis beweert te hebben. Deze proportie wordt berekend voor 6 cutoffs voor kennis: de proportie van onbestaande termen met een score van 2 of hoger, de proportie van onbestaande termen met een score van 3 of hoger enz. voor 4, 5, 6 en 7. Daarna wordt het gemiddelde genomen van deze proporties.

accuracy Accuraatheid, wordt als van intervalniveau verondersteld

Accuraatheid wordt bekomen door analoog als bij het bepalen van overclaiming, de proportie te bepalen van bestaande termen waarover de participant kennis beweerde te hebben en hiervan de proportie van onbestaande termen af te trekken.

FINRA_score Score op de test voor financiële geletterdheid, wordt als van intervalniveau verondersteld (indicator voor werkelijke kennis)

Onderzoeksvraag

De onderzoekers wensen na te gaan in welke mate gepercipieerde kennis overclaiming voorspelt, rekening houdend met (i.e. controlerend voor) accuraatheid en FINRA-score.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 18, 23, 26, 29, 32, 37, 39, 41, 52, 54 en 56.

1.2.2 Pijneducatie bij onderrugpijn

We beschouwen hier een voorbeeld van een quasi-experiment. De fictieve data zijn gebaseerd op de volgende studie:

Mosely, G.L., 2004, Evidence for a direct relationship between cognitive and psychological change during an education intervention in people with chronic low back pain. *European Journal of Pain, 8*, 39-45

In de oefeningensessies wordt hetzelfde voorbeeld hernomen.

Zowel de code die hoort bij de analyses (`pijneducatie.R`) als de data (`pijneducatie.csv`) zijn terug te vinden op Ufora.

Achtergrond

Men stelt vast dat naast puur fysieke factoren ook cognitieve factoren een rol spelen bij pijnperceptie. In deze studie stelt men zich de vraag of deze cognities een *actieve* rol spelen in de pijnperceptie bij mensen met klachten over pijn in de onderrug.

Bovendien bestaat er in de literatuur ook evidentie voor verschillende soorten van betrokken cognities: enerzijds meer algemene cognities over pijn en anderzijds meer specifieke pijn-locatie gerelateerde cognities.

Om dit in meer detail na te gaan voeren de onderzoekers een manipulatie van verschillende pijn-gerelateerde cognities uit en gaan ze na wat de effecten op pijnperceptie zijn. Daarnaast wensen de auteurs ook te controleren voor comorbiditeit van depressie en leeftijd.

Men wenst te onderzoeken of er evidentie is voor het bestaan van verschillen tussen de onderliggende cognities bij pijnperceptie.

In een poging om een zuivere pijnindicator te definiëren, wordt bij de patiënten gemeten hoe ver ze voorover kunnen buigen.

Methode en design

De patiënten zijn geselecteerd via een lokaal ziekenhuis ($n = 121$) waar één van de onderzoekers makkelijk toegang tot heeft. Subjecten worden geweerd indien er comorbiditeit met andere fysieke of zuiver neurologische aandoeningen is. Alle patiënten ondergaan een één-op-één sessie met een therapeut waarin een uitgebreide pijneducatie sessie plaats vindt.

Condities (pijn-educatie groepen)

We onderscheiden de volgende condities:

- Conditie 1: hier ligt de focus op cognities die gerelateerd zijn aan de algemene pijnfysiologie.
  Deze conditie belicht de werking van pijn-, druk- en andere receptoren binnen het centrale en perifere zenuwstelsel.
- Conditie 2: de nadruk ligt hier op de fysiologie van de ruggengraat.
  Er wordt belicht hoe de verschillende wervels samenzitten in de ruggengraat.
- Conditie 3: dit is de baseline conditie.
  In deze conditie gaat men dieper in op de algemene werking van het maag- en darmstelsel.

Toewijzing

De eerste 41 patiënten worden toegewezen aan de eerste conditie, de volgende 40 aan de tweede conditie. Tot slot worden de laatste 40 patiënten toegewezen aan de baseline conditie.

Data

Afhankelijke variabele:

> Er wordt bij iedere deelnemer 2 keer gemeten hoe ver men voorover kan buigen (in mm): 1 keer vóór en 1 keer na de experimentele conditie (conditie 1, 2 of 3). Er wordt vervolgens een verschilscore berekend (na-vóór). Positieve verschilscores wijzen op een verbetering.

Onafhankelijke variabelen en predictoren:

- Een nominale variabele duidt de 3 condities aan.
- Leeftijd van de deelnemers
- Depressiescore; deze score mag als van intervalniveau beschouwd worden.

De dataset bevat bijgevolg de volgende variabelen:

Buig Verschilscore hoe ver iemand voorover kan buigen (na-vóór), in mm uitgedrukt

Gender Geslacht; Vrouw =1, man =0

Leeft Leeftijd in jaren

Conditie Conditie=1: algemene pijn-fysiologie groep, Conditie=2: Ruggengraat-educatie groep, Conditie=3: baseline groep

Dep Depressiescore

Onderzoeksvragen

Men wil nagaan of er een verschil is in gemiddelde uitkomst tussen de verschillende onderliggende cognities bij pijnperceptie.

Aangezien het hier geen gerandomiseerde studie betreft (deelnemers zijn niet at random toegewezen aan de condities) is het zinvol om het onderzochte effect ook te corrigeren voor het effect van leeftijd en de graad van depressie. Er kan ook nagegaan worden of effecten variëren volgens leeftijd en/of depressie.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 44, 57 en 59.

1.2.3 Herstel na coma

We gebruiken een subset van de data uit de volgende paper:

Wong, P. P., Monette, G., & Weiner, N. I. (2001) Mathematical models of cognitive recovery. *Brain Injury, 15*, 519530.

Zowel de code die hoort bij de analyses (`coma.R`) als de data (`coma.csv`) zijn terug te vinden op Ufora.

Achtergrond

De data die we gebruiken zijn een onderdeel van een longitudinale studie waarin modellen voor IQ opgesteld worden om herstelverloop na een coma te voorspellen.

200 patiënten die als gevolg van een traumatisch hersenletsel in coma lagen voor bepaalde tijd, worden na het ontwaken opgevolgd en er worden periodisch standaard IQ-testen afgenomen. Deze testen kunnen afgenomen worden na de coma, na het uiteindelijke herstel en op alle tijdspunten tussenin. Op die manier kan het herstelverloop van IQ onderzocht worden.

Wij beschouwen voor iedere patiënt slechts 1 meting.

De originele volledige dataset is beschikbaar in het R-package `carData`.

Data

Wij beschouwen de volgende variabelen:

duration Duur van de coma (in dagen)

sex Geslacht van de patiënt

age Leeftijd (in jaren) op moment van trauma

piq Mathematisch (*performance*) IQ

viq Verbaal IQ

duration_cat Nominale variabele met categorieën volgens duur van coma

> Categorieën zijn: t.e.m. 1 dag, meer dan 1 dag t.e.m. 7 dagen, meer dan 7 dagen t.e.m. 14 dagen, meer dan 14 dagen t.e.m. 255 dagen

Wij gebruiken de dataset hoofdzakelijk op een exploratieve manier om interacties tussen variabelen beter te begrijpen. We beschouwen hierbij een lineair regressiemodel met mathematisch IQ (`piq`) als afhankelijke variabele.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden in sectie 5.4 op pagina 70.

1.2.4 Levenstevredenheid

We beschouwen hier een voorbeeld met fictieve data uit de paper van Preacher & Hayes (2004):

Preacher, K.J., Hayes, & A.F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717-731.

Men is geïnteresseerd in het onderzoeken van het proces of de manier waarop cognitieve gedragstherapie een invloed heeft op levenstevredenheid na pensionering bij klinisch depressieve personen.

Zowel de code die hoort bij de analyses (`satisfaction.R`) als de data (`satisfaction.csv`) zijn terug te vinden op Ufora.

Methode en design

Dertig klinisch depressieve patiënten worden at random toegewezen aan de conditie die bestaat uit 10 sessies van de nieuwe therapie of aan de conditie die bestaat uit 10 sessies van een standaardtherapie .

Na sessie 8 wordt de positiviteit bepaald van de attributies die de deelnemers maken voor een recente negatieve ervaring (attribueren = toekennen van oorzaken). Op het einde van sessie 10 wordt de algemene levenstevredenheid gemeten.

Data

De dataset bevat de volgende variabelen:

satis Algemene levenstevredenheid (wordt als van intervalniveau verondersteld), gemeten op sessie 10

therapy Therapie; `therapy = 1`: nieuwe therapie, `therapy = 0`: standaardtherapie

attrib Attributie (wordt als van intervalniveau verondersteld), gemeten op sessie 8

Onderzoeksvraag

De onderzoeksvraag is of het eventuele effect van de cognitieve therapie op de algemene levenstevredenheid (deels) verloopt via attributie, m.a.w. of cognitieve therapie een invloed heeft op attributie die dan op zijn beurt een invloed heeft op de algemene levenstevredenheid. De onderzoeksvraag is dus of het effect van de cognitieve gedragstherapie *gemedieerd* wordt door de positiviteit van de attributies.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 93.

1.3 Matrixnotaties

Aangezien een lineair regressiemodel met p predictoren voor n observaties als volgt is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

houdt dit in:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n \end{aligned}$$

Aan de hand van matrixnotaties kunnen we dit model compact als volgt noteren:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

waarbij:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

We hebben dus

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times(p+1)} \underbrace{\boldsymbol{\beta}}_{(p+1)\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n\times 1}$$

De matrix $\mathbf{X}$ wordt de **designmatrix** genoemd. Deze matrix is gedefinieerd als een kolom van 1-en en p kolommen die telkens de n observaties van de predictoren $X_1, X_2, \ldots, X_p$ bevatten. De kolom met 1-en wordt gebruikt als het lineaire model een intercept bevat (d.i. als $\beta_0 \neq 0$).

Wanneer we in het voorbeeld rond overclaiming (zie sectie 1.2.1), de uitkomst overclaiming regresseren op gepercipieerde kennis, accuraatheid en FINRA-score, dan ziet de designmatrix er als volgt uit (we tonen hier enkel de vier eerste rijen):

```
  (Intercept) self_perceived_knowledge    accuracy FINRA_score
1           1                      5.5  0.25000000           4
2           1                      4.5  0.19444444           4
3           1                      3.5  0.34722222           5
4           1                      6.0 -0.05555556           4
```

Verder geldt dat $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ aangezien:

$$E(\boldsymbol{\varepsilon}) = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Hieruit volgt dat $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$.

Aangezien $\text{Var}(\varepsilon_i) = \sigma^2$ voor alle i en $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ voor alle $i \neq j$, hebben we:

$$\text{Var}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 \mathbf{I}.$$

1.4 Parameterschattingen

β_0, β_1, …, β_p en σ^2 zijn onbekende parameters.

Op basis van n observaties wensen we de onbekende populatieparameters te schatten.

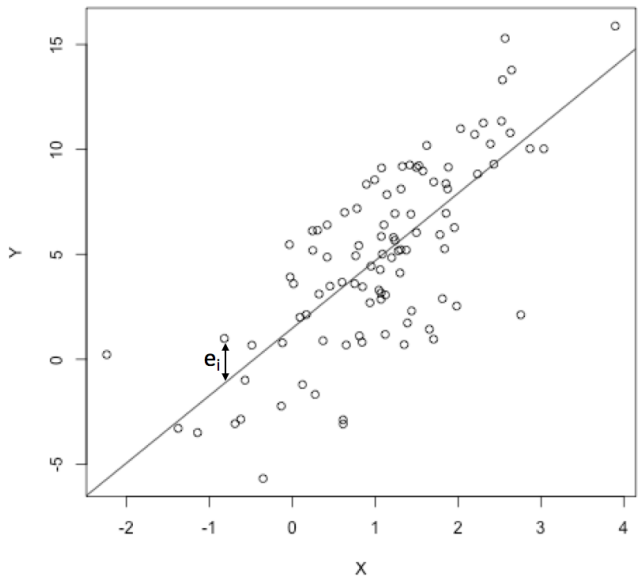
Concreet: gegeven een lukraak getrokken steekproef van n observaties $Y_1, \dots, Y_n$ en bijhorende predictoren $x_{1\ell}, \dots, x_{n\ell}$ ($\ell = 1, \dots, p$) wensen we β_0, β_1, …, β_p en σ^2 te schatten.

Twee courante methodes:

- Methode van de kleinste kwadraten
- Methode van de maximale aannemelijkheid

1.4.1 Kleinste kwadratenschatters voor β

Grafisch (2 dimensies, d.i. enkelvoudig):



De best passende rechte wordt door de puntenwolk getekend.

De kleinste kwadraten regressielijn is een unieke rechte die men bekomt door de som van de verticale kwadratische afstanden (afwijkingen) tussen ieder datapunt en de rechte te minimaliseren.

Analoog voor meerdere dimensies: de regressievergelijking die men bekomt via de methode van de kleinste kwadraten is de vergelijking van het hypervlak waarvoor de som van de kwadratische afwijkingen tussen ieder datapunt en dat vlak geminimaliseerd is.

Met **B** duiden we de vector van de kleinste kwadratenschatters aan:

$$\begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_p \end{bmatrix}.$$

B_0, B_1, …, B_p zijn de **schatters** voor respectievelijk β_0, β_1, …, β_p.

Schatters zijn kansvariabelen en worden dus genoteerd met grote letters: wanneer we herhaaldelijk een andere lukrake steekproef nemen waarbij de predictoren constant gehouden worden, zouden we telkens andere waarden bekomen voor deze schatters. De verdeling van schatters noemt men de **steekproevenverdeling**.

b_0, b_1, ..., b_p zijn de **puntschattingen**, i.e. een concreet getal in een gegeven steekproef.

Noteer $\hat{Y}_i = B_0 + B_1 x_{i1} + B_2 x_{i2} + \dots B_p x_{ip}$, $\hat{Y}_i$ is de gefitte waarde van Y_i. Merk op dat $\hat{Y}_i$ een schatter is voor $E(Y_i|x_{i1}, \dots, x_{ip})$ en niet voor Y_i!

De vector van gefitte waarden wordt voorgesteld door $\hat{\mathbf{Y}}$:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{XB}.$$

$E_i = Y_i - \hat{Y}_i$ is de afwijking tussen Y_i en $\hat{Y}_i$, E_i is het residu van Y_i.

Technisch gezien minimaliseert de kleinste kwadratenmethode $\sum_{i=1}^{n} E_i^2$.

$\sum_{i=1}^{n} E_i^2$ is de **residuele** of **fout kwadratensom**, in het Engels wordt deze term *residual/error sum of squares* genoemd of kortweg SSE [1].

De SSE kan men beschouwen als een maat voor de fout van het regressiemodel, of nog, een maat voor het verlies aan informatie door Y_i te vervangen door $\hat{Y}_i$.

De vector van residuen wordt voorgesteld door $\mathbf{E}$

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Er kan aangetoond worden dat:

$$\underbrace{\mathbf{B}}_{(p+1)\times 1} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{(p+1)\times(p+1)} \underbrace{\mathbf{X}'\mathbf{Y}}_{(p+1)\times 1}$$

Er bestaat slechts een unieke oplossing voor de kleinste kwadratenschatters indien de inverse van de matrix $\mathbf{X}'\mathbf{X}$ bestaat. Dit betekent dat deze matrix van volledige rang of niet-singulier moet zijn.

[1] In de literatuur wordt zowel de term 'residual sum of squares' als 'error sum of squares' gebruikt. Beiden zijn gelijk. Er geldt dus SSE=SS_{Res}.

Waarom zijn kleinste kwadratenschatters goede schatters?

1. Ee schatters hebben een zinvolle meetkundige betekenis (denk aan het minimaliseren van de som van de kwadratische afwijkingen van de datapunten tot de regressierechte in 2 dimensies).
2. Wanneer de fouttermen (ε_i) onafhankelijk en identisch normaal verdeeld zijn, zijn de kleinste kwadratenschatters gelijk aan de **maximum likelihood schatters**. Dit zijn de schatters die bekomen worden via de methode van de maximale aannemelijkheid.

   Via maximum likelihood worden de waarden voor de parameters gekozen die het meest aannemelijk zijn, gegeven de geobserveerde data.
3. Het Gauss-Markov theorema stelt dat de kleinste kwadratenschatters de beste lineaire zuivere schatters zijn (zie volgende sectie).

1.4.2 Het Gauss-Markov theorema

Het Gauss-Markov theorema stelt:

Indien $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ en de structurele component van het model correct is ($E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$) dan zijn de kleinste kwadratenschatters **zuiver**, **efficiënt** en **lineair**.

- Het **zuiver** of **onvertekend** zijn van de schatters impliceert:

  $$E(\mathbf{B}) = \boldsymbol{\beta}.$$

  Concreet: $E[B_0] = \beta_0$, $E[B_1] = \beta_1$, …, $E[B_p] = \beta_p$.
  Dit betekent dat de verwachte waarde van de schatters gelijk is aan de regressiecoëfficiënten (populatieparameters), de gezochte parameters.

- De schatters zijn **efficiënt** (of nog: de 'beste') omdat hun variantie (i.e. de variantie van hun steekproevenverdeling) minimaal is (in vergelijking met alle *onvertekende lineaire* schatters).

  $$\text{Var}(\mathbf{B}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

  σ^2 is ongekend en moet ook geschat worden (zie volgende sectie). Merk opnieuw op dat de standaarddeviatie (vierkantswortel van de variantie) van een schatter ook de **standaardfout** genoemd wordt.

- De schatters noemt men **lineair** omdat ze een lineaire functie zijn van de observaties Y_i.

Het theorema toont aan dat de kleinste kwadratenschatters een goede keuze zijn, maar wanneer de fouttermen bvb. gecorreleerd zijn of ongelijke varianties hebben, bestaan er betere schatters. We gaan daar in deze cursus niet verder op in.

1.4.3 Een schatter voor σ^2

Een schatter voor σ^2 kan als volgt bekomen worden:

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} E_i^2}{n-(p+1)} = \frac{\text{SSE}}{n-(p+1)}.$$

In totaal worden $p+1$ regressiecoëfficiënten geschat, namelijk β_0, β_1, …, β_p. Dit betekent dat dat we $p+1$ vrijheidsgraden verliezen, daarom delen we door $n-(p+1)$.

S^2 wordt ook wel MSE genoemd (*Mean Squared Error*[2]).

S^2 is een onvertekende schatter voor σ^2: $E[S^2] = \sigma^2$.

1.4.4 Voorbeeld: overclaiming

We bekijken het effect van gepercipieerde kennis op overclaiming (zie sectie 1.2.1) op basis van een lineair regressiemodel.

```
> fit1_expertise<-lm(overclaiming_proportion~self_perceived_knowledge,data=expertise)
> fit1_expertise

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge,
data = expertise)

Coefficients:
(Intercept)  self_perceived_knowledge
-0.11406                     0.09532
```

We lezen af dat $b_0 = -0.114$ en $b_1 = 0.0953$. Omdat b_1 positief is, hebben we een stijgende geschatte regressierechte: er is een (al dan niet significante) stijgende trend of een positief verband tussen gepercipieerde kennis en overclaiming. We schatten dat, indien gepercipieerde

[2] Alternatieve notatie: MS_{Res}

kennis met 1 eenheid stijgt, de gemiddelde proportie overclaiming met 0.0953 eenheden toeneemt.

Via de functie `summary` krijgen we meer informatie over het regressiemodel.

```
> summary(fit1_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge,
data = expertise)

Residuals:
Min       1Q   Median       3Q      Max
-0.50551 -0.15610  0.00662  0.12167  0.54215

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.11406    0.05624  -2.028   0.0439 *
self_perceived_knowledge  0.09532    0.01228   7.762 4.22e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2041 on 200 degrees of freedom
Multiple R-squared:  0.2315,Adjusted R-squared:  0.2277
F-statistic: 60.25 on 1 and 200 DF,  p-value: 4.225e-13
```

In de kolom `Estimate` staan de geschatte regressiecoëfficiënten. In de kolom `Std. Error` staan de geschatte standaardfouten. Hier is $s_{B_0} = 0.0562$ en $s_{B_1} = 0.0123$. We lezen verder af dat de `Residual standard error` gelijk is aan 0.204, dit is s met s^2 een schatting voor σ^2.

Hoewel de onderzoekers geïnteresseerd zijn in het verband tussen overclaiming en gepercipieerde kennis, zijn er meerdere mogelijke predictoren voor overclaiming. We voegen enkele van deze predictoren toe aan het regressiemodel. Dit laat o.a. toe om het effect van gepercipieerde kennis te bekijken, conditioneel op de andere predictoren.

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                        +FINRA_score,data=expertise)
> summary(fit3_expertise)

Call:
```

```
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)

Residuals:
Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

We zien nog steeds een positief verband tussen gepercipieerde kennis en overclaiming: we schatten dat voor een constante accuraatheid en FINRA-score, de gemiddelde proportie overclaiming met 0.0940 eenheden toeneemt als gepercipieerde kennis met 1 eenheid stijgt.

We zien verder een negatief effect van accuraatheid: we schatten dat voor een constante gepercipieerde kennis en FINRA-score, de gemiddelde proportie overclaiming met 0.793 eenheden afneemt als accuraatheid met 1 eenheid stijgt. Hoewel deze interpretatie wiskundig correct is, is ze niet zo nuttig in dit geval. De verdeling van de variabele `accuracy` ziet er als volgt uit:

```
> summary(expertise$accuracy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.1944  0.1562  0.2847  0.2953  0.4549  0.9306
```

Een toename van 1 eenheid is dus niet betekenisvol. Het is interessanter om het effect te bekijken van bvb. een toename van 0.10 eenheden in accuracy (terwijl de overige predictoren constant blijven). We schatten in dat geval dat de gemiddelde proportie overclaiming met $0.793 \times 0.10 = 0.0793$ eenheden afneemt.

1.5 Predictie

1.5.1 Predicties

Gegeven de schatters voor de regressiecoëfficiënten, kunnen we op basis van het regressiemodel de (verwachte) waarde van Y_* voorspellen op basis van nieuwe (niet eerder geobserveerde) waarden voor de set van predictoren x_{1*}, x_{2*}, ..., x_{p*}:

$$\hat{Y}_* = B_0 + B_1 x_{1*} + B_2 x_{2*} + \ldots + B_p x_{p*}.$$

"*Wat is de verwachte mate van overclaiming, gegeven de scores voor gepercipieerde kennis en accuraatheid en de FINRA-score?*"

Vaak wenst men predicties te doen voor een 'typisch' profiel. Dan kan men bvb. een predictie doen waarbij elke predictor gelijk gesteld wordt aan het steekproefgemiddelde. We hernemen het voorbeeld rond overclaiming (zie sectie 1.2.1).

```
> mean(expertise$self_perceived_knowledge)
[1] 4.428218
> mean(expertise$accuracy)
[1] 0.2953108
> mean(expertise$FINRA_score)
[1] 3.69802
```

De geschatte regressiecoëfficiënten van het model waarbij overclaiming geregresseerd wordt op gepercipieerde kennis, accuraatheid en FINRA-score, zijn als volgt:

```
> fit3_expertise

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)

Coefficients:
(Intercept)  self_perceived_knowledge                 accuracy
0.05779                   0.09407                  -0.79322
FINRA_score
0.01837
```

De voorspelde mate van overclaiming voor een typisch profiel is bijgevolg gelijk aan $0.05779 + 0.09407 \times 4.428218 - 0.79322 \times 0.2953108 + 0.01837 \times 3.69802 = 0.308$.

In R kunnen we predicties op een eenvoudige manier bekomen door een dataframe te maken waarin de waarden voor de predictoren ingevuld worden en op basis van het geschatte model de uitkomst te voorspellen voor de waarden in de nieuwe dataframe.

```
>avprofiel<-data.frame(self_perceived_knowledge=mean(expertise$self_perceived_knowledge),
                       accuracy=mean(expertise$accuracy),FINRA_score=mean(expertise$FINRA_score))
> predict(fit3_expertise,newdata=avprofiel)
1
0.3080308
```

Bij het voorspellen van de uitkomst op basis van nieuwe waarden voor de set van predictoren x_{1*}, x_{2*}, ..., x_{p*} in het model, kunnen we een onderscheid maken tussen 2 zaken:

- het voorspellen van de verwachting van een toekomstige waarde : $E(Y_*|x_{1*}, x_{2*}, \ldots, x_{p*})$, een gemiddelde
- het voorspellen van een toekomstige waarde: Y_*, een individuele observatie

De schatters voor beide gevallen zijn dezelfde en gelijk aan
$\hat{Y}_* = B_0 + B_1 x_{1*} + B_2 x_{2*} + \ldots + B_p x_{p*}$.

Bij intervalschattingen kunnen we een onderscheid maken tussen betrouwbaarheids- en predictie-intervallen.

1.5.2 Betrouwbaarheidsintervallen

Om betrouwbaarheidsintervallen voor predicties te kunnen opstellen maken we gebruik van distributionele assumpties en daarom is een extra assumptie m.b.t. de fouttermen noodzakelijk:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

(i.i.d.= onafhankelijk en identisch verdeeld)
Dit betekent dat de fouttermen ε_i allen onafhankelijk zijn van elkaar en normaal verdeeld zijn. Als gevolg daarvan zijn de schatters voor de regressiecoëfficiënten $\mathbf{B}$ normaal verdeeld met verwachtingswaarde en variantie zoals hier boven aangegeven. σ^2 is echter ongekend en wordt geschat door S^2.

$$\begin{aligned} E(\mathbf{B}) &= \boldsymbol{\beta} \\ \widehat{\mathrm{Var}}(\mathbf{B}) &= S^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Als bovenstaande assumptie m.b.t. de fouttermen geldt, dan geldt ook dat de individuele observaties Y_i normaal verdeeld zijn met verwachtingswaarde $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ en variantie σ^2.

Een $(1-\alpha) \times 100\%$ betrouwbaarheidsinterval voor $E(Y_*|x_{1*}, x_{2*}, \ldots, x_{p*})$ wordt gegeven door

$$\left[\hat{Y}_* - |t_{n-(p+1);\alpha/2}| S\left\{\hat{Y}_*\right\}, \hat{Y}_* + |t_{n-(p+1);\alpha/2}| S\left\{\hat{Y}_*\right\}\right]$$

met p het aantal predictoren in het model en $S\left\{\hat{Y}_*\right\}$ de standaardfout van het geschatte gemiddelde (de uitdrukking kan afgeleid worden maar komt hier niet aan bod).

Dit interval bevat met kans $1-\alpha$ de gemiddelde waarde voor Y_* (op populatieniveau!), gegeven de set van p predictoren. Op basis van het geschatte regressiemodel kunnen we bijgevolg naast een schatter voor $E(Y_*|x_{1*}, x_{2*}, \ldots, x_{p*})$ ook een interval opstellen waarvan met een bepaalde betrouwbaarheid gesteld kan worden dat het de verwachte waarde bevat.

In R kunnen we naast de predicties ook een bijhorend betrouwbaarheidsinterval opvragen.

```
> predict(fit3_expertise,newdata=avprofiel,interval='confidence')
    fit       lwr       upr
1 0.3080308 0.2905969 0.3254647
```

Onder `lwr` lezen we de ondergrens van het interval af, onder `upr` de bovengrens. Standaard krijgen we een 95% betrouwbaarheidsinterval. Het betrouwbaarheidsniveau kan ook aangepast worden, bvb. een 90% betrouwbaarheidsinterval bekomen we als volgt:

```
> predict(fit3_expertise,newdata=avprofiel,interval='confidence',level=0.90)
    fit       lwr       upr
1 0.3080308 0.2934209 0.3226407
```

1.5.3 Predictie-intervallen

Bij het opstellen van predictie-intervallen wordt ook de assumptie van normaal verdeelde fouttermen gemaakt:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Een $(1-\alpha) \times 100\%$ predictie-interval voor Y_* wordt gegeven door

$$\left[\hat{Y}_* - |t_{n-(p+1);\alpha/2}| S_{\text{pred}}\left\{\hat{Y}_*\right\}, \hat{Y}_* + |t_{n-(p+1);\alpha/2}| S_{\text{pred}}\left\{\hat{Y}_*\right\}\right]$$

met $S_{\text{pred}}\left\{\hat{Y}_*\right\}$ de standaardfout van de predictie. We moeten hier niet enkel de onzekerheid van de geschatte regressiecoëfficiënten in rekening brengen maar ook de spreiding op de individuele observaties. Bijgevolg is $S_{\text{pred}}\left\{\hat{Y}_*\right\} > S\left\{\hat{Y}_*\right\}$ en is een predictie-interval breder dan een betrouwbaarheidsinterval.

Een $(1-\alpha)\times 100\%$ predictie-interval bevat met kans $1-\alpha$ de (toekomstige) uitkomst Y_* (gegeven de set van p predictoren).

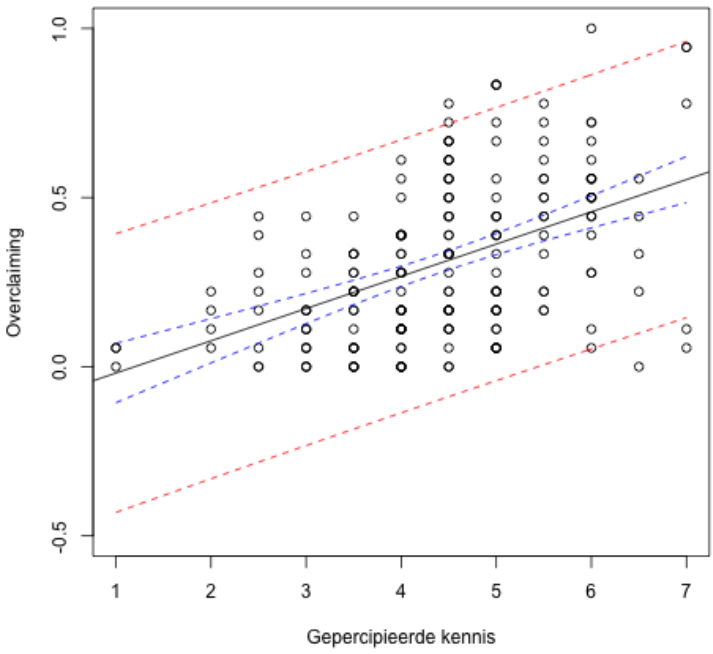
Voor het voorbeeld rond overclaiming (zie sectie 1.2.1) krijgen we volgend 95% predictie-interval voor een 'typisch' profiel:

```
> predict(fit3_expertise,newdata=avprofiel,interval='prediction')
    fit        lwr       upr
1 0.3080308 0.05963625 0.5564254
```

Een 99% predictie-interval bekomen we als volgt:

```
> predict(fit3_expertise,newdata=avprofiel,interval='prediction',level=0.99)
    fit         lwr       upr
1 0.3080308 -0.01957595 0.6356376
```

Om grafisch het verschil tussen een betrouwbaarheidsinterval en predictie-interval te demonstreren, kijken we naar het enkelvoudig regressiemodel waarbij overclaiming geregresseerd wordt op gepercipieerde kennis. De figuur toont voor elke waarde van gepercipieerde kennis de voorspelde mate van overclaiming (i.e. de regressierechte), met bijbehorend betrouwbaarheidsinterval (blauw) en predictie-interval (rood).



Voor elke waarde van gepercipieerde kennis is het betrouwbaarheidsinterval inderdaad smaller dan het predictie-interval. Verder zien we dat het betrouwbaarheidsinterval smaller is in het midden: de standaardfouten van de predicties zijn het kleinst in het midden van de puntenwolk.

1.6 Modelassumpties en invloedrijke observaties

Om de assumpties van het regressiemodel na te gaan kunnen diagnostische plots van de geobserveerde residuen gemaakt worden.

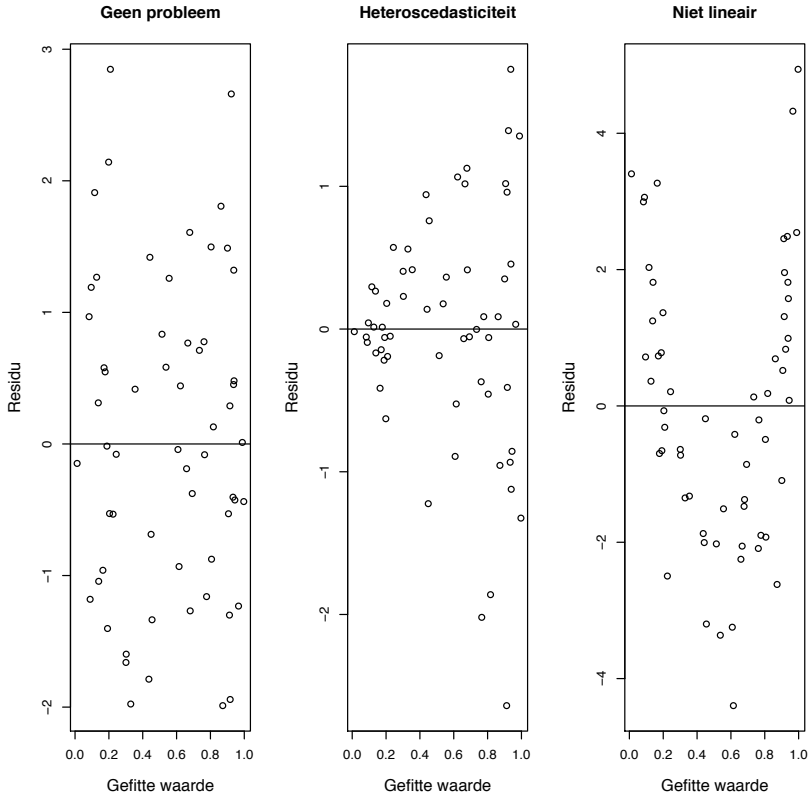
Eén van de belangrijkste plots is een spreidingsdiagram van de geobserveerde residuen e_i in functie van de gefitte waarden $\hat{y}_i$. Op deze plot kunnen we nagaan of

- er sprake is van heteroscedasticiteit (niet-constante variantie van de residuen).
- er sprake is van een niet-lineaire relatie tussen de predictoren en de uitkomst.

  Dit is het geval als er op de figuur duidelijk een trend waar te nemen is, bvb. een kwadratische trend.

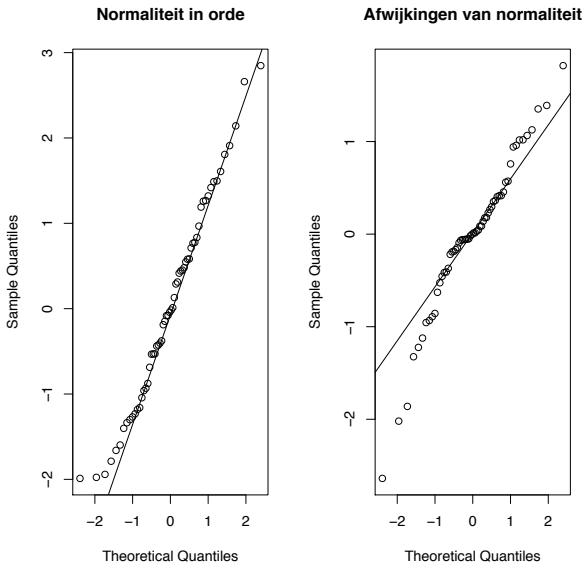
Dit kan men verhelpen door het structurele deel van het model aan te passen, bvb. door ook kwadraten van sommige predictoren in het model op te nemen.

Als alles in orde is, liggen de residuen symmetrisch rond 0 gespreid. De onderstaande figuur illustreert de verschillende gevallen.



Verder moeten we ook nagaan of de residuen normaal verdeeld zijn (indien we gebruik maken van intervalschattingen en/of toetsen):

- Maak een histogram of een boxplot van de gestandaardiseerde residuen, hiermee kan de symmetrie van de verdeling bekeken worden.
- Maak een normale QQ-plot (quantile-quantile plot) van de gestandaardiseerde residuen. Op deze figuur worden de geobserveerde kwantielen t.o.v. de verwachte kwantielen onder de normale verdeling geplot. Systematische afwijkingen van de rechte betekenen afwijkingen van de normaliteit.



In R kunnen diagnostische plots op een eenvoudige manier verkregen worden via de functie `plot` van het object dat het gefitte lineair regressiemodel omvat. Herneem het voorbeeld rond overclaiming (zie sectie 1.2.1).

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                    +FINRA_score,data=expertise)
> plot(fit3_expertise)
```

Dit commando geeft ons 4 verschillende plots:

Residuals vs fitted Zie hierboven. Op de plot geeft een rode lijn de geobserveerde trend in de residuen weer. We verwachten een horizontale lijn (op 0).

Normal QQ QQ-plot van de residuen.

Scale-Location Deze plot toont de vierkantswortel van de absolute waarde van de gestandaardiseerde residuen in functie van de gefitte waarden. Een rode lijn geeft de geobserveerde trend weer. In het geval van homoscedasticiteit verwachten we een horizontale lijn (niet noodzakelijk rond 0!).

Residuals vs Leverage De plot toont de gestandaardiseerde residuen t.o.v. de *leverage*. Deze plot laat toe om invloedrijke observaties te detecteren. Dit zijn observaties die een (grote) invloed hebben op de geschatte regressievergelijking.

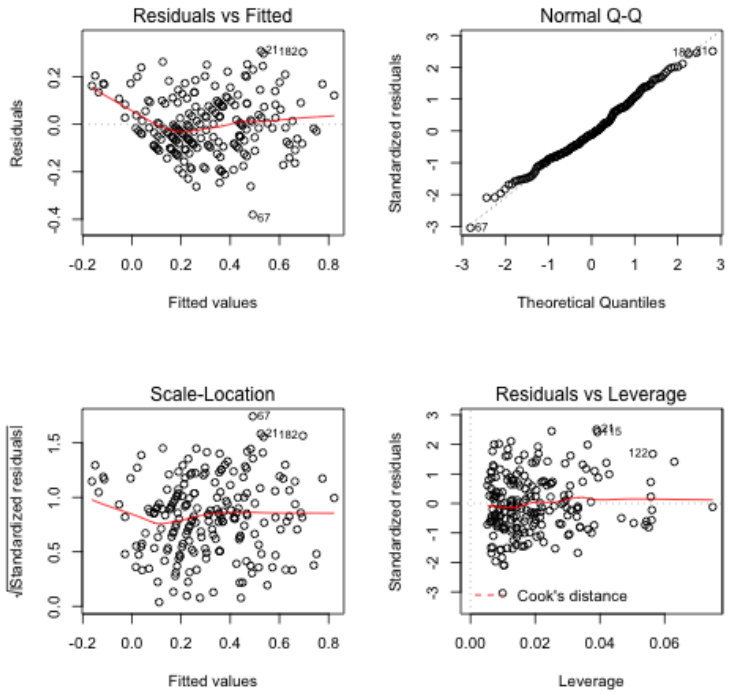
Observaties met een hoge leverage zijn observaties met extreme waarden voor de predictor(en). In het geval van p predictoren is de leverage een afstandsmaat tussen de vector van de p geobserveerde scores voor observatie i en de vector met de gemiddeldes over alle observaties. Een vuistregel stelt dat de leverage groot is als ze groter is dan $2p/n$ met n het aantal observaties. Een punt met extreme waarden voor de predictoren heeft potentieel een grote invloed, zelfs al ligt het niet zo ver van de regressielijn of -vlak; het is immers best mogelijk dat dat punt de regressielijn- of vlak al sterk naar zich toe getrokken heeft.

Het is ook mogelijk dat punten met extreme waarden voor de predictoren toch weinig invloed hebben op de geschatte regressievergelijking. Om de invloed van een welbepaalde observatie na te gaan, kan men de regressie uitvoeren met en zonder die observatie. Cook's distance biedt inzicht in de impact van een welbepaalde observatie i op de predictie van alle andere punten (en niet louter op de predictie van de observatie zelf). Een vuistregel luidt om waarden vanaf 0.8 (voor kleine n) en zeker vanaf 1 als een aanduiding van sterke invloed te zien.

Op deze plot zijn observaties buiten een rode gestreepte lijn observaties met een hoge Cook's distance. Resultaten kunnen drastisch wijzigen wanneer deze observaties uit de analyse weggelaten worden.

Merk op dat R op bovenstaande plots soms het observatienummer naast een punt zet. Dit punt kan dan als een outlier of afwijkende waarde beschouwd worden (maar dit is niet noodzakelijk een invloedrijke observatie). Het loont de moeite om in sommige gevallen deze observaties in meer detail te bekijken.

Onderstaande figuur toont de output voor het regressiemodel `fit3_expertise` (overclaiming):



2 De grootte van een effect

Het is niet ongebruikelijk bij lineaire regressie dat de focus ligt op het nagaan of de effecten van de predictoren op de uitkomst statistisch significant zijn (zie verder bij toetsing). Echter, statistische significantie is niet equivalent met praktische significantie.

Bij het rapporteren van resultaten is het noodzakelijk om naast de resultaten van een toets ook de grootte van effecten mee te geven. We beschouwen een aantal mogelijke r-maten (associatiematen) hiervoor alsook een betrouwbaarheidsinterval voor de regressiecoëfficiënten. In de praktijk kunnen beiden naast elkaar gerapporteerd worden, ze verschaffen andere informatie.

2.1 De determinatiecoëfficiënt R^2

De determinatiecoëfficiënt laat toe om te kwantificeren hoeveel van de variatie in de uitkomst verklaard wordt door het regressiemodel.

Als we denken aan een enkelvoudig regressiemodel waarbij Y op X geregresseerd wordt, weten we reeds dat de lineaire regressielijn de tendens van de lineaire relatie aangeeft. We kunnen echter niet verwachten dat de punten van de observaties op de scatterplot perfect op deze lijn liggen. Analoog zijn bij meervoudige lineaire regressie de predicties $\hat{Y}$ typisch niet perfect gelijk aan de geobserveerde uitkomsten. Dit impliceert dat de predictoren in een regressiemodel de uitkomst Y niet volledig verklaren.

Een maat voor de sterkte van de lineaire regressie:

- is de bekwaamheid van de onafhankelijke variabelen om de variantie van de afhankelijke variabele Y te verklaren.
- wordt bepaald door de grootte van de afwijkingen van de observaties tot de predicties.

  Bij enkelvoudige lineaire regressie wordt de lineaire relatie tussen predictor X en uitkomst Y sterker naarmate de observaties zich dichter bij de regressielijn bevinden.

De totale variatie in Y kan opgesplitst worden in 2 delen: het gedeelte dat verklaard wordt door de regressievergelijking en het gedeelte dat niet kan verklaard worden door de regressievergelijking. De variatie in Y wordt bepaald door de *totale kwadratensom (total sum of squares)* SST:

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Concreet kan men aantonen dat:

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

SSR is de **regressie kwadratensom (regression sum of squares)** en stelt het deel voor van de variatie in Y dat verklaard wordt door de lineaire regressie van Y op de predictoren[3]. SSE, de **fout kwadratensom** (zie ook hierboven), is het deel van de variatie in Y dat niet verklaard kan worden door de regressie, bij enkelvoudige lineaire regressie is dit de spreiding van de punten rond de regressielijn.

[3] De regressie kwadratensom stelt het deel van de variatie in Y voor dat verklaard wordt door het regressiemodel. SSR wordt daarom ook vaak genoteerd als SS_{Model} of SS_{Mod}.

De determinatiecoëfficiënt R^2 wordt gegeven door de verhouding van de verklaarde variatie op de totale variatie:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST-SSE}}{\text{SST}}$$

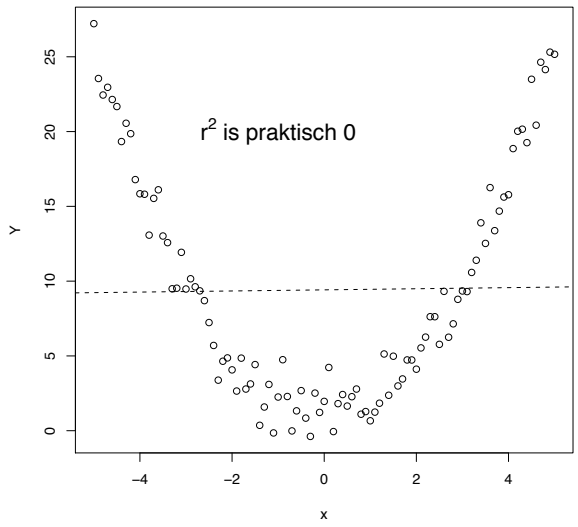
Interpretatie:

- R^2 is de proportie van de totale variatie in de uitkomst Y die verklaard wordt door de predictoren in de lineaire regressie van Y op de predictoren, $0 \leq R^2 \leq 1$.
- R^2 is het kwadraat van de steekproefcorrelatie $\text{Cor}(Y, \hat{Y})$, $-1 \leq R \leq 1$.
- Als $R^2 = 1 (= 100\%)$ dan is SST = SSR m.a.w. SSE = 0, dit wil zeggen dat alle residuen e_i gelijk zijn aan 0. De predicties op basis van het model komen dan perfect overeen met de observaties.
- Als $R^2 = 0 (= 0\%)$ dan is SSR = 0 m.a.w. geen enkel stukje van de variatie in Y wordt door de regressie verklaard. De predictoren zoals opgenomen in het lineair regressiemodel hebben dus geen enkele invloed bij het verklaren van de variatie van Y.

Bij enkelvoudige lineaire regressie met 1 predictor X wordt R^2 soms voorgesteld als r^2 aangezien in dat geval $r = \text{cor}(X, Y)$.

Let op:

- Een hoge R^2 betekent niet noodzakelijk dat nuttige predicties gemaakt kunnen worden.
  R^2 zegt niets over de precisie waarmee predicties gemaakt worden.
- Een hoge R^2 impliceert niet automatisch dat de geschatte regressievergelijking een goede fit is voor de data.
  Bij modellen met een hoge R^2 is het bvb. mogelijk dat modellen waarbij een niet-lineair verband tussen predictoren en uitkomst verondersteld wordt, een betere fit zijn voor de data.
- Een R^2 die dicht bij 0 ligt betekent niet automatisch dat er geen verband is tussen de uitkomst en de set van predictoren.
  R^2 geeft enkel de sterkte van het *lineaire* verband tussen Y en de (lineaire) combinatie van de predictoren weer.



R^2 stijgt naarmate men meer predictoren of onafhankelijke variabelen in het model toevoegt; bovendien is R^2 een vertekende schatter van de ware determinatiecoëfficiënt in de populatie. Om te corrigeren voor het aantal predictoren in het model en een meer onvertekende schatter te bekomen, maakt men in de praktijk vaak gebruik van de aangepaste (Engels: 'adjusted') meervoudige determinatiecoëfficiënt R_a^2:

$$R_a^2 = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$$

Merk op dat R_a^2 steeds kleiner is dan R^2 en niet steeds positief. Merk ook op dat R_a^2 niet dezelfde betekenis heeft als R^2, wees voorzichtig bij het interpreteren van deze statistiek.

In R wordt via de de `summary` van het gefitte model R^2 en R_a^2 van het model weergegeven.

Wanneer we in het voorbeeld rond overclaiming (zie sectie 1.2.1) de uitkomst overclaiming regresseren op gepercipieerde kennis, accuraatheid en FINRA-score, bekomen we:

```
> summary(fit3_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)
```

```
Residuals:
Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

We zien dat $R^2 = 0.7116$ (`Multiple R-squared`) wat betekent dat 71.16% van de variatie in overclaiming verklaard wordt door het model. $R^2_a = 0.7073$ (`Adjusted R-squared`).

We kunnen R^2 uit bovenstaande output als volgt bekomen:

```
> summary(fit3_expertise)$r.squared
[1] 0.7116225
```

2.2 Semi-partiële en partiële correlatie

R^2 geeft de proportie variatie in de uitkomst verklaard door het volledige model. Partiële maten laten toe om de bijdrage van de afzonderlijke predictoren te kwantificeren. We beschouwen hier 2 r-maten die de samenhang tussen de uitkomst en een individuele predictor geven, conditioneel op de overige predictoren (i.e. gegeven dat deze constant gehouden worden).

Laat $R^2_{Y|x_1,x_2,\ldots,x_p}$ de determinatiecoëfficiënt voorstellen van de regressie van Y op de p predictoren X_1, X_2, $\ldots$, X_p.

Analoog stelt $R^2_{Y|x_1,\ldots,x_{\ell-1},x_{\ell+1},\ldots,x_p}$ de determinatiecoëfficiënt voor van de regressie met dezelfde predictoren maar zonder X_ℓ in het model.

De **semi-partiële correlatie van** X_ℓ met Y is gelijk aan

$$sr_\ell = \sqrt{R^2_{Y|x_1,x_2,\ldots,x_p} - R^2_{Y|x_1,\ldots,x_{\ell-1},x_{\ell+1},\ldots,x_p}}.$$

sr_ℓ^2 is het gedeelte van de variatie van Y dat X_ℓ toelaat te verklaren bovenop het gedeelte dat door de andere $p-1$ predictoren voorspeld kan worden.

De **partiële correlatie** van X_ℓ met Y is gelijk aan

$$pr_\ell = \sqrt{\frac{sr_\ell^2}{1 - R^2_{Y|x_1,\ldots,x_{\ell-1},x_{\ell+1},\ldots,x_p}}}.$$

Dit geeft het verband weer tussen Y en X_ℓ nadat beide uitgezuiverd zijn voor het gedeelte dat met $X_1, \ldots, X_{\ell-1}, X_{\ell+1}, \ldots, X_p$ samenhangt.

Merk op dat een vierkantswortel zowel positief als negatief kan zijn; hier is het **teken** van sr_ℓ en pr_ℓ gelijk aan het teken van de geschatte regressiecoëfficiënt van predictor X_ℓ!

Aangezien β_ℓ de verwachte verandering in Y weergeeft indien X_ℓ met één eenheid stijgt terwijl de andere predictoren constant blijven, wordt hiernaar soms verwezen met de term **partiële regressiecoëfficiënt**.

Een andere (maar in dit geval equivalente) manier om sr_ℓ en pr_ℓ te bekomen is via kwadratensommen. Laat SSR en SSE respectievelijk de verklaarde kwadratensom en fout kwadratensom voorstellen van het model waarin ook predictor X_ℓ opgenomen is en SSR_ℓ de verklaarde kwadratensom van het model zonder X_ℓ. De kwadratensom die bij predictor X_ℓ hoort is dan gelijk aan $\text{SSR} - \text{SSR}_\ell$ en stellen we voor door SS_{X_ℓ}. Er geldt dat

$$\begin{aligned} sr_\ell^2 &= \frac{\text{SS}_{X_\ell}}{SST} \\ pr_\ell^2 &= \frac{\text{SS}_{X_\ell}}{\text{SS}_{X_\ell} + SSE} \end{aligned}$$

In R kan gebruik maakt worden van het package `lsr` om op een eenvoudige manier de semi-partiële en partiële correlatie te bekomen. We gebruiken hiervoor het commando `etaSquared`. We specifiëren hierbij dat we wensen gebruik te maken van Type III kwadratensommen (i.e. de kwadratensommen die we beschouwen in deze cursus).

Voor het voorbeeld rond overclaiming (zie sectie 1.2.1) krijgen we:

```
> etaSquared(fit3_expertise,type=3)
                              eta.sq    eta.sq.part
self_perceived_knowledge 0.200467989  0.41008454
accuracy                 0.439646003  0.60388983
FINRA_score              0.006681956  0.02264613
```

In de kolom `eta.sq` lezen we voor iedere predictor sr^2_ℓ af. Zo kunnen we afleiden dat de semi-partiële correlatie tussen FINRA-score en de uitkomst gelijk is aan $\sqrt{0.00668} = 0.0817$. In de kolom `eta.sq.part` lezen we pr^2_ℓ. De partiële correlatie tussen FINRA-score en de uitkomst is bijgevolg gelijk aan $\sqrt{0.0226} = 0.15$. Merk op dat de correlaties positief zijn aangezien het teken van de geschatte regressiecoëfficiënt voor de FINRA-score positief is.

De output kan ook verkregen worden samen met de kwadratensommen:

```
> etaSquared(fit3_expertise,type=3,anova=TRUE)
                             eta.sq eta.sq.part       SS    df         MS
self_perceived_knowledge 0.200467989  0.41008454 2.17303527   1 2.17303527
accuracy                 0.439646003  0.60388983 4.76567991   1 4.76567991
FINRA_score              0.006681956  0.02264613 0.07243114   1 0.07243114
Residuals                0.288377524          NA 3.12595808 198 0.01578767
                                F          p
self_perceived_knowledge 137.641316 0.00000000
accuracy                 301.860933 0.00000000
FINRA_score                4.587831 0.03342126
Residuals                       NA         NA
```

De waarden voor de verschillende kwadratensommen lezen we af onder `SS` (we concentreren ons hier enkel op dit stuk van de output). De fout kwadratensom kunnen we aflezen bij `Residuals`, SSE=3.12596. De totale kwadratensom SST kunnen we in R berekenen:

```
> sst<-sum((expertise$overclaiming_proportion-mean(expertise$overclaiming_proportion))^2)
> sst
[1] 10.83981
```

Voor FINRA-score lezen we af dat

- $sr^2_\ell = \frac{0.07243}{10.8398} = 0.006682$ en
- $pr^2_\ell = \frac{0.07243}{0.07243+3.1260} = 0.02265$.

Merk op dat bij `eta.sq` bij de `Residuals` SSE/SST weergegeven wordt, dit is dus $1 - R^2$ (de proportie van de variatie in de uitkomst die niet verklaard wordt door het model).

2.3 Betrouwbaarheidsintervallen voor $\boldsymbol{\beta}$

Om betrouwbaarheidsintervallen voor de regressiecoëfficiënten te kunnen opstellen wordt ook de assumptie van normaal verdeelde fouttermen gemaakt:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Een betrouwbaarheidsinterval voor een parameter heeft typisch de volgende vorm:

$$\text{schatter} \pm \text{ (kritische waarde) } \times \text{(Standaardfout van schatter)}$$

Onder de bovenstaande assumptie m.b.t. de fouttermen, kan aangetoond worden dat een $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval voor β_ℓ $(\ell = 1, \ldots, p)$ er als volgt uit ziet:

$$\left[B_\ell - |t_{n-(p+1);\alpha/2}| S\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{\ell,\ell}}, B_\ell + |t_{n-(p+1);\alpha/2}| S\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{\ell,\ell}}\right]$$

waarbij p het aantal predictoren in het model is en $(\mathbf{X}'\mathbf{X})_{\ell,\ell}$ het element op rij ℓ en kolom ℓ van de matrix $(\mathbf{X}'\mathbf{X})^{-1}$ voorstelt. $S\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{\ell,\ell}}$ is bijgevolg de standaardfout van B_ℓ.

Hoewel de steekproevenverdeling van B_ℓ een normale verdeling is, moeten we gebruik maken van de kristische waarde $t_{n-(p+1);\alpha/2}$ uit de t-verdeling met $n-(p+1)$ vrijheidsgraden aangezien we moeten corrigeren voor het feit dat σ^2 geschat wordt door S^2.

Dit betrouwbaarheidsinterval omvat β_ℓ met kans $1-\alpha$.

Deze intervallen geven ons naast de grootte van een effect informatie over de nauwkeurigheid van de schatter.

Wanneer we in het voorbeeld rond overclaiming (zie sectie 1.2.1), de uitkomst overclaiming regresseren op gepercipieerde kennis, accuraatheid en FINRA-score, kunnen we als volgt een 95% betrouwbaarheidsinterval voor het effect van elke predictor bekomen:

```
> confint(fit3_expertise)
                            2.5 %      97.5 %
(Intercept)             -0.019521646  0.13509525
self_perceived_knowledge  0.078257283  0.10988105
accuracy                -0.883251647 -0.70318621
FINRA_score              0.001457144  0.03528216
```

De geschatte regressiecoëfficiënt die het geschatte effect van gepercipieerde kennis, na correctie voor accuraatheid en FINRA score, op de gemiddelde uitkomst weergeeft, is gelijk aan 0.09407 (zie output hieronder). Een 95% betrouwbaarheidsinterval voor het effect is gelijk aan $[0.0783; 0.1099]$.

```
> fit3_expertise

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)

Coefficients:
(Intercept)  self_perceived_knowledge                  accuracy
0.05779                     0.09407                   -0.79322
FINRA_score
0.01837
```

Standaard geeft R via `confint` een 95% betrouwbaarheidsinterval, maar we kunnen eenvoudig ook andere niveaus opvragen, bvb. een 90% betrouwbaarheidsinterval:

```
> confint(fit3_expertise, level=0.90)
                                5 %        95 %
(Intercept)              -0.006999055  0.12257266
self_perceived_knowledge  0.080818526  0.10731981
accuracy                 -0.868667951 -0.71776990
FINRA_score               0.004196669  0.03254263
```

3 Regressie met nominale predictoren

Hoewel de term 'regressie' in de klassieke terminologie inhoudt dat alle predictoren van minstens intervalniveau zijn, is lineaire regressie technisch perfect mogelijk met nominale predictoren. Een variabele van nominaal niveau wordt een **factor** genoemd (dit is ook de terminologie gehanteerd door R).

Wanneer alle predictoren van nominaal niveau zijn, spreekt men van **variantie-analyse**. In sectie 8.2 gaan we daar dieper op in.

3.1 Lineaire regressie met hulpveranderlijken

In het algemeen geldt dat we een nominale predictor met I niveaus moeten hercoderen tot $I - 1$ nieuwe hulpveranderlijken die we vervolgens in het regressiemodel kunnen stoppen.

We kunnen hierbij een onderscheid maken tussen **dummy-codering** en **effect-codering**. We bekijken dit aan de hand van een voorbeeld.

Veronderstel dat we het effect van het type onderwijs op de uiteindelijke studieresultaten wensen te modelleren waarbij er in totaal 4 types van onderwijs beschouwd worden. De variabele die het type onderwijs weergeeft is nominaal.

- Bij **dummy-codering** kiest men één van de I niveaus als referentieniveau en worden de andere niveaus via een 0-1 variabele gecodeerd.

  In het geval van het voorbeeld betekent dit dat we 3 hulpveranderlijken X_1, X_2 en X_3 moeten aanmaken. Wanneer we type 4 als referentieniveau beschouwen, dan bekomen we de volgende codering:

Type onderwijs	X_1	X_2	X_3
Type 1	1	0	0
Type 2	0	1	0
Type 3	0	0	1
Type 4	0	0	0

  Dit betekent concreet dat voor een individu i

  - die type 1 van onderwijs volgt, geldt: $x_{i1} = 1$, $x_{i2} = 0$, $x_{i3} = 0$.
  - die type 2 van onderwijs volgt, geldt: $x_{i1} = 0$, $x_{i2} = 1$, $x_{i3} = 0$.
  - die type 3 van onderwijs volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 1$.
  - die type 4 van onderwijs volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 0$.

- **Effect-codering** is analoog aan dummy-codering behalve dat de referentiegroep steeds met -1 gecodeerd wordt i.p.v. met 0. Voor het voorbeeld bekomen we:

Type onderwijs	X_1	X_2	X_3
Type 1	1	0	0
Type 2	0	1	0
Type 3	0	0	1
Type 4	-1	-1	-1

  Dit betekent dat de codering hetzelfde is als de dummy-codering voor individuen die type 1, 2 of 3 van het onderwijs volgen maar voor een individu i die type 4 van onderwijs volgt, geldt: $x_{i1} = -1$, $x_{i2} = -1$, $x_{i3} = -1$.

Het effect van het type onderwijs op de verwachtingswaarde van de studieresultaten Y kunnen we als volgt modelleren:

$$E(Y_i|x_{i1}, x_{i2}, x_{i3}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Naargelang het coderingsschema dat gehanteerd wordt, hebben de regressieparameters een andere betekenis.

- Dummy-codering
  - $E(Y_i|\text{Type 4}) = E(Y_i|x_{i1} = 0, x_{i2} = 0, x_{i3} = 0) = \beta_0$. β_0 stelt dus het verwachte studieresultaat voor bij type 4 van onderwijs.
  - $E(Y_i|\text{Type 1}) = E(Y_i|x_{i1} = 1, x_{i2} = 0, x_{i3} = 0) = \beta_0 + \beta_1$. β_1 stelt dus het verschil voor van het verwachte studieresultaat bij type 1 en het verwachte studieresultaat bij type 4 van onderwijs.

    Analoog stellen β_2 en β_3 het verschil in verwachte studieresultaat tussen type 2 en type 4 en tussen type 3 en type 4.
- Effect-codering
  - In dit geval kan aangetoond worden dat β_0 het marginale gemiddelde van het studieresultaat voorstelt, i.e. het gemiddelde van de gemiddelde studieresultaten over de verschillende onderwijstypes heen:
    $(E(Y_i|\text{Type 1}) + E(Y_i|\text{Type 2}) + E(Y_i|\text{Type 3}) + E(Y_i|\text{Type 4}))/4$.
  - β_ℓ ($\ell = 1, 2, 3$) drukt dan het verschil uit tussen het verwachte studieresultaat in onderwijstype ℓ en het marginale gemiddelde.
  - Het verwachte studieresultaat in type 4 van het onderwijs is

    $$E(Y_i|\text{Type 4}) = E(Y_i|x_{i1} = -1, x_{i2} = -1, x_{i3} = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3.$$

    Dit betekent dat het verschil tussen het verwachte resultaat in onderwijstype 4 en het marginale gemiddelde gelijk is aan $\beta_4 = -\beta_1 - \beta_2 - \beta_3$. Bijgevolg geldt dat $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$.

3.2 Voorbeeld: pijneducatie

Wanneer we via lineaire regressie het effect van `conditie` (nominaal, 3 niveaus) op de uitkomst `Buig` (verschilscore in voorover buigen) modelleren (zie sectie 1.2.2; we laten de overige variabelen hier buiten beschouwing), vergelijken we de gemiddelde uitkomst over de 3 condities.

Na het inlezen van de data, zien we dat de variabele `conditie` in R als een factor met 3 niveaus gedefinieerd is.

```
> class(pijneducatie$Conditie)
[1] "factor"
> levels(pijneducatie$Conditie)
[1] "Algemene pijneducatie" "Baseline"              "Rugpijneducatie"
```

Via `contrasts()` kunnen we opvragen welke restrictieschema gehanteerd wordt.

```
> contrasts(pijneducatie$Conditie)
                          Baseline Rugpijneducatie
Algemene pijneducatie        0               0
Baseline                     1               0
Rugpijneducatie              0               1
```

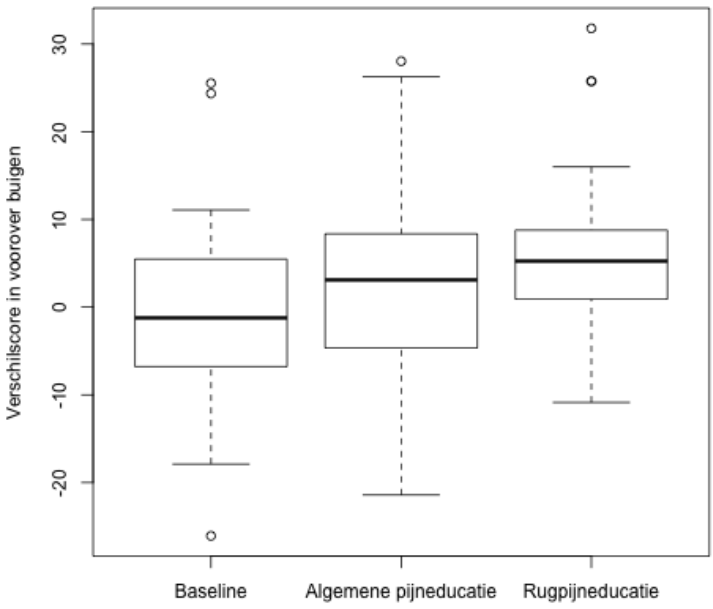
We zien dat dit standaard de dummy-codering is. R zal als referentieniveau altijd het eerste niveau kiezen. Echter, de niveaus werden in dit geval in R alfabetisch gerangschikt wat betekent dat we `conditie=Baseline` en `conditie=Rugpijneducatie` vergelijken met `conditie=Algemene pijneducatie`. Gezien we hier 2 'behandelingen' hebben en 1 baseline, is het logisch om de baseline als referentieniveau te kiezen.

```
pijneducatie$Conditie<-factor(pijneducatie$Conditie,levels=c("Baseline","Algemene
                                    pijneducatie","Rugpijneducatie"))
```

Op die manier geven we aan in R dat het eerste niveau (het referentieniveau) `conditie=Baseline` is.

```
> contrasts(pijneducatie$Conditie)
                          Algemene pijneducatie Rugpijneducatie
Baseline                                    0               0
Algemene pijneducatie                       1               0
Rugpijneducatie                             0               1
```

De verdeling van de uitkomst per groep kan grafisch als volgt voorgesteld worden:



De steekproefgemiddeldes per groep zijn:

```
Baseline          Algemene pijneducatie       Rugpijneducatie
-0.765494               2.608108              5.063635
```

We stellen vast dat de geobserveerde gemiddelde uitkomst het laagst is in de baseline conditie en het hoogst in de conditie met rugpijneducatie.

Het marginale (ongewogen) steekproefgemiddelde is gelijk aan $(-0.765494 + 2.608108 + 5.063635)/3 = 2.30208$.

3.2.1 Dummy-codering voor conditie

We weten reeds dat `conditie` in R als een factor gedefinieerd is en dat standaard de dummy-codering gebruikt zal worden. We hoeven dus verder niets te specifiëren wanneer we `Buig` op `Conditie` regresseren.

```
> fit_pijneducatie_dummy<-lm(Buig~Conditie,data=pijneducatie)
> summary(fit_pijneducatie_dummy)

Call:
lm(formula = Buig ~ Conditie, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.7655     1.5885  -0.482   0.6308
ConditieAlgemene pijneducatie  3.3736     2.2464   1.502   0.1358
ConditieRugpijneducatie        5.8291     2.2327   2.611   0.0102 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

We concentreren ons hier op de geschatte regressiecoëfficiënten. We lezen af dat de geschatte gemiddelde uitkomst in de baseline conditie gelijk is aan -0.7655. In de conditie met algemene pijneducatie is dit $-0.7655 + 3.3736 \times 1 + 5.8291 \times 0 = 2.60814$; in de conditie met rugpijneducatie is dit $-0.7655 + 3.3736 \times 0 + 5.8291 \times 1 = 5.0636$. Dit komt overeen met de geobserveerde steekproefgemiddeldes.

Omgekeerd kunnen de geschatte regressiecoëfficiënten afgeleid worden uit de geobserveerde steekproefgemiddeldes.

Het is belangrijk om zeker te zijn dat de nominale variabelen als factoren gedefinieerd zijn en dat men goed weet welk coderingsschema gebruikt wordt. In R staat `contr.treatment` voor dummy-codering. Je kan dit ook zelf instellen voor een factor (dit kan nodig zijn als het coderingsschema niet goed ingesteld staat).

```
> contrasts(pijneducatie$Conditie)<-contr.treatment
> contrasts(pijneducatie$Conditie)
                     2 3
Baseline             0 0
```

```
Algemene pijneducatie  1 0
Rugpijneducatie        0 1
```

Merk op dat bij de naamgeving van de hulpveranderlijken de oorspronkelijke labels in dit geval niet overgenomen worden. Uit het coderingsschema kunnen we afleiden dat `Conditie=2` overeen komt met de algemene pijneducatie en `Conditie=3` met de rugpijneducatie.

```
> fit2_pijneducatie_dummy<-lm(Buig~Conditie,data=pijneducatie)
> summary(fit2_pijneducatie_dummy)

Call:
lm(formula = Buig ~ Conditie, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7655     1.5885  -0.482   0.6308
Conditie2     3.3736     2.2464   1.502   0.1358
Conditie3     5.8291     2.2327   2.611   0.0102 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

Het is, zeker wanneer er meerdere nominale predictoren zijn, handig en veilig om coderingsschema's rechtstreeks via de functie `lm` mee te geven.

```
lm(Buig~Conditie,data=pijneducatie,contrasts=list(Conditie=contr.treatment))
```

3.2.2 Effect-codering voor conditie

Wanneer we effect-codering willen hanteren, doen we dit via `contr.sum` in R.

```
> contrasts(pijneducatie$Conditie)<-contr.sum
> contrasts(pijneducatie$Conditie)
                       [,1] [,2]
Baseline                 1    0
Algemene pijneducatie    0    1
Rugpijneducatie         -1   -1
```

We zien dat beide hulpveranderlijken op -1 gezet worden in de rugpijneducatie.

We kunnen het coderingsschema ook meegeven in de functie `lm`.

```
> fit_pijneducatie_effect<-lm(Buig~Conditie,data=pijneducatie,
                              contrasts=list(Conditie=contr.sum))
> summary(fit_pijneducatie_effect)

Call:
lm(formula = Buig ~ Conditie, data = pijneducatie,
                 contrasts = list(Conditie = contr.sum))

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3021     0.9134   2.520   0.0131 *
Conditie1    -3.0676     1.2943  -2.370   0.0194 *
Conditie2     0.3060     1.2943   0.236   0.8135
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

We zien dat de schatting voor het intercept inderdaad gelijk is aan het marginale (ongewogen) steekproefgemiddelde van de gemiddelde uitkomst binnen de 3 condities. Verder leiden we af dat het geschatte gemiddelde binnen de conditie met algemene pijneducatie (`conditie=2`) gelijk is aan $2.3021 - 3.0676 \times 0 + 0.3060 \times 1 = 2.6081$ en dat het geschatte gemiddelde in de

baseline conditie (`conditie=1`) gelijk is aan $2.3021 - 3.0676 \times 1 + 0.3060 \times 0 = -0.7655$. Dit komt overeen met de geobserveerde steekproefgemiddeldes.

Voor de conditie met rugpijneducatie kunnen we afleiden dat het geschatte gemiddelde gelijk is aan $2.3021 - 3.0676 \times (-1) + 0.3060 \times (-1) = 5.0637$.

4 Toetsing

In dit stuk gaan we dieper in op hypothesetoetsen voor de parameters in een regressiemodel. Hierbij wordt ook de assumptie van normaal verdeelde fouttermen gemaakt:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

4.1 Modelvergelijkingen

Wanneer we meerdere (p) predictoren in het regressiemodel hebben, kunnen we ons afvragen of het nodig is die allemaal in het model op te nemen.

Om dit te toetsen kunnen we het model met p predictoren vergelijken met een model met k predictoren die een subset zijn van de p predictoren.

Dit is een speciaal geval van een algemene toets die nagaat of twee lineaire modellen, met het ene model genest in het andere, significant van elkaar verschillen wat betreft de mogelijkheid om de uitkomst Y te voorspellen.

Een lineair model B is genest in model A wanneer bij model B bijkomende lineaire restricties m.b.t. de te schatten parameters opgelegd worden.

Als we de 2 modellen A (zonder restricties) en B (met restricties) met elkaar wensen te vergelijken, kunnen we gebruik maken van deze toetsingsgrootheid:

$$F = \frac{(\text{SSE}_B - \text{SSE}_A)/(\text{df}_B - \text{df}_A)}{\text{SSE}_A/\text{df}_A} \tag{2}$$

SSE_A stelt de fout kwadratensom van model A voor en SSE_B de fout kwadratensom van model B. df_A en df_B stellen de overeenkomstige vrijheidsgraden in respectievelijk model A en model B voor.

Met p predictoren in het model dienen op basis van n observaties $p + 1$ coëfficiënten geschat te worden (vergeet het intercept niet). Dit betekent dat in dit geval $\text{df}_A = n - (p + 1)$.

Wanneer model B slechts een subset van k predictoren bevat, is $\text{df}_B = n - (k + 1)$.

Er kan aangetoond worden dat onder H_0 (beide modellen zijn niet verschillend van elkaar), de toetsingsgrootheid (2) F-verdeeld is met $\text{df}_B - \text{df}_A$ vrijheidsgraden voor de teller en df_A vrijheidsgraden voor de noemer ($F \sim F(\text{df}_B - \text{df}_A, \text{df}_A)$). Dit is de nuldistributie van F in (2), dit betekent dat we de bijhorende p-waarde van de toets kunnen berekenen en beslissen om H_0 al dan niet te verwerpen.

Dit is een algemene beschrijving voor toetsing aan de hand van modelvergelijkingen. In de volgende secties worden enkele concrete gevallen in meer detail besproken.

4.2 Toets voor alle predictoren

We stellen ons hier de vraag: is er tenminste 1 predictor nuttig in het voorspellen van de uitkomst?

We vergelijken het model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n \tag{3}$$

met het model zonder predictoren:

$$Y_i = \beta_0 + \varepsilon_i.$$

Dit model noemen we het **nulmodel**. Het aantal vrijheidsgraden dat geassocieerd is met het nulmodel is $n - 1$ (er dient enkel een intercept geschat te worden). Voor het nulmodel geldt dat $\hat{Y}_i = B_0 = \bar{Y}$ $(i = 1, \ldots, n)$.

Of nog: we toetsen de volgende nulhypothese

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

tegenover de alternatieve hypothese H_1 die stelt dat tenminste 1 regressiecoëfficiënt verschillend is van 0. Het verwerpen van H_0 betekent dus dat we verder moeten nagaan voor welke predictoren de regressiecoëfficiënt significant verschillend is van 0.

Er geldt dat de fout kwadratensom voor het nulmodel gelijk is aan de totale kwadratensom SST van het volledig model in (3). Het aantal vrijheidsgraden dat overeenkomt met SST is $n - 1$.

Bijgevolg ziet de toetsingsgrootheid in (2) er in dit geval als volgt uit:

$$F = \frac{(\text{SST} - \text{SSE})/(\text{df}_0 - \text{df}_A)}{\text{SSE}/\text{df}_A} = \frac{\text{SSR}/(\text{df}_0 - \text{df}_A)}{\text{SSE}/\text{df}_A}$$

met df_0 het aantal vrijheidsgraden van het nulmodel $(= n - 1)$ en df_A het aantal vrijheidsgraden van het volledige model in (3) $(= n - (p + 1))$. Bijgevolg is $\text{df}_0 - \text{df}_A = p$. Dit is

het aantal vrijheidsgraden dat hoort bij de verklaarde kwadratensom SSR van een model en is gelijk aan het aantal parameters van het volledige model, intercept niet meegerekend.

Beschouw het voorbeeld rond overclaiming (zie sectie 1.2.1). We voorspellen overclaiming op basis van gepercipieerde kennis, accuraatheid en FINRA-score.

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
             +FINRA_score,data=expertise)
> summary(fit3_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge + accuracy
         + FINRA_score, data = expertise)

Residuals:
    Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

In deze output kunnen we op de onderste lijn het resultaat van de F-toets voor alle predictoren aflezen. Hier is de geobserveerde toetsingsgrootheid f^* gelijk aan 162.9, bij de bijhorende p-waarde duidt `< 2.2e-16` aan dat deze p-waarde heel klein is. We hebben sterke evidentie tegen de nulhypothese wat impliceert dat we evidentie hebben voor het feit dat minstens 1 predictor een invloed heeft op de uitkomst. In dit voorbeeld is het aantal observaties $n = 202$. De toetsingsgrootheid volgt onder de nulhypothese een F-verdeling met 3 vrijheidsgraden (aantal parameters dat getoetst wordt) voor de teller en 198 (i.e. 202-(3+1)) voor de noemer.

In dit stukje van de output kunnen we de kwadratensommen zelf niet aflezen. Deze informatie

krijgen we wel als we zelf de modelvergelijkingstoets uitvoeren via het commando `anova`.

Het nulmodel kunnen we als volgt definiëren:

```
fit0_expertise<-lm(overclaiming_proportion~1,data=expertise)
```

Het resultaat van de modelvergelijking die het effect voor alle predictoren toetst is als volgt:

```
> anova(fit0_expertise,fit3_expertise)
Analysis of Variance Table

Model 1: overclaiming_proportion ~ 1
Model 2: overclaiming_proportion ~ self_perceived_knowledge + accuracy +
FINRA_score
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    201 10.840
2    198  3.126  3    7.7139 162.87 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Bovenaan in de output staat weergegeven welke 2 modellen met elkaar vergeleken worden.

In de kolom `RSS` lezen we de fout kwadratensommen af. Voor het nulmodel is deze som gelijk aan 10.840. Dit is ook de SST van het volledige model. Het overeenkomstige aantal vrijheidsgraden kan afgelezen worden in de kolom `Res.Df` en is gelijk aan $n - 1 = 202 - 1 = 201$.

Voor het volledige model is de fout kwadratensom SSE gelijk aan 3.126, het overeenkomstig aantal vrijheidsgraden is gelijk aan $n - (3 + 1) = 202 - 4 = 198$. We lezen op de onderste lijn onder `Sum of Sq` verder af dat SST-SSE=7.7139=SSR en dat het overeenkomstig aantal vrijheidsgraden gelijk is aan 3, i.e. het verschil in aantal vrijheidsgraden tussen het volledige model en het nulmodel of nog, het aantal parameters voor de predictoren in het volledige model, intercept niet inbegrepen.

We zien dat de geobserveerde toetsingsgrootheid inderdaad overeenkomt met wat we voordien bekwamen, namelijk $f^* = (7.7139/3)/(3.126/198) = 162.87$.

4.3 Toets voor een subset van predictoren

De toets voor alle predictoren uit de vorige sectie is een speciaal geval van de modelvergelijkingstoets waarbij een set van predictoren getoetst wordt. Wanneer we in het

voorbeeld rond overclaiming (zie sectie 1.2.1) wensen te toetsen of het model waar naast de gepercipieerde kennis ook accuraatheid en FINRA-score als predictoren opgenomen zijn de variatie in overclaiming beter verklaart, vergelijken we het model met de 3 predictoren met het model met enkel gepercipieerde kennis als predictor.

Als β_2 en β_3 de regressiecoëfficiënten voor respectievelijk accuraatheid en FINRA-score voorstellen, dan toetsen we $H_0 : \beta_2 = \beta_3 = 0$ versus de alternatieve hypothese die stelt dat minstens 1 van beide parameters niet 0 is. De specifiëring in R van de 2 modellen gebeurt als volgt:

```
fit1_expertise<-lm(overclaiming_proportion~self_perceived_knowledge,data=expertise)
fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                   +FINRA_score,data=expertise)
```

Wanneer we beide modellen met elkaar vergelijken, krijgen we:

```
> anova(fit1_expertise,fit3_expertise)
Analysis of Variance Table

Model 1: overclaiming_proportion ~ self_perceived_knowledge
Model 2: overclaiming_proportion ~ self_perceived_knowledge + accuracy +
FINRA_score
   Res.Df  RSS Df   Sum of Sq    F    Pr(>F)
1    200 8.3303
2    198 3.1260  2    5.2044 164.82 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Model 1 is hier genest in model 2, i.e. in de notatie in (2) is model 1 dus gelijk aan model B terwijl model 2 gelijk is aan model A.

Het aantal vrijheidsgraden (`Res.Df`) voor model 1 is 200 (df_B), i.e. 202-(1+1) terwijl dit 198 (df_A) voor model 2 is.

In kolom `RSS` zien we dat $\text{SSE}_B = 8.3303$ en $\text{SSE}_A = 3.1260$.

Op de onderste lijn zien we verder dat er $200 - 198 = 2 = \text{df}_B - \text{df}_A$ parameters getoetst worden.

De geobserveerde toetsingsgrootheid is $f^* = (5.2044/2)/(3.1260/198) = 164.82$. Onder H_0 is deze toetsingsgrootheid F-verdeeld met 2 en 198 vrijheidsgraden. De bijhorende p-waarde

(`Pr(>F)`) is heel klein waardoor we evidentie hebben tegen de nulhypothese en aannemen dat model 2 de variatie in overclaiming beter verklaart dan model 1.

4.4 Toets voor 1 predictor

We stellen ons hier de vraag: kan 1 welbepaalde predictor uit het model weggelaten worden?

Hiervoor kunnen we opnieuw een modelvergelijkingstoets uitvoeren: het volledige model wordt vergeleken met het model zonder de predictor.

Als deze predictor van intervalniveau is, komt dit overeen met het toetsen van 1 enkele parameter.

Als de predictor van nominaal niveau is, zal het aantal parameters dat getoetst wordt overeenkomen met het aantal hulpveranderlijken voor deze predictor. M.a.w. als we willen nagaan of een nominale predictor een invloed heeft op de uitkomst, vergelijken we de modellen met en zonder de hulpveranderlijken die coderen voor de nominale predictor. Het resultaat van deze toets zal onafhankelijk zijn van het coderingsschema (dummy- of effect-codering) dat gebruikt wordt voor de nominale predictor.

4.4.1 Predictor van intervalniveau

Beschouw het volgende model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_\ell x_{i\ell} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

Wanneer we wensen te toetsen of predictor ℓ uit het model weggelaten kan worden, toetsen we $H_0 : \beta_\ell = 0$ tegenover $H_1 : \beta_\ell \neq 0$.

In het geval dat er slechts 1 parameter getoetst wordt, is de F-toets uit de modelvergelijking equivalent aan een t-toets. Hiervoor kunnen we gebruik maken van de volgende toetsingsgrootheid:

$$T = \frac{B_\ell}{S_{B_\ell}}$$

waarbij S_{B_ℓ} de standaardfout van B_ℓ voorstelt. Er kan aangetoond worden dat onder H_0, $T \sim t_{n-(p+1)}$ of nog: dat T onder H_0 een t-verdeling volgt met $n - (p + 1)$ vrijheidsgraden.

Onder H_0 is de absolute waarde van de toetsingsgrootheid klein (dicht bij 0). Als H_1 geldt, zal de absolute waarde groot zijn.

Op basis van de geobserveerde toetsingsgrootheid $t^* = b_\ell / s_{B_\ell}$ kunnen we de p-waarde berekenen:

$$2 \times P\left(T \geq \left|\frac{b_\ell}{s_{B_\ell}}\right|\right) \text{ met } T \sim t_{n-(p+1)}.$$

Via de output van `lm` in R krijgen we standaard het resultaat van de t-toets voor alle parameters in het model (ook voor het intercept, maar hieraan wordt in de praktijk doorgaans geen aandacht aan besteed).

Herneem het voorbeeld rond overclaiming (zie sectie 1.2.1).

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
+FINRA_score,data=expertise)
> summary(fit3_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge + accuracy
+ FINRA_score, data = expertise)

Residuals:
Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

Voor iedere predictor lezen we in de kolom `t value` de geobserveerde toetsingsgrootheid af en in de kolom `Pr(>|t|)` de overeenkomstige p-waarde.

Zo zien we bvb. voor FINRA-score dat de geobserveerde toetsingsgrootheid $t^* = 0.018370/0.008576 = 2.142$. De bijhorende p-waarde is 0.0334, dit is kleiner dan 5% wat

betekent dat we de nulhypothese dat FINRA-score geen effect heeft op overclaiming kunnen verwerpen op het 5% significantieniveau.

Via `anova` waarbij we de modelvergelijking zelf definiëren komen we tot hetzelfde resultaat:

```
fit2_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy,
                   data=expertise)
fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                  +FINRA_score,data=expertise)
> anova(fit2_expertise,fit3_expertise)
Analysis of Variance Table

Model 1: overclaiming_proportion ~ self_perceived_knowledge + accuracy
Model 2: overclaiming_proportion ~ self_perceived_knowledge + accuracy +
FINRA_score
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    199 3.1984
2    198 3.1260  1  0.072431 4.5878 0.03342 *
```

Er geldt dat $f^* = t^{*2} = (2.142)^2 = 4.59$.

4.4.2 Predictor van nominaal niveau

Om het effect van een nominale predictor met I niveaus te toetsen, voeren we een modelvergelijkingstoets uit met als nulhypothese dat de regressiecoëfficiënten die horen bij de $(I-1)$ hulpveranderlijken allen 0 zijn.

Dit komt neer op toetsen of de gemiddelde uitkomst gelijk is over de verschillende niveaus (conditioneel op de overige predictoren in het model).

Merk op dat wanneer $I = 2$ er slechts 1 parameter getoetst moet worden. In dat geval is de F-toets ook equivalent aan de t-toets zoals in voorgaande sectie.

We hernemen het voorbeeld van de pijneducatie (sectie 1.2.2) waarbij we het effect van `conditie` (3 niveaus) op de uitkomst `buig` regresseren. De output bij dummy-codering is als volgt:

```
> fit_pijneducatie_dummy<-lm(Buig~Conditie,data=pijneducatie)
> summary(fit_pijneducatie_dummy)
```

```
Call:
lm(formula = Buig ~ Conditie, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.7655     1.5885  -0.482   0.6308
ConditieAlgemene pijneducatie   3.3736     2.2464   1.502   0.1358
ConditieRugpijneducatie         5.8291     2.2327   2.611   0.0102 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

Aangezien `conditie` de enige predictor in het model is, lezen we hier onmiddellijk het resultaat voor de modelvergelijkingstoets voor het effect van `conditie` af (door vergelijking van het model met nulmodel): de geobserveerde toetsingsgrootheid f^* is gelijk aan 3.432 met een bijhorende p-waarde gelijk aan 0.036. We kunnen bijgevolg de nulhypothese die stelt dat de gemiddelde verschilscore in voorover buigen gelijk is in de 3 condities, verwerpen.

Dit resultaat kunnen we ook als volgt bekomen:

```
> fit0_pijneducatie_dummy<-lm(Buig~1,data=pijneducatie)
> anova(fit0_pijneducatie_dummy,fit_pijneducatie_dummy)
Analysis of Variance Table

Model 1: Buig ~ 1
Model 2: Buig ~ Conditie
Res.Df   RSS     Df Sum of Sq   F  Pr(>F)
1    120 12602
2    118 11910  2    692.76 3.4319 0.03559 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Kijk zelf na dat het resultaat hetzelfde blijft wanneer effect-codering gebruikt wordt voor

conditie.

4.4.3 Algemene strategie: Anova-tabel in R

Van zodra er meerdere predictoren (waaronder nominale predictoren) in een regressiemodel opgenomen zijn, is het praktischer om de resultaten voor de toetsen van alle predictoren afzonderlijk via 1 enkel commando te verkrijgen. Het commando `Anova` (let op de hoofdletter!) uit het R-package `car` geeft deze resultaten in 1 anova-tabel weer. De reden dat we werken met dit commando is dat het toelaat verschillende types toetsen uit voeren waaronder deze die overeenkomen met wat andere software zoals SPSS standaard zou weergeven.

Wij maken altijd gebruik van Type III-toetsen bij het toetsen van de predictoren afzonderlijk. Het onderscheid tussen de verschillende types toetsen speelt voornamelijk een rol wanneer ook interacties in het model opgenomen zijn. We komen er op terug in sectie 5 over interacties.

Een belangrijke opmerking bij het gebruik van de Type III-toetsen bij Anova is dat we voor het toetsen zelf eerst moeten overgaan op effect-codering van de nominale variabelen in het model. Onthoud dat een model met effect-codering in essentie hetzelfde is als een model met dummy-codering; enkel de interpretatie van de parameters wijzigt. Het betreft dus louter een technisch aspect om correcte en zinvolle resultaten bij het toetsen te krijgen. Dit is opnieuw enkel van belang als er ook interacties met nominale predictoren opgenomen zijn (zie verder).

We hernemen het voorbeeld rond pijneducatie (sectie 1.2.2) . In deze studie zijn de participanten niet at random toegewezen aan de verschillende condities. Daarom is het zinvol om te corrigeren voor het effect van leeftijd en de graad van depressie. Leeftijd (`Leeft`) en Depressiescore (`Dep`) mogen als van intervalniveau verondersteld worden.

```
> fit3_pijneducatie_dummy<-lm(Buig~Conditie+Leeft+Dep,data=pijneducatie)
> summary(fit3_pijneducatie_dummy)

Call:
lm(formula = Buig ~ Conditie + Leeft + Dep, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-20.2287  -4.9537  -0.5254   5.1798  18.7101

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 16.42433    4.41878   3.717 0.000312 ***
```

```
ConditieAlgemene pijneducatie  3.73261    1.72868   2.159 0.032891 *
ConditieRugpijneducatie        4.78874    1.72026   2.784 0.006276 **
Leeft                         -0.10238    0.10749  -0.952 0.342840
Dep                           -0.65121    0.07187  -9.061 3.75e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.723 on 116 degrees of freedom
Multiple R-squared:  0.451,Adjusted R-squared:  0.4321
F-statistic: 23.82 on 4 and 116 DF,  p-value: 2.133e-14
```

Uit bovenstaande output kunnen we aflezen dat leeftijd in het model geen statistisch significante invloed heeft op de uitkomst ($p = 0.34$) en depressiescore wel ($p < 0.001$), maar voor conditie kunnen we dit niet rechtstreeks aflezen. Wat we aflezen is het resultaat van de toetsen waarbij de conditie met algemene pijneducatie en rugpijneducatie afzonderlijk met de baseline conditie vergeleken worden.

We vragen nu de bijhorende anova-tabel op om de resultaten voor alle predictoren afzonderlijk te zien. Hiervoor herdefiniëren we eerst het model aan de hand van effect-codering.

```
> library(car)
> fit3_pijneducatie_test<-lm(Buig~Conditie+Leeft+Dep,data=pijneducatie,
                                  contrasts=list(Conditie=contr.sum))
> Anova(fit3_pijneducatie_test,type=3)
Anova Table (Type III tests)

Response: Buig
            Sum Sq  Df F value    Pr(>F)
(Intercept) 1208.3   1 20.2587 1.618e-05 ***
Conditie     509.2   2  4.2685   0.01626 *
Leeft         54.1   1  0.9072   0.34284
Dep         4897.2   1 82.1057 3.746e-15 ***
Residuals   6918.8 116
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In deze output lezen we voor iedere predictor het resultaat af van de modelvergelijkingstoets waarbij de modellen met en zonder de predictor vergeleken worden (de geobserveerde toetsingsgrootheid in `F value` en p-waarde in `Pr(>F)`). We krijgen de overeenkomstige

kwadratensom voor deze toetsen (`Sum Sq`) en het aantal vrijheidsgraden (`Df`), i.e. het aantal parameters dat getoetst wordt (1 voor leeftijd en 1 voor depressiescore maar 2 voor conditie).

Voor `Leeft` en `Dep` bekomen we (uiteraard) dezelfde resultaten voor de F-toets als voor de t-toets. We lezen verder af dat het effect van conditie na correctie voor leeftijd en depressiescore nog steeds statistisch significant is op het 5% significantieniveau ($p = 0.016$).

Verifieer zelf dat via `Anova` dezelfde resultaten bekomen worden voor de analyses uitgevoerd in sectie 4.4.1 en 4.4.2!

5 Interactie (moderatie)

5.1 Wat is interactie?

Bij een regressiemodel met 2 predictoren X_1 en X_2 waarvoor

$$E(Y_i|x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ \ (i = 1, \ldots, n)$$

zijn de effecten van de predictoren additief: het effect van de ene predictor hangt niet af van het niveau van de andere, hun gezamenlijk effect kan gemodelleerd worden als een som.

Wanneer er een **interactie** is tussen beide predictoren betekent dit dat het effect van een combinatie van de 2 predictoren groter of kleiner is dan de som van de afzonderlijke effecten. Het effect van de ene predictor is nu anders voor elk niveau van de andere predictor. We hebben dan:

$$E(Y_i|x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} \ \ (i = 1, \ldots, n)$$

De interactieterm is gelijk aan het product van beide predictoren. β_3 is het interactie-effect.

De termen **effect-modificatie** en **moderatie** worden ook vaak gebruikt om te verwijzen naar een interactie. X_1 modificeert het effect van X_2 op Y (en vice versa). X_1 is een moderator voor het effect van X_2 op Y (en vice versa).

De regressiecoëfficiënten β_1 en β_2 hebben nu een andere interpretatie dan voorheen. Wanneer X_1 1 eenheid stijgt terwijl $X_2 = x_2$ constant blijft, neemt de verwachte uitkomst toe met $\beta_1 + \beta_3 x_2$.

Analoog: wanneer X_2 1 eenheid stijgt terwijl $X_1 = x_1$ constant blijft, neemt de verwachte uitkomst toe met $\beta_2 + \beta_3 x_1$.

Voorbeeld

Beschouwen we het verwachte aantal telefonisch verkochte abonnementen per dag in functie van de ervaring van de verkoper (X_1, gemeten op een schaal van 1 tot 7) en het instituut waarbij de verkoper een opleiding gekregen heeft (X_2, Instituut 1 en Instituut 2).

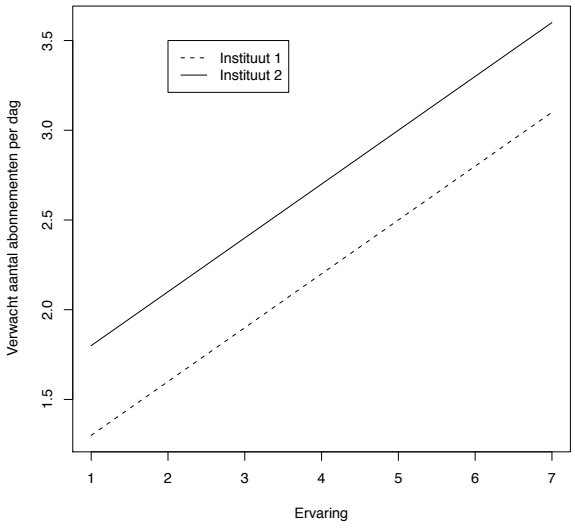
- Veronderstel dat $E(Y_i|x_{i1}, x_{i2}) = 1 + 0.3x_{i1} + 0.5x_{i2}$ waarbij $x_{i2} = 0$ voor instituut 1 en $x_{i2} = 1$ voor instituut 2 (dummy-codering). In dit geval is er geen interactie tussen de ervaring en het instituut waar de verkoper de opleiding gekregen heeft. We vinden immers dat het verwacht aantal verkochte abonnementen per dag in functie van de ervaring van een verkoper die een opleiding in instituut 1 kreeg als volgt is:

  $$E(Y_i|x_{i1}, x_{i2} = 0) = 1 + 0.3x_{i1}$$

  terwijl dit voor een verkoper die zijn opleiding in instituut 2 kreeg als volgt is:

  $$E(Y_i|x_{i1}, x_{i2} = 1) = 1 + 0.3x_{i1} + 0.5 = 1.5 + 0.3x_{i1}.$$

  Het effect van de ervaring blijft gelijk (0.3), alleen het intercept verandert. Grafisch kunnen we het verwacht aantal abonnementen per dag in functie van de ervaring als volgt weergeven:



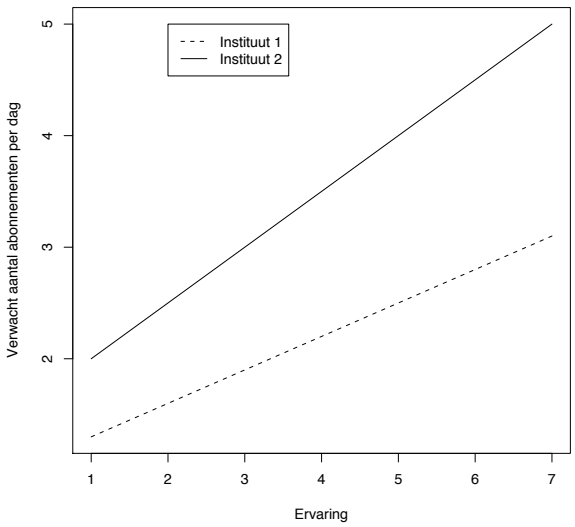
- In het geval dat $E(Y_i|x_{i1}, x_{i2}) = 1 + 0.3x_{i1} + 0.5x_{i2} + 0.2x_{i1}x_{i2}$ waarbij $x_{i2} = 0$ voor instituut 1 en $x_{i2} = 1$ voor instituut 2, is er wel een interactie tussen beide predictoren. Het verwacht aantal abonnementen per dag in functie van de ervaring van de verkoper die een opleiding kreeg in instituut 1 is immers

$$E(Y_i|x_{i1}, x_{i2} = 0) = 1 + 0.3x_{i1}$$

terwijl dit voor een verkoper die een opleiding kreeg in instituut 2 als volgt is

$$E(Y_i|x_{i1}, x_{i2} = 1) = 1 + 0.3x_{i1} + 0.5 + 0.2x_{i1} = 1.5 + 0.5x_{i1}.$$

We zien dat het effect van de ervaring groter is voor een verkoper die een opleiding in instituut 2 kreeg. Dit wordt getoond op onderstaande figuur.



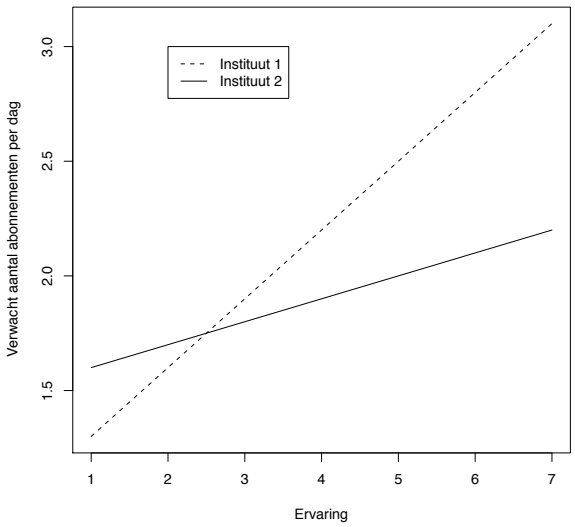
- In het geval dat $E(Y_i|x_{i1}, x_{i2}) = 1 + 0.3x_{i1} + 0.5x_{i2} - 0.2x_{i1}x_{i2}$ waarbij $x_{i2} = 0$ voor instituut 1 en $x_{i2} = 1$ voor instituut 2, is er ook een interactie tussen beide predictoren. Het verwacht aantal abonnementen per dag in functie van de ervaring van de verkoper die een opleiding kreeg in instituut 1 is dan

$$E(Y_i|x_{i1}, x_{i2} = 0) = 1 + 0.3x_{i1}$$

terwijl dit voor een verkoper die een opleiding kreeg in instituut 2 als volgt is

$$E(Y_i|x_{i1}, x_{i2} = 1) = 1 + 0.3x_{i1} + 0.5 - 0.2x_{i1} = 1.5 + 0.1x_{i1}.$$

Hier is het effect van de ervaring kleiner voor een verkoper die een opleiding in instituut 2 kreeg. Dit wordt getoond op onderstaande figuur.



Op de bovenstaande figuren zien we dat de regressierechten (voor verwacht aantal in functie van ervaring) voor instituut 1 en instituut 2 parallel zijn wanneer er geen interactie aanwezig is. Bij interactie zijn deze rechten niet langer parallel.

In de praktijk is het nuttig om een dergelijke plot te maken op basis van de geschatte regressievergelijkingen en/of geschatte groepsgemiddeldes. Op die manier kan onderzocht worden of er aanwijzingen zijn voor interacties (zie verder).

5.2 Hoofd- en interactie-effecten

In bovenstaand voorbeeld representeren β_1 en β_2 de *hoofdeffecten* van respectievelijk X_1 en X_2; β_3 stelt het *interactie-effect* voor. De aanwezigheid van een interactie impliceert dat effecten niet eenduidig te interpreteren zijn: het effect van X_1 hangt immers af van het niveau van X_2 en vice versa.

5.3 Implementatie en toetsen van interactie-effecten

Een term voor de interactie tussen 2 predictoren kan eenvoudig in een model toegevoegd worden door een nieuwe variabele te creëren waarvan ieder element het product is van de overeenkomstige elementen van de predictoren en die variabele in het model te stoppen.

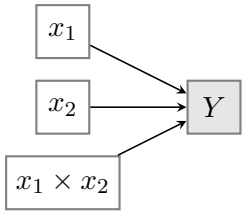
Het implementeren van interactie-effecten is routine en is vrij eenvoudig in statistische software zoals R.

Wanneer men echter interactietermen toevoegt aan het model, is het mogelijk dat er problemen met collineariteit ontstaan door de correlatie tussen de predictoren en de interactietermen. Collineariteit treedt op als twee (of meer) predictoren sterk met elkaar correleren. Dit wordt ook multicollineariteit genoemd en heeft als gevolg dat de standaardfouten van schatters heel groot kunnen worden. Dit euvel wordt gedeeltelijk verholpen door de predictoren van intervalniveau waarvoor een interactieterm aanwezig is eerst te centreren of nog: door de oorspronkelijke waarden van de predictoren $x_{i\ell}$ te vervangen door hun deviatiescore $x_{i\ell} - \bar{x}_\ell$. Dit zorgt er ook voor dat de parameters voor de hoofdeffecten makkelijker te interpreteren zijn.

Het toetsen van interactie-effecten gebeurt aan de hand van modelvergelijkingen zoals beschreven in de voorgaande sectie. In de aanwezigheid van interacties, maakt het wel een verschil uit welk type toets we gebruiken. De volgende types bestaan:

- **Type III toetsen**: effecten worden getoetst terwijl gecorrigeerd wordt voor alle andere effecten; i.e. het model met een effect wordt vergeleken met het model zonder dat effect.

  Concreet: stel dat in een regressiemodel met 2 predictoren een interactie tussen X_1 en X_2 opgenomen is, dan zal de toets voor het hoofdeffect van X_1 het volledige model (2 hoofdeffecten + interactie-effect) vergelijken met het model zonder het hoofdeffect van X_1 dus een model met een hoofdeffect van X_2 en het interactie-effect. De toets voor het interactie-effect vergelijkt het volledige model met het model zonder interactie-effect.

  Toetsing van hoofdeffect van X_2:

  - Via modelvergelijking:
    - model 1: $X_1 + X_2 + (X_1 \times X_2)$
    - model 2: $X_1 + (X_1 \times X_2)$

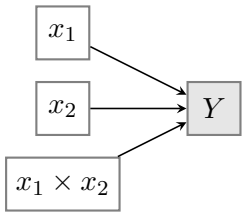
  Toetsing van interactie-effect:

  - Via modelvergelijking:
    - model 1: $X_1 + X_2 + (X_1 \times X_2)$

    * model 2: $X_1 + X_2$

- **Type II toetsen**: effecten worden getoetst terwijl gecorrigeerd wordt voor alle effecten van dezelfde of een lagere orde maar niet voor hogere orde effecten die het te toetsen effect omvatten.

  Concreet: stel dat in een regressiemodel met 2 predictoren een interactie tussen X_1 en X_2 opgenomen is, dan zal de toets voor het hoofdeffect van X_1 de interactie tussen X_1 en X_2 niet in rekening brengen aangezien dit een hogere orde term is die X_1 omvat. Hier wordt dan een model met X_1 en X_2 (zonder de interactieterm) vergeleken met een model met enkel X_2 (zonder de interactieterm). De toets voor het interactie-effect vergelijkt het volledige model (2 hoofdeffecten + interactie-effect) met het model zonder interactie-effect.



  Toetsing van hoofdeffect van X_2:

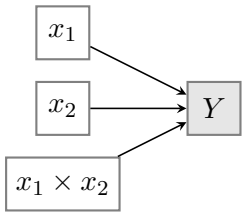
  – Via modelvergelijking:
    * model 1: $X_1 + X_2$
    * model 2: X_1

  Toetsing van interactie-effect:

  – Via modelvergelijking:
    * model 1: $X_1 + X_2 + (X_1 \times X_2)$
    * model 2: $X_1 + X_2$

- **Type I toetsen**: hier volgt het toetsen een sequentiële strategie en hangt dus af van de volgorde waarin predictoren aan een model toegevoegd worden.

  Concreet: als we eerst X_1 als predictor in een model voegen, dan X_2 en vervolgens het interactie-effect, zal de toets voor X_1 het model met enkel X_1 vergelijken met het nulmodel. De toets voor X_2 zal het model met X_1 en X_2 vergelijken met het model met enkel X_1 terwijl de toets voor het interactie-effect het volledige model (2 hoofdeffecten + interactie-effect) vergelijkt met het model zonder de interactie.



Toetsing van hoofdeffect van X_1:

- Via modelvergelijking:
  - model 1: X_1
  - model 2: nulmodel (geen predictoren)

Toetsing van hoofdeffect van X_2:

- Via modelvergelijking:
  - model 1: $X_1 + X_2$
  - model 2: X_1

Toetsing van interactie-effect:

- Via modelvergelijking:
  - model 1: $X_1 + X_2 + (X_1 \times X_2)$
  - model 2: $X_1 + X_2$

In R worden standaard de Type I toetsen gehanteerd. Daarom maken wij gebruik van `Anova` (package `car`) waarbij we kunnen aangeven of we Type II of Type III toetsen gebruiken. Uit de redenering hierboven kan afgeleid worden dat de toets voor het interactie-effect hetzelfde is maar er is een verschil tussen beide types voor het toetsen van de hoofdeffecten.

Wij gebruiken **Type III toetsen** aangezien dit de standaard is binnen andere softwarepakketten zoals SPSS.

Opmerking met betrekking tot nominale predictoren

Wanneer we interacties met nominale predictoren beschouwen, dienen ook extra hulpveranderlijken aangemaakt te worden.

In het algemeen, voor 2 nominale variabelen met respectievelijk I en J niveaus, hebben we $(I-1) \times (J-1)$ hulpveranderlijken nodig om het interactie-effect te representeren.

Om te onderzoeken of een interactie-effect statistisch significant is, moeten we een modelvergelijkingstoets uitvoeren om na te gaan of de regressiecoëfficiënten die horen bij de $(I-1) \times (J-1)$ hulpveranderlijken van de interactie 0 zijn.

Ongeacht of men dummy- of effect-codering gebruikt voor de nominale variabelen, geeft de toets voor het interactie-effect hetzelfde resultaat. Wanneer een interactieterm aanwezig is, zullen de toetsen voor de hoofdeffecten wel verschillen naargelang de codering die gehanteerd wordt. Bij dummy-codering wordt het effect van een nominale variabele getoetst binnen het referentieniveau van de andere nominale variabele. Bij effect-codering wordt het effect van een nominale variabele getoetst, uitgemiddeld over de niveaus van de andere nominale variabele.

Merk opnieuw op dat we in deze cursus toetsen aan de hand van **Type III toetsen** en dat we hierbij voor het toetsen overgaan op **effect-codering** van de nominale predictoren.

Opmerking met betrekking tot een combinatie van nominale predictoren en predictoren van intervalniveau

Als we een interactie tussen een nominale predictor van I niveaus en een predictor van intervalniveau beschouwen, zijn er $(I-1)$ hulpveranderlijken die dit interactie-effect representeren. Nagaan of het interactie-effect statistisch significant is, houdt dus in dat we een modelvergelijkingstoets uitvoeren om na te gaan of de regressiecoëfficiënten die horen bij de $(I-1)$ hulpveranderlijken van de interactie 0 zijn.

De hierboven beschreven situatie waarbij een verschil bestaat tussen toetsen voor hoofdeffecten wanneer men dummy- of effect-codering gebruikt, doet zich ook voor wanneer de interactie tussen een nominale predictor en een predictor van intervalniveau onderzocht wordt.

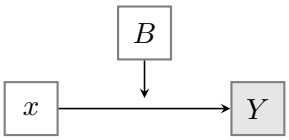
Opnieuw geldt hier dat we toetsen aan de hand van **Type III toetsen** en dat we hierbij voor het toetsen overgaan op **effect-codering** van de nominale predictoren.

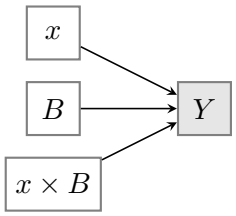
Verschillende gevallen

1. Interactie tussen 2 nominale predictoren (factoren) A en B

- toetsing: via modelvergelijking:
  - model 1: $A + B + (A \times B)$
  - model 2: $A + B$
- indien interactie $(A \times B)$ significant: moderatie

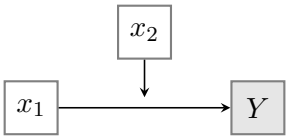
2. Interactie tussen een nominale predictor B en een predictor van intervalniveau x

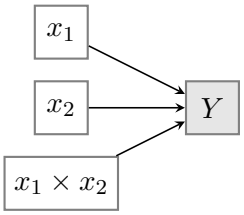




- toetsing: via modelvergelijking:
  - model 1: $x + B + (x \times B)$
  - model 2: $x + B$
- indien interactie $(x \times B)$ significant: moderatie

3. Interactie tussen 2 predictoren van intervalniveau x_1 en x_2





- toetsing: via modelvergelijking:
  - model 1: $x_1 + x_2 + (x_1 \times x_2)$
  - model 2: $x_1 + x_2$
- indien interactie $(x_1 \times x_2)$ significant: moderatie

5.4 Voorbeeld: herstel na coma

We illustreren de interpretatie en het toetsen van interactie-effecten in de verschillende gevallen aan de hand van het voorbeeld over het herstel na een coma (zie sectie 1.2.3).

We modelleren het wiskundig IQ (`piq`) in functie van verbaal IQ (`viq`), duur van de coma in dagen (hier gebruiken we de versie waarbij de 4 intervallen onderscheiden worden, `duration_cat`), gender van de patiënt (`sex`), leeftijd van de patiënt (`age`) .

`viq` en `age` zijn predictoren van intervalniveau, `sex` is een predictor van nominaal niveau en `duration_cat` beschouwen we ook als een predictor van nominaal niveau.

```
> class(coma$duration_cat)
[1] "factor"
> contrasts(coma$duration_cat)
          (1,7] (7,14] (14,255]
[0,1]         0      0        0
(1,7]         1      0        0
(7,14]        0      1        0
(14,255]      0      0        1

> class(coma$sex)
[1] "factor"
> contrasts(coma$sex)
       Male
Female    0
Male      1
```

Bij het interpreteren van de parameters zullen we dummy-codering gebruiken. Bij `duration_cat` is de kortste duur de referentiecategrie en bij `sex` is vrouw de referentiecategorie.

5.4.1 Het lineair regressiemodel zonder interacties

We bekijken eerst het lineair regressiemodel zonder interacties.

```
> fit1_coma<-lm(piq~viq+duration_cat+sex+age,data=coma)
> summary(fit1_coma)
```

```
Call:
lm(formula = piq ~ viq + duration_cat + sex + age, data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-34.639  -5.847  -0.503   6.532  29.852

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          30.60636    6.41685   4.770 3.63e-06 ***
viq                   0.60328    0.06085   9.914  < 2e-16 ***
duration_cat(1,7]    -2.26030    2.15909  -1.047 0.296466
duration_cat(7,14]   -4.33226    2.28508  -1.896 0.059468 .
duration_cat(14,255] -8.28546    2.19158  -3.781 0.000209 ***
sexMale              -0.72454    1.91321  -0.379 0.705322
age                   0.03617    0.05705   0.634 0.526768
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.01 on 193 degrees of freedom
Multiple R-squared:  0.3962,Adjusted R-squared:  0.3774
F-statistic: 21.11 on 6 and 193 DF,  p-value: < 2.2e-16
```

De verschillende effecten in het model kunnen ook weergegeven worden aan de hand van de functie `effect` in het R package `effects`.

Voor `viq` zien we een positief effect ($\hat{b} = 0.60$). Dit betekent dat wanneer alle overige predictoren constant blijven, de gemiddelde `piq` toeneemt als `viq` toeneemt.

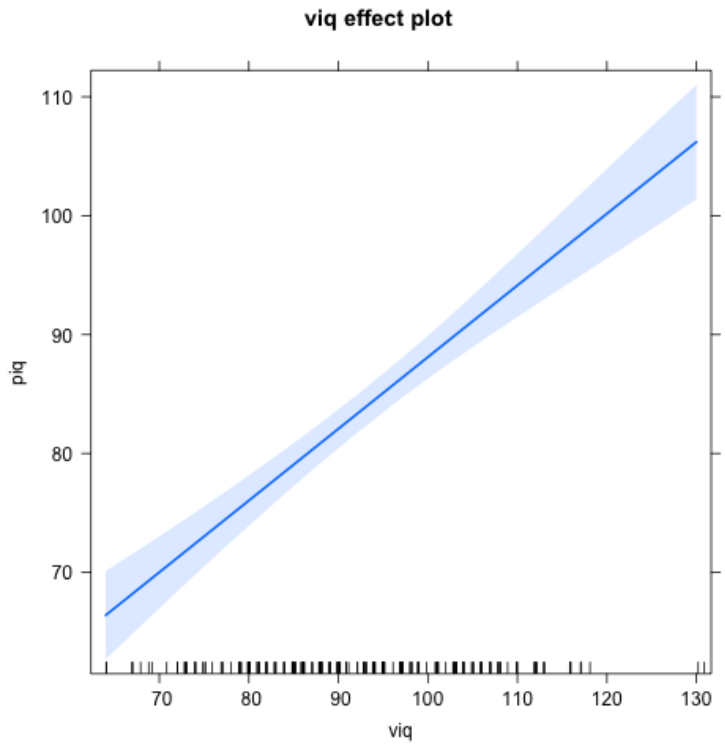
```
> library(effects)
> effect("viq",fit1_coma)

viq effect
viq
64        81        98       110       130
66.38895  76.64466  86.90037  94.13969 106.20524
```

Via `effects` wordt voor een aantal waarden van verbaal IQ (`viq` = 64, 81, 98, 110 , 130) het wiskundige IQ voorspeld: predictie `piq` = 66.39, 76.65, 86.90, 94.14, 106.21. Hierbij worden de

overige predictoren constant gehouden. Bij de predictie worden predictoren van intervalniveau gelijk gesteld aan hun gemiddelde waarde, terwijl een gewogen gemiddelde van de uitkomst berekend wordt over de verschillende groepen gevormd door de (combinaties van) nominale predictoren. Aangezien er geen interactieterm met `viq` in het model zit, maakt het niveau waarop de overige predictoren constant gehouden worden niet uit voor het effect van `viq`, enkel voor de voorspelde waarde van de uitkomst `piq`.

Het effect kan ook grafisch voorgesteld worden (met betrouwbaarheidsinterval):

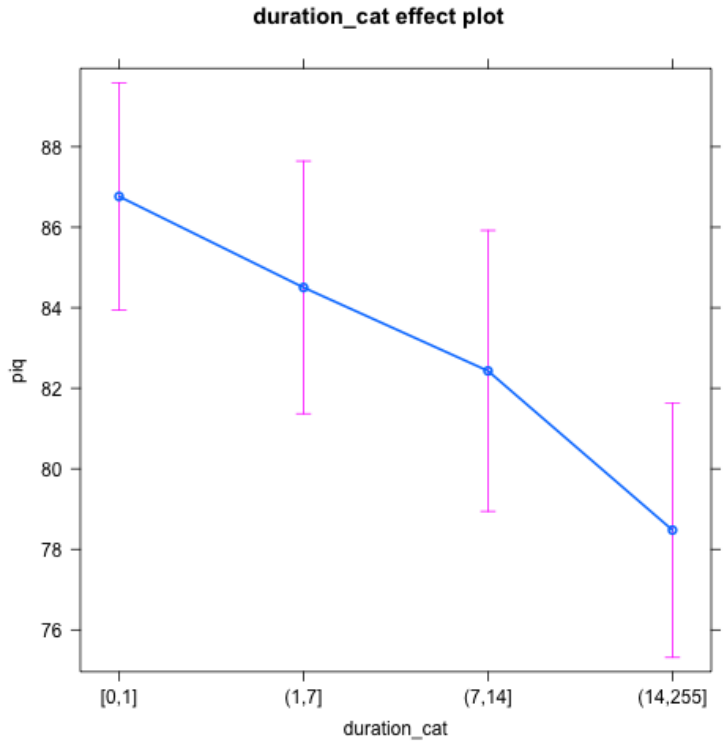
```
plot(effect("viq",fit1_coma))
```



Voor het effect van de duur van de coma (gecategoriseerd) krijgen we:

```
> effect("duration_cat",fit1_coma)

duration_cat effect
duration_cat
```

```
[0,1]    (1,7]    (7,14]  (14,255]
86.76350 84.50320 82.43124 78.47804
```



Hier zien we, net als op basis van de geschatte regressiecoëfficiënten, dat het geschatte gemiddelde wiskundig IQ afneemt naarmate de coma langer geduurd heeft.

5.4.2 Interactie tussen 2 nominale predictoren

We voegen de interactie tussen de duur van de coma (`duration_cat`) en gender (`sex`) in het model toe. Dit kunnen we doen door de term `duratio_cat:sex` in het model toe te voegen.

Opmerking: wanneer we `duration_cat*sex` *in het model toevoegen, zal R automatisch ook alle hoofdeffecten van de termen die in de interactie opgenomen zijn, toevoegen. Hier maakt dit geen verschil omdat deze hoofdeffecten al in het model zitten. Bovendien beschouwen wij geen modellen waar een interactieterm in zit zonder de hoofdeffecten van de termen die in de interactie opgenomen zijn.*

```
> fit2_coma<-lm(piq~viq+duration_cat+sex+age+duration_cat:sex,data=coma)
> summary(fit2_coma)

Call:
lm(formula = piq ~ viq + duration_cat + sex + age + duration_cat:sex,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-33.802  -6.378  -0.374   5.962  27.361

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   35.26586    6.95253   5.072  9.3e-07 ***
viq                            0.58947    0.06127   9.620  < 2e-16 ***
duration_cat(1,7]             -7.29674    4.18644  -1.743  0.08296 .
duration_cat(7,14]            -6.49533    5.16375  -1.258  0.20998
duration_cat(14,255]         -15.80542    5.01042  -3.155  0.00187 **
sexMale                       -5.17123    3.44390  -1.502  0.13487
age                            0.04026    0.05717   0.704  0.48211
duration_cat(1,7]:sexMale      6.67411    4.84814   1.377  0.17025
duration_cat(7,14]:sexMale     2.76867    5.74587   0.482  0.63046
duration_cat(14,255]:sexMale   9.22301    5.50591   1.675  0.09556 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11 on 190 degrees of freedom
Multiple R-squared:  0.407,Adjusted R-squared:  0.3789
F-statistic: 14.49 on 9 and 190 DF,  p-value: < 2.2e-16
```

Om het interactie-effect te toetsen via `Anova` en Type III toetsen, moeten we zorgen dat bij de nominale predictoren effect-codering gebruikt wordt. Dit is louter om te toetsen, om het interactie-effect te interpreteren maken we gebruik van het bovenstaande geschatte model (dummy-codering).

```
> fit2_coma_test<-lm(piq~viq+duration_cat+sex+age+duration_cat:sex,
                  contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova(fit2_coma_test,type=3)
Anova Table (Type III tests)
```

```
Response: piq
                  Sum Sq  Df F value    Pr(>F)
(Intercept)       2585.6   1 21.3623 7.005e-06 ***
viq              11202.0   1 92.5523 < 2.2e-16 ***
duration_cat      1947.3   3  5.3630  0.001446 **
sex                  7.9   1  0.0652  0.798744
age                 60.0   1  0.4960  0.482108
duration_cat:sex   417.7   3  1.1504  0.330054
Residuals        22996.6 190
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 3 en 190 vrijheidsgraden. De geobserveerde waarde is gelijk aan (417.7/3)/(22996.6/190)=1.15. De bijhorende p-waarde is gelijk aan 0.33.

Hoewel de interactie niet statistisch significant is op het 5% significantieniveau, gaan we ter illustratie na wat deze (niet-signifcante) interactie inhoudt.

1. **Geschatte effect van gender**: we bekijken op basis van het model het geschatte effect van gender voor verschillende categorieën van de duur van de coma. Een interactie-effect in het model impliceert dat dit geschatte effect zal wijzigen naargelang de duur van de coma.

   De overige predictoren (`viq`, `age`) die niet in het interactie-effect zitten, houden we constant en gelijk aan hun gemiddelde. Zoals de berekeningen hieronder tonen, hangt het geschatte effect van gender binnen een bepaalde duur niet af van het niveau waarop we deze variabelen constant houden aangezien er geen interactie met deze variabelen is.

   ```
   > mean(coma$viq)
   [1] 92.09
   > mean(coma$age)
   [1] 32.34668
   ```

   (a) Kortste duur van coma (`[0,1]`)

   - Voorspelde waarde uitkomst (`piq`) voor een vrouw in de categorie kortste duur:

   $$\hat{E}_1(Y) = 35.26586 + 0.58947 * 92.09$$
   $$\textit{intercept + verbaal IQ}$$

$$
\begin{aligned}
& -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 0 \\
& \qquad \textit{kortste duur is referentie dus 3 hulpveranderlijken = 0} \\
& -5.17123 * 0 \\
& \qquad \textit{vrouw is referentie dus hulpveranderlijke = 0} \\
& +0.04026 * 32.34668 \\
& \qquad \textit{leeftijd} \\
& +6.67411 * 0 * 0 + 2.76867 * 0 * 0 + 9.22301 * 0 * 0 \\
& \qquad \textit{interactie tussen duur en gender} \\
= \; & 90.85243
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
                          sex="Female", age=mean(coma$age))
> predict(fit2_coma,newdata)
90.85273
```

- Voorspelde waarde uitkomst (`piq`) voor een man in de categorie kortste duur:

$$
\begin{aligned}
\hat{E}_2(Y) \; = \; & 35.26586 + 0.58947 * 92.09 \\
& \qquad \textit{intercept + verbaal IQ} \\
& -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 0 \\
& \qquad \textit{kortste duur is referentie dus 3 hulpveranderlijken = 0} \\
& -5.17123 * 1 \\
& \qquad \textit{vrouw is referentie dus hulpveranderlijke voor man = 1} \\
& +0.04026 * 32.34668 \\
& \qquad \textit{leeftijd} \\
& +6.67411 * 0 * 1 + 2.76867 * 0 * 1 + 9.22301 * 0 * 1 \\
& \qquad \textit{interactie tussen duur en gender} \\
= \; & 85.6812
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
                        sex="Male",age=mean(coma$age))
> predict(fit2_coma,newdata)
1
85.6815
```

Effect van gender (mannen versus vrouwen) bij kortste duur:
85.6815-90.85273=-5.17123. Dit kunnen we ook rechtstreeks aflezen uit de geschatte

regressiecoëfficiënten in de output van het gehanteerde regressiemodel. Het geschatte hoofdeffect van gender geeft bij dummy-codering het effect van gender weer binnen het referentieniveau van de duur van de coma.

(b) Langste duur van coma (`(14,255]`)

- Voorspelde waarde uitkomst (`piq`) voor een vrouw in de categorie langste duur:

$$\begin{aligned}\hat{E}_3(Y) &= 35.26586 + 0.58947 * 92.09 \\ &\qquad \textit{intercept + verbaal IQ} \\ &\quad -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 1 \\ &\qquad \textit{langste duur dus 3e hulpveranderlijke=1, de rest 0} \\ &\quad -5.17123 * 0 \\ &\qquad \textit{vrouw is referentie dus hulpveranderlijke = 0} \\ &\quad +0.04026 * 32.34668 \\ &\qquad \textit{leeftijd} \\ &\quad +6.67411 * 0 * 0 + 2.76867 * 0 * 0 + 9.22301 * 1 * 0 \\ &\qquad \textit{interactie tussen duur en gender} \\ &= 75.0473\end{aligned}$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="(14,255]",
                                  sex="Female",age=mean(coma$age))
> predict(fit2_coma,newdata)
1
75.04731
```

- Voorspelde waarde uitkomst (`piq`) voor een man in de categorie langste duur:

$$\begin{aligned}\hat{E}_4(Y) &= 35.26586 + 0.58947 * 92.09 \\ &\qquad \textit{intercept + verbaal IQ} \\ &\quad -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 1 \\ &\qquad \textit{langste duur dus 3e hulpveranderlijke=1, de rest 0} \\ &\quad -5.17123 * 1 \\ &\qquad \textit{vrouw is referentie dus hulpveranderlijke voor man = 1} \\ &\quad +0.04026 * 32.34668 \\ &\qquad \textit{leeftijd} \\ &\quad +6.67411 * 0 * 1 + 2.76867 * 0 * 1 + 9.22301 * 1 * 1 \\ &\qquad \textit{interactie tussen duur en gender}\end{aligned}$$

$= 79.09879$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="(14,255]",
                              sex="Male",age=mean(coma$age))
> predict(fit2_coma,newdata)
1
79.09909
```

Effect van gender (mannen versus vrouwen) bij langste duur: 79.09909-75.04731 =4.05178. Dit kunnen we ook afleiden uit de geschatte regressiecoëfficiënten in de output van het gehanteerde regressiemodel: -5.171232 (hoofdeffect gender) +9.22301 (interactie-effect met duur = langste duur)=4.051778.

2. **Geschatte effect van de duur van de coma**: we bekijken op basis van het model het geschatte effect van de duur van de coma voor mannen en vrouwen apart. Een interactie-effect in het model impliceert dat dit geschatte effect zal wijzigen naargelang gender.

   De overige predictoren (`viq`, `age`) die niet in het interactie-effect zitten, houden we opnieuw constant en gelijk aan hun gemiddelde.

   (a) Vrouwen

   Het geschatte verschil in gemiddeld wiskundig IQ bij vrouwen voor de langste versus de kortste duur is $\hat{E}_3(Y) - \hat{E}_1(Y) = 75.04731 - 90.85273 = -15.80542$. Dit is op afronding na de geschatte regressiecoëfficiënt die we aflezen bij het hoofdeffect van de hoogste categorie (i.e. het geschatte effect van de hoogste versus de laagste categorie bij vrouwen).

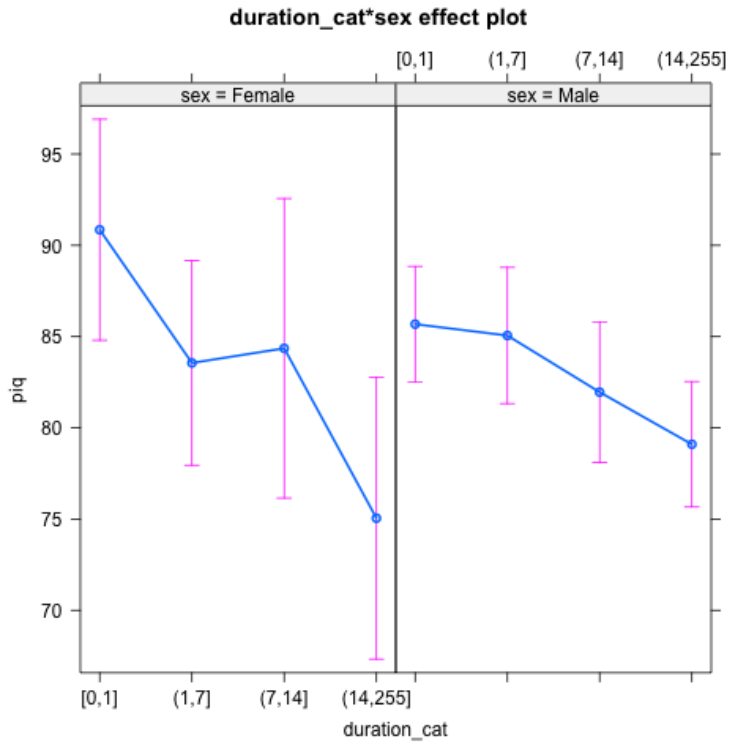
   (b) Mannen

   Het geschatte verschil in gemiddeld wiskundig IQ bij mannen voor de langste versus de kortste duur is $\hat{E}_4(Y) - \hat{E}_2(Y) = 79.09879 - 85.6812 = -6.58241$. Dit kunnen we (op afronding na) afleiden uit de geschatte regressiecoëfficiënten: -15.80542 (hoofdeffect hoogste categorie) + 9.22301 (interactie-effect met gender=man) = -6.58241.

In R krijgen we de volgende weergave van het interactie-effect:

```
> effect("duration_cat:sex", fit2_coma)

duration_cat*sex effect
```

```
sex
duration_cat   Female     Male
[0,1]         90.85273 85.68150
(1,7]         83.55599 85.05887
(7,14]        84.35740 81.95484
(14,255]      75.04731 79.09909
```



5.4.3 Interactie tussen een nominale predictor en een predictor van intervalniveau

Geval 1: de nominale predictor bestaat uit 2 niveaus

We voegen de interactie tussen gender (`sex`) en leeftijd (`age`) in het model toe.

```
> fit3_coma<-lm(piq~viq+duration_cat+sex+age+sex:age,data=coma)
> summary(fit3_coma)
```

```
Call:
lm(formula = piq ~ viq + duration_cat + sex + age + sex:age,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-34.726  -5.869  -0.563   6.470  30.375

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          29.46295    6.94603   4.242 3.45e-05 ***
viq                   0.59959    0.06157   9.739  < 2e-16 ***
duration_cat(1,7]    -2.28337    2.16429  -1.055 0.292741
duration_cat(7,14]   -4.41770    2.29828  -1.922 0.056064 .
duration_cat(14,255] -8.35007    2.20121  -3.793 0.000199 ***
sexMale               1.20384    4.82652   0.249 0.803300
age                   0.08334    0.12249   0.680 0.497094
sexMale:age          -0.05973    0.13719  -0.435 0.663789
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.04 on 192 degrees of freedom
Multiple R-squared:  0.3968,Adjusted R-squared:  0.3748
F-statistic: 18.04 on 7 and 192 DF,  p-value: < 2.2e-16
```

De resultaten van de toetsen van de verschillende effecten kunnen hieronder afgelezen worden. Opnieuw: om het interactie-effect te interpreteren maken we gebruik van het bovenstaande geschatte model.

```
> fit3_coma_test<-lm(piq~viq+duration_cat+sex+age+age:sex,
                   contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova((fit3_coma_test),type=3)
Anova Table (Type III tests)

Response: piq
              Sum Sq  Df F value    Pr(>F)
(Intercept)   2359.0   1 19.3629 1.791e-05 ***
viq          11554.9   1 94.8446 < 2.2e-16 ***
```

```
duration_cat  1863.0   3  5.0974  0.002042 **
sex              7.6   1  0.0622  0.803300
age             71.9   1  0.5899  0.443394
sex:age         23.1   1  0.1895  0.663789
Residuals    23391.2 192
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 1 en 192 vrijheidsgraden. De geobserveerde waarde is gelijk aan (23.1)/(23391/192)=0.19. De bijhorende p-waarde is gelijk aan 0.66.

Hoewel de interactie niet statistisch significant is op het 5% significantieniveau, gaan we ter illustratie opnieuw na wat deze (niet-signifcante) interactie inhoudt.

We leiden af wat het geschatte effect is van leeftijd op het gemiddelde wiskundig IQ. Een interactie-effect tussen gender en leeftijd impliceert dat dit geschatte effect anders zal zijn bij vrouwen en mannen.

1. **Geschatte effect van leeftijd bij vrouwen**: bij het afleiden van dit effect worden de predictoren die niet in de interactie zitten constant gehouden. Verbaal IQ (`viq`) stellen we gelijk aan het gemiddelde en we kijken binnen de laagste categorie van duur (`duration_cat`). Het geschatte effect van leeftijd hangt niet af van het niveau waarop we deze variabelen constant houden aangezien er geen interactie met deze variabelen is.

   ```
   > mean(coma$viq)
   [1] 92.09
   > mean(coma$age)
   [1] 32.34668
   ```

   (a) Voorspelde waarde uitkomst (`piq`) voor een vrouw met gemiddelde leeftijd in de categorie kortste duur:

   $$\begin{aligned}\hat{E}_{\overline{\text{age}}}(Y) &= 29.46295 + 0.59959 * 92.09 \\ &\qquad \textit{intercept + verbaal IQ} \\ &\quad -2.28337 * 0 - 4.41770 * 0 - 8.35007 * 0 \\ &\qquad \textit{kortste duur is referentie dus 3 hulpveranderlijken = 0} \\ &\quad +1.20384 * 0 \\ &\qquad \textit{vrouw is referentie dus hulpveranderlijke = 0}\end{aligned}$$

$$
\begin{aligned}
& +0.08334 * 32.34668 \\
& \qquad \textit{leeftijd} \\
& -0.05973 * 0 * 32.34668 \\
& \qquad \textit{interactie tussen gender en leeftijd} \\
= \; & 87.37497
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
            sex="Female",age=mean(coma$age))
> predict(fit3_coma,newdata)
1
87.37468
```

(b) Voorspelde waarde uitkomst (`piq`) voor een vrouw in de categorie kortste duur wanneer leeftijd met 1 eenheid toeneemt:

$$
\begin{aligned}
\hat{E}_{\overline{\text{age}}+1}(Y) \; = \; & 29.46295 + 0.59959 * 92.09 \\
& \qquad \textit{intercept + verbaal IQ} \\
& +0.08334 * (32.34668 + 1) \\
& \qquad \textit{leeftijd} \\
= \; & 87.45831
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
             sex="Female",age=(mean(coma$age)+1))
> predict(fit3_coma,newdata)
1
87.45802
```

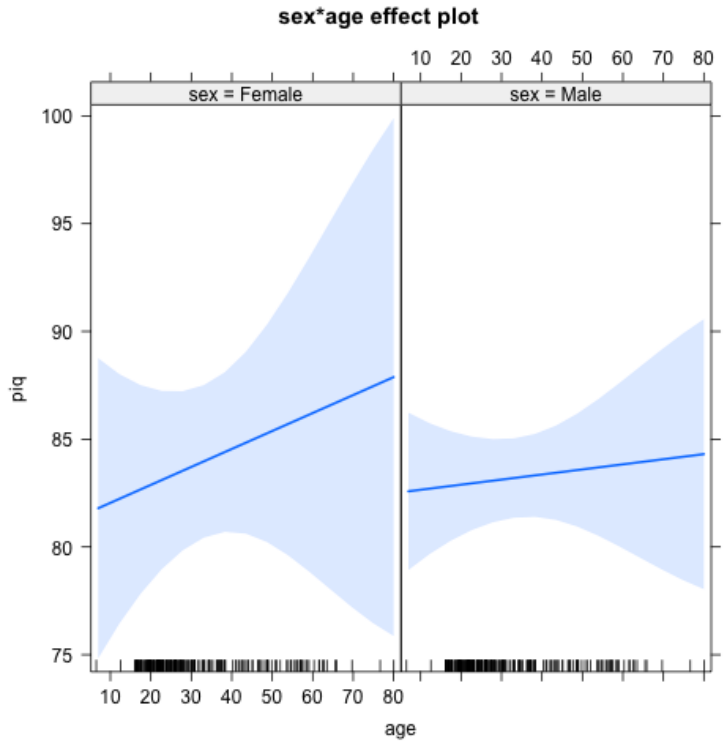
Het geschatte effect van leeftijd is gelijk aan 87.45802-87.37468=0.08334. Dit is de regressiecoëfficiënt die we aflezen bij het hoofdeffect van leeftijd (i.e. geschatte effect van leeftijd binnen referentieniveau van gender).

2. **Geschatte effect van leeftijd bij mannen**: opnieuw houden we verbaal IQ (`viq`) constant en gelijk aan het gemiddelde en we kijken binnen de laagste categorie van duur (`duration_cat`).

   Analoog aan voorgaande berekeningen kunnen we afleiden dat het geschatte effect gelijk is aan 0.08334 (hoofdeffect van leeftijd)-0.05973 (interactie-effect met gender=man)=0.02361. Dit geschatte effect van leeftijd bij mannen is kleiner dan dit effect bij vrouwen.

In R krijgen we de volgende weergave van het interactie-effect:

```
> effect("sex:age",fit3_coma)

sex*age effect
         age
sex         7       20       40       60       80
Female 81.79573 82.87911 84.54584 86.21257 87.87930
Male   82.58148 82.88839 83.36056 83.83273 84.30489
```



Geval 2: de nominale predictor bestaat uit meer dan 2 niveaus

We voegen de interactie tussen de duur van de coma (`duration_cat`) en leeftijd (`age`) in het model toe.

```
> fit4_coma<-lm(piq~viq+duration_cat+sex+age+duration_cat:age,data=coma)
```

```
> summary(fit4_coma)

Call:
lm(formula = piq ~ viq + duration_cat + sex + age + duration_cat:age,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-34.907  -6.010  -0.659   6.805  29.498

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               31.70527    6.64696   4.770 3.66e-06 ***
viq                        0.60052    0.06175   9.725  < 2e-16 ***
duration_cat(1,7]         -2.02242    5.68100  -0.356   0.7222
duration_cat(7,14]        -0.39656    6.18241  -0.064   0.9489
duration_cat(14,255]     -13.33837    5.11249  -2.609   0.0098 **
sexMale                   -0.79532    1.91966  -0.414   0.6791
age                        0.01567    0.08890   0.176   0.8603
duration_cat(1,7]:age     -0.01379    0.16101  -0.086   0.9319
duration_cat(7,14]:age    -0.12992    0.17290  -0.751   0.4533
duration_cat(14,255]:age   0.17453    0.14568   1.198   0.2324
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.02 on 190 degrees of freedom
Multiple R-squared:  0.4055,Adjusted R-squared:  0.3774
F-statistic:  14.4 on 9 and 190 DF,  p-value: < 2.2e-16
```

Op basis van bovenstaande output zien we dat het geschatte effect van leeftijd (voor een constant verbaal IQ en gender) binnen de kortste duur (0 tot 1 dag) gelijk is aan 0.01567, binnen de categorie met een duur van meer dan 1 dag tot 7 dagen 0.01567-0.01379=0.00188, binnen de categorie met een duur van meer dan 7 dagen tot 14 dagen 0.01567-0.12992=-0.11425 en binnen de categorie met een duur van meer dan 14 dagen 0.01567+0.17453=0.1902. Het opnemen van een interactie-effect tussen duur en leeftijd impliceert dat het geschatte effect van leeftijd varieert naargelang de duur van de coma.

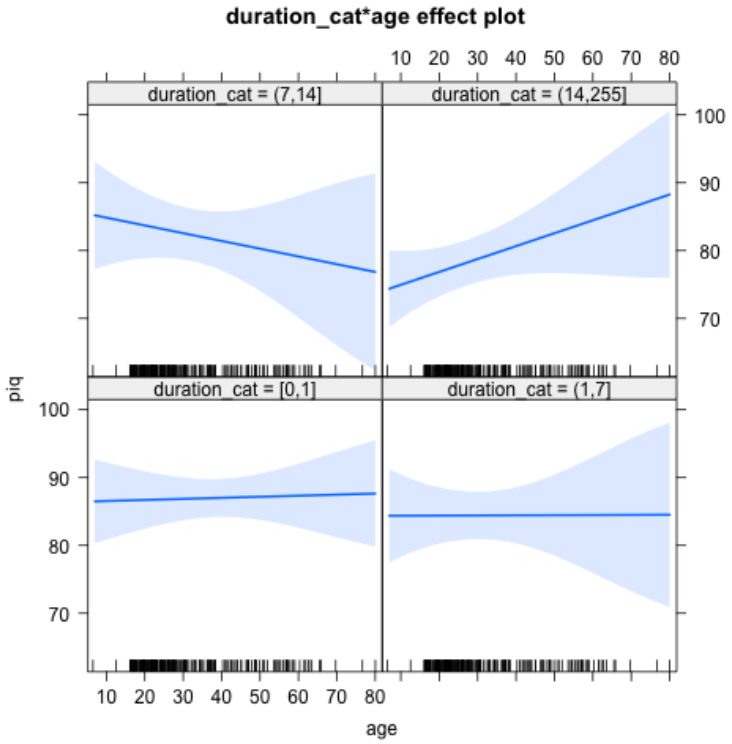
De resultaten voor de toetsen van de verschillende effecten in het model kunnen we aflezen in onderstaande output:

```
> fit4_coma_test<-lm(piq~viq+duration_cat+sex+age+duration_cat:age,
               contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova((fit4_coma_test),type=3)
Anova Table (Type III tests)

Response: piq
                   Sum Sq  Df F value    Pr(>F)
(Intercept)       2406.5   1 19.8338 1.439e-05 ***
viq              11474.2   1 94.5691 < 2.2e-16 ***
duration_cat      1058.3   3  2.9074   0.03591 *
sex                 20.8   1  0.1716   0.67912
age                 17.6   1  0.1450   0.70378
duration_cat:age   361.4   3  0.9930   0.39728
Residuals        23052.9 190
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 3 en 190 vrijheidsgraden. De geobserveerde waarde is gelijk aan (361.4/3)/(23052.9/190)=0.99. De bijhorende p-waarde is gelijk aan 0.397.

De weergave van deze (niet-significante) interactie is als volgt:

```
> effect("duration_cat:age",fit4_coma)

duration_cat*age effect
                  age
duration_cat  7       20       40       60       80
[0,1]    86.49294 86.69667 87.01010 87.32353 87.63696
(1,7]    84.37403 84.39855 84.43628 84.47401 84.51174
(7,14]   85.18694 83.70172 81.41676 79.13181 76.84685
(14,255] 74.37626 76.84886 80.65285 84.45684 88.26083
```



5.4.4 Interactie tussen 2 predictoren van intervalniveau

We voegen de interactie tussen verbaal IQ `viq` en leeftijd `age` in het model toe.

```
> fit5_coma<-lm(piq~viq+duration_cat+sex+age+viq:age,data=coma)
> summary(fit5_coma)

Call:
lm(formula = piq ~ viq + duration_cat + sex + age + viq:age,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-33.762  -5.993  -0.340   6.467  30.651

Coefficients:
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          38.121018  14.374328   2.652 0.008670 **
viq                   0.517736   0.158542   3.266 0.001294 **
duration_cat(1,7]    -2.071854   2.186682  -0.947 0.344581
duration_cat(7,14]   -4.094544   2.324836  -1.761 0.079793 .
duration_cat(14,255] -8.123493   2.212750  -3.671 0.000313 ***
sexMale              -0.580379   1.932286  -0.300 0.764228
age                  -0.197554   0.403956  -0.489 0.625366
viq:age               0.002572   0.004400   0.584 0.559586
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.03 on 192 degrees of freedom
Multiple R-squared:  0.3973,Adjusted R-squared:  0.3753
F-statistic: 18.08 on 7 and 192 DF,  p-value: < 2.2e-16
```

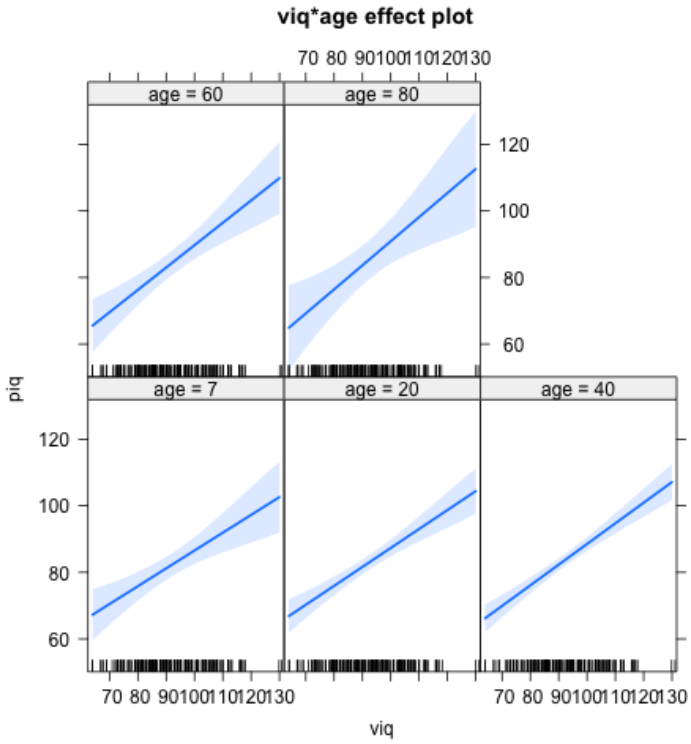
Het opnemen van het interactie-effect tussen verbaal IQ en leeftijd in het model impliceert dat het geschatte effect van verbaal IQ zal variëren naargelang de leeftijd en vice versa.

Analoog aan de afleidingen in sectie 5.1, zien we dat het geschatte effect van verbaal IQ gelijk is aan 0.517736+0.002572*`age`. Het geschatte effect van `viq` op het wiskundig IQ is bijgevolg positief en neemt lichtjes toe naarmate de leeftijd toeneemt.

Omgekeerd zien we ook dat het geschatte effect van leeftijd gelijk is aan -0.197554+0.002572*`viq`. Voor kleine waarden van `viq`, is het geschatte effect van leeftijd negatief; bij hoge waarden van `viq` zal dit positief worden.

Deze bevindingen zien we ook in de volgende output:

```
> effect("viq:age",fit5_coma)

viq*age effect
      age
viq        7        20        40        60        80
64   67.27360  66.84527  66.18629  65.52732  64.86835
81   76.38118  76.52125  76.73674  76.95223  77.16772
98   85.48876  86.19723  87.28718  88.37714  89.46710
110  91.91763  93.02733  94.73456  96.44178  98.14901
130 102.63243 104.41084 107.14684 109.88285 112.61886
```



Op de figuur zien we dat de rechten die het geschatte effect (slope) van `viq` weergeven voor verschillende leeftijd quasi parallel zijn. Dit komt overeen met een interactie-effect dat quasi 0 is. Op basis van de plots vinden we geen evidentie voor een bestaand interactie-effect tussen verbaal IQ en leeftijd m.b.t. hun effect op het gemiddeld wiskundig IQ.

Wanneer we het interactie-effect toetsen, zien we inderdaad dat dit niet statistisch significant is op het 5% significantieniveau ($p = 0.56$).

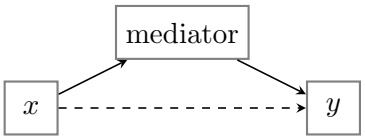
```
> fit5_coma_test<-lm(piq~viq+duration_cat+sex+age+viq:age,
            contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova(fit5_coma_test,type=3)
Anova Table (Type III tests)

Response: piq
              Sum Sq  Df F value   Pr(>F)
(Intercept)    678.6   1  5.5748 0.019222 *
viq           1298.2   1 10.6642 0.001294 **
```

```
duration_cat  1758.0   3  4.8139 0.002960 **
sex             11.0   1  0.0902 0.764228
age             29.1   1  0.2392 0.625366
viq:age         41.6   1  0.3416 0.559586
Residuals    23372.7 192
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

6 Mediatie

6.1 Wat is mediatie?

- Mediatie = interventie:
  - De relatie tussen y en x wordt *gemedieerd* door een derde variabele
  - of nog: de mediator *intervenieert* in de relatie tussen y en x.



- Mediatie impliceert een conceptuele causale hypothese (*the mediational hypothesis*): de onafhankelijke variabele x beïnvloedt de mediator, en de mediator beïnvloedt de afhankelijke variabele y.
- Mediatie probeert te verklaren waarom x een invloed heeft op y.
- Mediatie-effecten zijn alomtegenwoordig in de gedragswetenschappen!
- In wat volgt beschouwen we zowel y als m (de mediator) als variabelen van minstens intervalniveau.

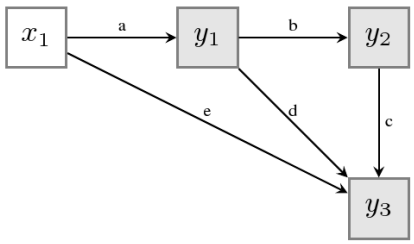
Een 'klassieke' paper waarbij het onderscheid tussen mediatie en moderatie aan bod komt, is:

> Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1185.

- Deze paper is al enorm veel geciteerd (zie Web of Science).
- De paper geeft een uitstekende beschrijving van het begrip *moderatie* en *mediatie*.
- Webpagina's:
  - `http://davidakenny.net/cm/moderation.htm`
  - `http://davidakenny.net/cm/mediate.htm`
- De auteurs beschrijven ook een 'statistische procedure' om na te gaan of er inderdaad sprake is van mediatie. Deze aanpak is bijzonder populair aangezien ze makkelijk uitvoerbaar is.
- In dit stuk nemen we de notatie van de auteurs over.

Paddiagrammen

- De schatting van een direct of rechtstreeks effect is een **padcoëfficiënt** (analoog met regressiecoëfficiënten)



- e representeert het direct effect van x_1 op y_3.
- Het indirect effect van x_1 op y_3:

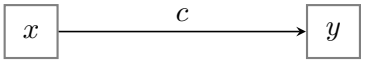
$$(a \times b \times c) + (a \times d)$$

- Het totaal effect van x_1 op y_3: direct + indirect:

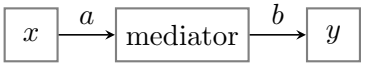
$$e + (a \times b \times c) + (a \times d)$$

Drie mogelijkheden m.b.t. mediatie

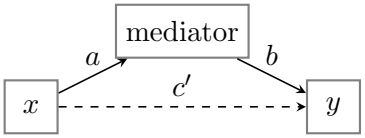
1. Geen mediatie: c = totaal effect van x op y:



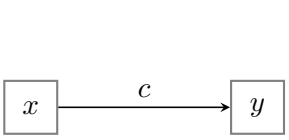
2. Volledige mediatie: $a \times b$ is het gemedieerd effect van x op y

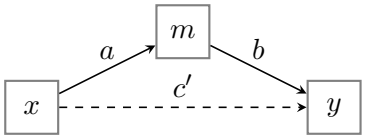


3. Gedeeltelijke mediatie: *totaal* effect: $c' + a \times b$, *direct* effect: c'



6.2 De Baron & Kenny methode





1. Regresseer y op x: is er wel een verband tussen y en x?

   1. $H_0 : c = 0$?

2. Regresseer m op x: is er wel een verband tussen m en x?

   2. $H_0 : a = 0$?

3. Regresseer y op x en m: is er wel een effect van m op y na controle voor x?

   3. $H_0 : b = 0$?

4. Is er na controle van m nog wel een effect van x op y?

   4. $H_0 : c' = 0$?

- Er is sprake van volledige mediatie indien:
  1. $a \neq 0$, $b \neq 0$ en $c \neq 0$,
  2. $c' = 0$
- Er is sprake van gedeeltelijke mediatie indien:
  1. $a \neq 0$, $b \neq 0$ en $c \neq 0$,
  2. $c' < c$
- Bij deze mediatie-analyse maakt men de assumptie van lineaire relaties tussen de verschillende variabelen.
- Het verschil $(c - c')$ hanteert men vaak als een maat voor het *gemedieerd* effect.
- Voor lineaire modellen geldt: onder normale omstandigheden (geen missing values, zelfde covariaten in beide modellen)
$$(c - c') = a \times b.$$
- Stap 1 is eigenlijk overbodig: indien de mediator fungeert als een 'suppressor', dan zal er geen verband zijn tussen x en y, terwijl er wel sprake is van mediatie; een symptoom is dat het teken van $a \times b$ omgekeerd is aan c'

  Een 'suppressor' of onderdrukkende variabele onderdrukt of verbergt de samenhang tussen 2 variabelen zodat deze geen verband met elkaar lijken te hebben.

  Omwille van deze reden gebruiken we in deze cursus de Baron & Kenny methode als volgt: we toetsen het totale effect c in stap 1, maar ongeacht het resultaat gaan we over naar de volgende stappen, omdat er nog steeds sprake kan zijn van mediatie. We voeren dus altijd de 4 stappen uit en gaan op basis van de resultaten na of er (1) evidentie is voor mediatie (stappen 2 en 3) (2) indien wel, of er sprake is van gedeeltelijke of volledige mediatie of dat er sprake is van suppressie (stappen 1 en 4).

6.3 De Sobel test

Enkele problemen met de Baron & Kenny methode:

1. Conceptueel: het is een indirecte wijze om het mediatie effect na te gaan: we focussen op het verschil tussen c en c', terwijl we eigenlijk geïnteresseerd zijn in a en b.
2. Statistisch:

  - in vaak voorkomende situaties: weinig power (MacKinnon et al., 2002)
  - meerdere toetsen na elkaar, dus een inflatie van Type I fouten
  - mogelijkheid tot inconsistente resultaten over de verschillende regressies heen:
    (a) a en b kunnen beide significant zijn, maar c' is niet kleiner dan c.
    (b) c' is veel kleiner dan c, maar a en b zijn niet significant
  - We bekomen enkel evidentie voor mediatie maar geen zekerheid; er zijn mogelijks andere modellen mogelijk! (bvb. m en y wisselen van plaats)

- Meer directe test: is het zo dat x een effect heeft op m ($= a$) en m een effect heeft op y ($= b$); met andere woorden, is $a \times b$ verschillend van nul?

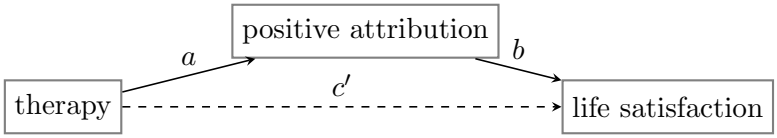
$$H_0 : a \times b = 0$$

  Deze test is de Sobel test.

- Probleem: de steekproevenverdeling van $(a \times b)$ is doorgaans niet normaal verdeeld (vooral bij kleinere steekproefgroottes).
- Via bootstrap (subsamplen uit een dataset) is het mogelijk om een 'empirische' steekproevenverdeling voor $(a \times b)$ te berekenen.
- MacKinnon et al. (2004) hebben een exacte verdeling afgeleid voor de steekproevenverdeling van $(a \times b)$.

6.4 Voorbeeld

We beschouwen de fictieve data uit de paper van Preacher & Hayes (2004) waarbij men geïnteresseerd is in het effect van een nieuwe cognitieve therapie (`therapy`, nominaal) op de levenstevredenheid (`satisfaction`, van intervalniveau) na pensionering (zie sectie 1.2.4). De onderzoeksvraag is of het effect van de cognitieve gedragstherapie gemedieerd wordt door de positiviteit van de attributies (`attribution`, van intervalniveau).



We voeren de 4 stappen van de Baron & Kenny methode uit om na te gaan of er al dan niet sprake is van mediatie.

- Stap 1

```
> satis_fitxy<-lm(satis~therapy,data=satisfaction)
> summary(satis_fitxy)

Call:
lm(formula = satis ~ therapy, data = satisfaction)

Residuals:
Min      1Q  Median      3Q     Max
-1.5669 -0.7319  0.3171  0.5121  1.3131

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3271     0.2233  -1.465   0.1541
therapy       0.7640     0.3058   2.498   0.0186 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8356 on 28 degrees of freedom
Multiple R-squared:  0.1823,Adjusted R-squared:  0.1531
F-statistic: 6.242 on 1 and 28 DF,  p-value: 0.01862
```

De p-waarde van de toets voor $H_0 : c = 0$ (totaal effect van therapie op satisfactie) is gelijk aan 0.019; we kunnen H_0 bijgevolg verwerpen op het 5% significantieniveau.

- Stap 2

```
> satis_fitxm<-lm(attrib~therapy,data=satisfaction)
> summary(satis_fitxm)

Call:
lm(formula = attrib ~ therapy, data = satisfaction)

Residuals:
Min      1Q  Median      3Q     Max
-1.4864 -0.5939 -0.0650  0.2611  1.8850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -0.3536     0.2184  -1.619   0.1166
therapy       0.8186     0.2990   2.738   0.0106 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8171 on 28 degrees of freedom
Multiple R-squared:  0.2111,Adjusted R-squared:  0.183
F-statistic: 7.494 on 1 and 28 DF,  p-value: 0.01064
```

De p-waarde voor de toets $H_0 : a = 0$ (effect van therapie op attributie) is gelijk aan 0.011; we kunnen H_0 bijgevolg verwerpen op het 5% significantieniveau.

- Stap 3 + 4

```
> satis_fitxmy<-lm(satis~therapy+attrib,data=satisfaction)
> summary(satis_fitxmy)

Call:
lm(formula = satis ~ therapy + attrib, data = satisfaction)

Residuals:
Min       1Q   Median       3Q      Max
-1.36527 -0.60758  0.02416  0.54923  1.29091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1843     0.2185  -0.844    0.406
therapy       0.4334     0.3221   1.346    0.190
attrib        0.4039     0.1808   2.234    0.034 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7818 on 27 degrees of freedom
Multiple R-squared:  0.3098,Adjusted R-squared:  0.2587
F-statistic:  6.06 on 2 and 27 DF,  p-value: 0.006697
```

De p-waarde voor de toets $H_0 : b = 0$ (effect van attributie op satisfactie, na controle voor therapie) is gelijk aan 0.034; we kunnen H_0 bijgevolg verwerpen op het 5% significantieniveau. Op basis van de tot nu toe uitgevoerde stappen kunnen we besluiten dat er volgens de Baron & Kenny methode aanwijzingen voor mediatie zijn.

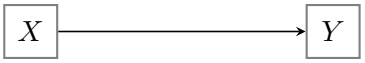
> Verder zien we dat, na correctie voor attributie, het effect van therapie op satisfactie niet statistisch significant is op het 5% significantieniveau (p-waarde gelijk aan 0.19). We hebben dus evidentie voor volledige mediatie.

We kunnen afleiden dat $\hat{c} = 0.76$ (i.e. het geschatte totaal effect van therapie op satisfactie) en $\hat{c}' = 0.43$ (i.e. het geschatte effect van therapie op satisfactie na correctie voor attributie). Het geschatte gemedieerde effect is bijgevolg $0.76 - 0.43 = 0.33$.

We bekijken in een volgende stap enkel de resultaten voor de Sobel test. In R krijgen we voor de bootstrap methode het volgende 95% betrouwbaarheidsinterval voor $a \times b$: $[0.081, 0.803]$ (we gaan niet dieper in op de manier waarop deze resultaten bekomen worden). Er is geen sprake van mediatie indien $a \times b = 0$. Aangezien het betrouwbaarheidsinterval 0 niet omvat, kunnen we $H_0 : a \times b = 0$ verwerpen op het 5% significantieniveau en aannemen dat er sprake is van mediatie.

7 De 'derde' variabele

Veronderstel dat we geïnteresseerd zijn in het effect van een predictor X op een uitkomst Y.



Een andere predictor Z (de 'derde' variabele) kan X en/of Y of de relatie tussen X en Y op verschillende manieren beïnvloeden. We bekijken hier enkele mogelijkheden, merk op dat Z een set van predictoren kan voorstellen.

7.1 Confounding

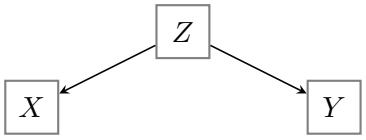
Regressie wordt vaak gebruikt om causale verbanden te onderzoeken, maar causale verklaringen zijn niet zomaar gerechtvaardigd.

Regressiemodellen beschrijven associaties die niet noodzakelijk interpreteerbaar zijn als causale effecten. Causale besluitvorming is mogelijk onder specifieke assumpties die niet getest kunnen worden a.d.h.v. data maar die soms gegarandeerd worden door het design (denk aan experimentele versus observationele designs).

Voorbeeld

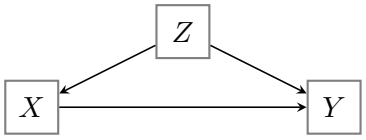
Via een regressie-analyse stelt men vast dat landen met een groter aantal televisies per persoon (X) een hogere levensverwachting (Y) hebben. Dit impliceert niet dat een hoger aantal televisies per persoon een hogere levensverwachting veroorzaakt. X is een indicator voor de welvaart. Welvaart beïnvloedt zowel X als Y. Wat men waarneemt is dus niet het rechtreekse verband tussen predictor en uitkomst.

In bovenstaand voorbeeld is er sprake van confounding.



Er is geen verband tussen X en Y, enkel een associatie omwille van de gemeenschappelijke oorzaak Z. Men spreekt van een spurieuze assocatie.

Confounding betekent niet noodzakelijk dat X geen effect heeft op Y, enkel dat minstens een deel van de associatie tussen X en Y verklaard wordt door het feit dat Z een confounder is voor de relatie tussen X en Y.



Om het effect van X op Y conditioneel op Z te schatten, moet Z als predictor in het regressiemodel opgenomen worden. Zie ook verder in sectie 7.4.

7.2 Moderatie

Het effect van X op Y hangt af van Z. Z is moderator voor de relatie tussen X en Y.

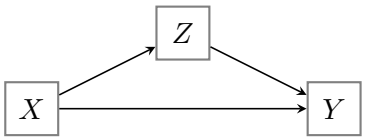
Voorbeeld

Cognitieve gedragstherapie is efficiënter bij adolescenten dan bij volwassenen.

Moderatie betekent dat er een interactie is tussen X en Z. De term 'interactie' is algemener aangezien de term 'moderatie' expliciet een onderscheid maakt tussen de rol van de variabelen. Hier: we zijn geïnteresseerd in het effect van X (predictor) op Y, maar deze relatie wijzigt naargelang het niveau van moderator Z. Zie sectie 5 voor het opnemen van interacties in een regressiemodel.

7.3 Mediatie

X is oorzaak van Z, Z is oorzaak van Y (Z is een *mediator* voor de relatie tussen X en Y)



Voorbeeld

Opleidingsniveau bepaalt motivatie en motivatie bepaalt leergierigheid.

Zie sectie 6 over het opsplitsen van het totaal effect van X op Y in een indirect of gemedieerd effect en een direct effect.

Merk op dat de data niet toelaten om de richting van de relaties na te gaan. Dit impliceert ook dat de data geen uitsluitsel kunnen bieden over het feit of Z een confounder of mediator is. De richting van de relaties kan gebaseerd zijn op onderliggende theorieën of gegarandeerd worden via het design via temporele ordening (X is gemeten op tijdstip 1, Z is gemeten op een later tijdstip 2 en Y is laatst gemeten op tijdstip 3).

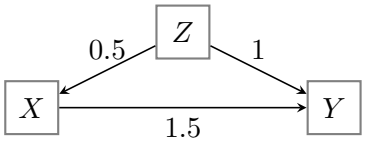
7.4 Omitted variable bias

De term 'omitted variable bias' verwijst naar de vertekening of bias bij het schatten van het effect van X op Y wanneer een variabele Z die geassocieerd is met zowel X als Y niet in het model opgenomen is.

Veronderstel bvb. dat Z een confounder is voor de relatie tussen X en Y en dat geldt voor $i = 1, \ldots, n$:

$$X_i \quad = \quad 1 + 0.5 \times Z_i + \varepsilon_i^* \text{ met } \varepsilon_i^* \sim N(0, 1)$$

$$Y_i \quad = \quad 1 + 1.5 \times X_i + Z_i + \varepsilon_i \text{ met } \varepsilon_i \sim N(0,1)$$

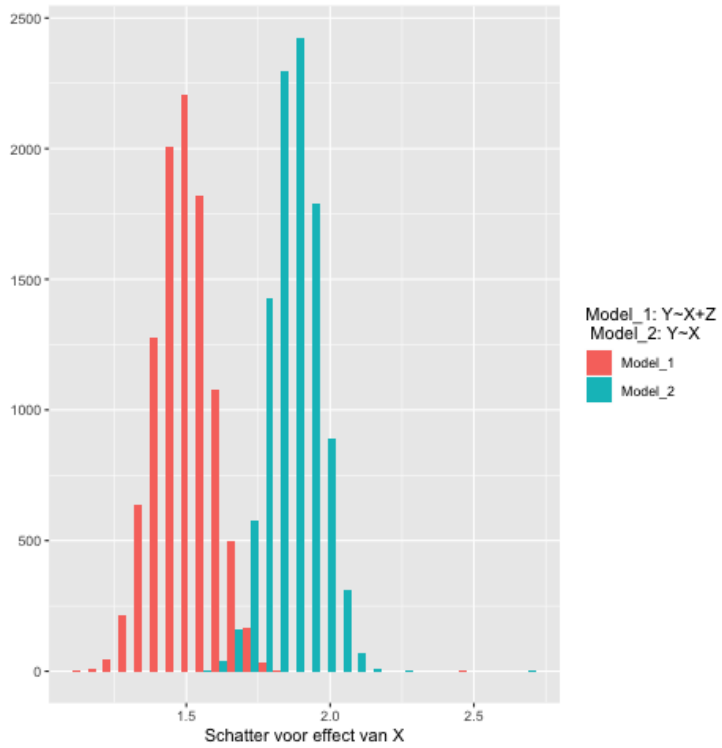


In dit voorbeeld kennen we de onderliggende waarheid, namelijk dat de regressieparameter die het effect van X op Y weergeeft gelijk is aan 1.5.

In een volgende stap genereren we 10 000 datasets waarbij X en Z vast zijn en gegenereerd volgens bovenstaande model en waarbij in iedere dataset de uitkomst Y volgens bovenstaand model gegenereerd is.

Voor iedere gegenereerde dataset schatten we 2 regressiemodellen. Model 1 is correct en bevat zowel X en Z als predictoren; model 2 bevat enkel X en is dus verkeerd gespecifieerd.

Onderstaande figuur toont voor beide modellen de steekproevenverdeling van de schatter voor het effect van X op Y.

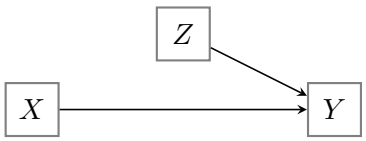


We zien dat de schatter op basis van model 2 vertekend is, het gemiddelde van de steekproevenverdeling is niet gelijk aan 1.5. De schatter op basis van model 1 is onvertekend. De reden voor de vertekening in model 2 is dat door het weglaten van Z, de assumptie van geen correlatie tussen X en de fouttermen geschonden wordt. De verdeling van de fouttermen bevat nu immers ook nog variatie verklaard door Z. Het geschatte effect van X op Y in model 2 bevat ook een deel van de associatie van Z met zowel X als Y. De R-code voor deze simulatie-oefening kan teruggevonden worden in `confounding.R` (optioneel).

Corrigeren voor confounders is bijgevolg belangrijk. In observationele studies kan het aantal covariaten en potentiële confounders echter groot zijn. Dit maakt het moeilijk om de relatie met de uitkomst goed te modelleren. Wanneer de associatie tussen uitkomst en confounder verkeerd gespecifieerd is, kan de schatter voor het effect van X alsnog vertekend zijn of kunnen toetsen voor het effect van X mogelijks niet valide zijn. Daarenboven moet men zich ook bewust zijn van de mogelijke aanwezigheid van ongemeten confounders die niet in het model opgenomen kunnen worden.

Is correctie voor Z altijd een goede zaak? Dit hangt af van de relatie tussen X, Y en Z. We beschouwen nu een aantal andere mogelijkheden dan confounding.

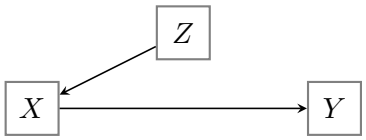
- In onderstaande situatie is er geen associatie tussen X en Z maar wel een effect van Z op Y.



  Wanneer Z niet in het model opgenomen wordt, zal het effect van X op Y onvertekend geschat kunnen worden. De precisie zal echter lager zijn. Algemeen geldt dat correctie voor predictoren van de uitkomst de precisie van schatters ten goede komt (kleinere standaardfouten). De verklaarde kwadratensom zal immers toenemen.

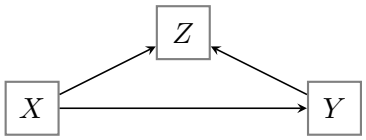
- Wanneer Z niet geassocieerd is met X en niet met Y, dan heeft het al dan niet opnemen van Z in het model geen invloed op de schatter van X op Y, deze zal in beide gevallen onvertekend zijn. Er is wel een verlies aan vrijheidsgraden als Z in het model opgenomen wordt, wat ook de precisie van de schatter beïnvloedt.

- In onderstaande situatie beïnvloedt Z de predictor X maar bestaat er geen assocatie tussen Z en Y.



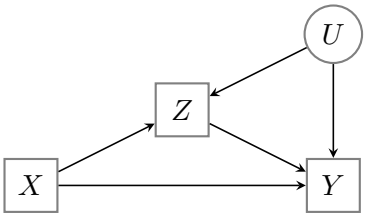
In dit geval is het niet wenselijk om Z in het model op te nemen: er kunnen o.a. problemen optreden met collineariteit (zie Statistiek II). Dit betekent dat de variantie van de schatter voor het effect van X zal toenemen. Bovendien kan er in bepaalde gevallen ook vertekening optreden.

- In onderstaande situatie wordt Z beïnvloed door zowel X als Y.



Wanneer Z in het lineair regressiemodel opgenomen wordt, zal het effect van X op Y vertekend worden. Dit is bekend als *collider bias* of *selection bias*. In dat geval is correctie voor Z dus niet wenselijk.

Een andere situatie waar ditzelfde probleem zich voordoet, is als volgt:



Hierbij stelt U alle ongemeten gemeenschappelijke oorzaken van Z en Y voor (ongemeten confounders). Dit probleem van collider bias verklaart (gedeeltelijk) de 'obesity paradox'. Men stelde vast dat bij patiënten met cardiovasculaire aandoeningen een hoog BMI een beschermende invloed leek te hebben: er werd een lagere mortaliteit vastgesteld dan bij patiënten met een lager BMI. In bovenstaande figuur stelt X BMI voor, Y mortaliteit en Z het al dan niet hebben van een cardiovasculaire aandoening. Veronderstel dat een hoger BMI geassocieerd is met een hogere kans op een cardiovasculaire aandoening. Wanneer men dan gaat kijken bij personen met een cardiovasculaire aandoening (en dus conditioneert op Z), zullen personen met een laag BMI bijgevolg de cardiovasculaire aandoening waarschijnlijk gekregen hebben door een andere oorzaak (U) dan BMI. Deze oorzaak kan bijvoorbeeld een ernstige, onderliggende aandoening zijn die op zijn beurt

een hogere mortaliteit veroorzaakt wat het 'beschermende' effect van een hoog BMI verklaart.

Bovenstaande situatie toont ook dat mediatie-analyses vertekende resultaten opleveren als er ongemeten confouders zijn voor de relatie tussen mediator (hier voorgesteld door Z) en uitkomst.

Een manier om deze vorm van vertekening te voorkomen, is om ook U te meten en in het model op te nemen.

In het algemeen geldt dat men heel voorzichtig moet zijn wanneer men gaat corrigeren voor variabelen die beïnvloed worden door X.

De assumpties die men bij mediatie-analyses in sectie 6 maakt, zijn als volgt:

(A1) geen ongemeten confounding voor de X - M relatie

(A2) geen ongemeten confounding voor de X - Y relatie

(A3) geen ongemeten confounding voor de M - Y relatie

(A4) geen confounders voor de M - Y relatie die beïnvloed zijn door X

Wanneer X gerandomiseerd is (bvb. at random toewijzing aan condities van een behandeling), is voldaan aan (A1) en (A2). Zelfs als X gerandomiseerd is, kunnen er nog confounders zijn voor de $M - Y$ relatie!

8 Analyse van experimentele designs

8.1 Het experiment

8.1.1 Designs

Bij een **zuiver experiment** is voldaan aan 3 essentiële kenmerken:

- de veronderstelde oorzaak van een gevolg wordt gemanipuleerd;
- willekeurige toewijzing van de deelnemers aan condities (randomisering);
- alle andere factoren worden constant gehouden.

Het doel is het onderzoeken van causale relaties. Wanneer aan de voorwaarden van een zuiver of **gerandomiseerd** experiment voldaan is, kunnen eventuele verschillen/veranderingen in de

afhankelijke variabele toegeschreven worden aan de verschillen in de niveaus van de onafhankelijke (gemanipuleerde) variabele.

Als aan één van de 3 genoemde eisen niet voldaan is (typisch randomisering), dan is er sprake van een **quasi-experiment**.

Een basisexperiment kan bvb. als volgt zijn:

- Verdeel deelnemers in 2 groepen door opgooien munststuk: experimentele en controlegroep
- Laat beide groepen andere procedure volgen = experimentele manipulatie
- Bekijk nadien het verschil tussen beide groepen m.b.t. de uitkomst of afhankelijke variabele

Dit basisexperiment kan als volgt voorgesteld worden:

$$\begin{array}{ccc} R & X & O \\ R & & O \end{array}$$

Hierbij wordt de volgende notatie gehanteerd:

- R: random toewijzing
- O: observatie
- X: experimentele behandeling / gebeurtenis

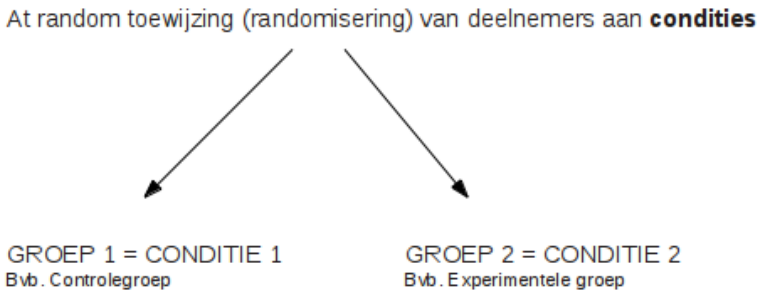
Twee belangrijke onderzoeksontwerpen / designs:

- **between-subjects design** (tussen-proefpersonenontwerp): verschillende deelnemers toegewezen aan verschillende condities

  Synoniemen: between-participants, independent-groups, unrelated groups, uncorrelated groups design

  Voorbeeld

  vergelijken van aantal fouten bij het invoeren van gegevens op computer bij het beluisteren van harde popmuziek in de ene groep en met 'witte ruis' van hetzelfde volume in de andere groep

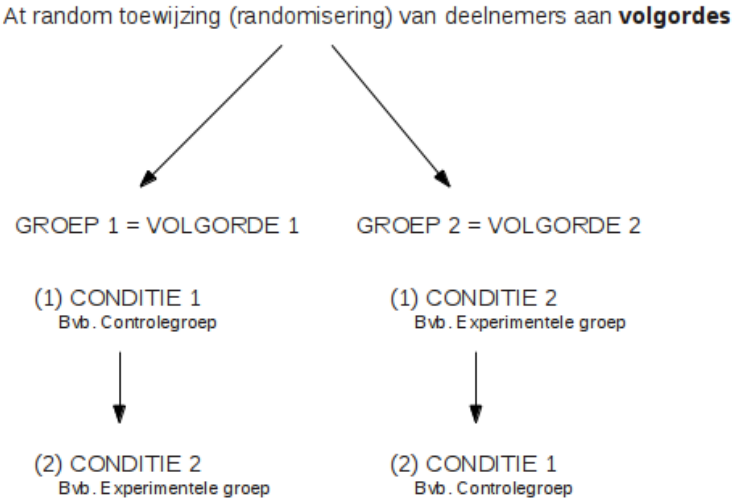


- **within-subjects design** (binnen-proefpersonenontwerp): dezelfde deelnemers toegewezen aan alle (of enkele) condities

  Synoniemen: within-participants, repeated-measures, dependent-groups, related-groups, correlated-groups design

  Voorbeeld

  aantal tikfouten bij beluisteren muziek en zonder muziek



Belang van onderscheid: verschillende statistische technieken om data te analyseren!

De onderstaande tabel (tabel 4.1, pag. 77 uit Kline, 2009) geeft een overzicht van de hoofdtypes van experimentele designs.

TABLE 4.1. Major Types of Experimental Designs

Type	Representation					
Basic	R		X	O		
	R			O		
Factorial	R		X_{A1B1}	O		
	R		X_{A1B2}	O		
	R		X_{A2B1}	O		
	R		X_{A2B2}	O		
Pretest–posttest	R	O_1	X	O_2		
	R	O_1		O_2		
Solomon Four Group	R	O_1	X	O_2		
	R	O_1		O_2		
	R		X	O_2		
	R			O_2		
Switching replications	R	O_1	X	O_2		O_3
	R	O_1		O_2	X	O_3
Crossover	R	O_1	X_A	O_2	X_B	O_3
	R	O_1	X_B	O_2	X_A	O_3
Longitudinal	R	$O \ldots O$	X	O	$O \ldots O$	
	R	$O \ldots O$		O	$O \ldots O$	

Note. R, random assignment; O, observation; X, treatment.

In deze cursus bekijken we de analyse van een voorbeeld van een factorieel design binnen een between-subjects design.

Een factorieel design is een design waarin alle mogelijke combinaties van niveaus van twee (of meer) variabelen voorkomen. Er kan een onderscheid gemaakt worden tussen een *gebalanceerd* factorieel design (even grote groepen) en een *niet-gebalanceerd* factorieel design. In een gebalanceerd design zijn de hoofd- en interactie-effecten allemaal onafhankelijk (dit betekent dat de effecten volledig gescheiden kunnen worden). Daarom noemt men dergelijke designs vaak *orthogonale* designs.

In toegepast onderzoek valt het echter vaak voor dat factoriële designs niet gebalanceerd (*niet-orthogonaal*) zijn.

8.1.2 Voorbeeld: motivatie

In een fictief experiment wenst men de motivatie van personen bij het uitvoeren van taken te onderzoeken. De onderzoeker wenst de invloed na te gaan van een externe beloning (het

toekennen van een geldbedrag). Het effect van de taakinteresse wordt ook onderzocht: er worden vervelende, matig interessante en interessante opdrachten aangeboden.

- De uitkomst of afhankelijke variabele Y is het aantal taken dat een persoon succesvol uitvoert, we noemen dit de `score`.
- De nominale variabelen `beloning` en `taakinteresse` bestaan uit respectievelijk 2 niveaus (het al dan niet toekennen van een beloning) en 3 niveaus (vervelend, matig interessant en interessant).

Veronderstel dat er 24 proefpersonen deelnemen aan het experiment m.b.t. motivatie. Bij dit experiment zijn er 6 (2×3) verschillende combinaties van `beloning` en `taakinteresse` (onderzoekscondities of cellen).

Er worden lukraak 4 personen aan elke cel toegewezen.

Niveaus van `beloning`: geen geldbeloning (1) en wel geldbeloning (2)

Niveaus van `taakinteresse`: vervelend (1), matig interessant (2) en interessant (3)

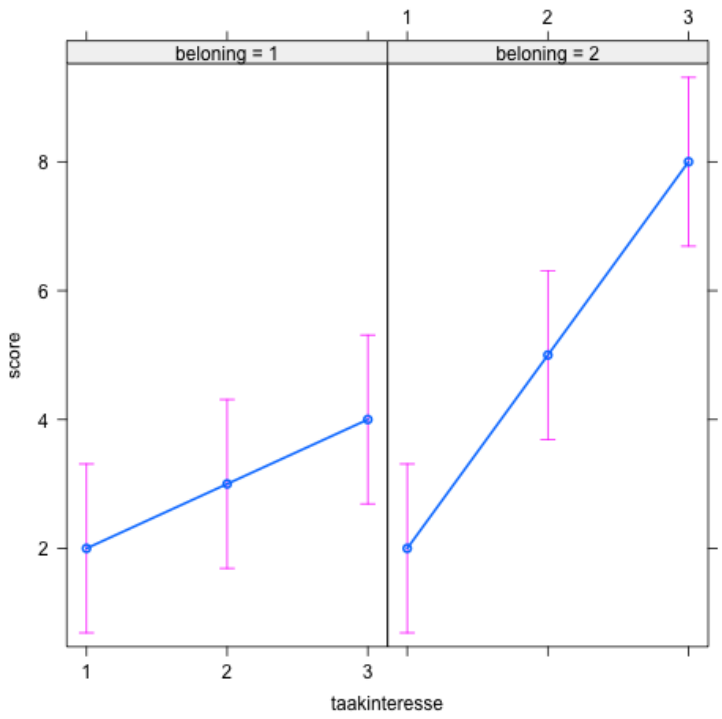
De onderstaande tabel bevat de resultaten:

score	beloning	taakinteresse	score	beloning	taakinteresse
3	1	1	4	2	1
2	1	1	2	2	1
2	1	1	1	2	1
1	1	1	1	2	1
4	1	2	7	2	2
3	1	2	5	2	2
3	1	2	4	2	2
2	1	2	4	2	2
6	1	3	9	2	3
4	1	3	9	2	3
3	1	3	8	2	3
3	1	3	6	2	3

In dit voorbeeld correspondeert een *cel* of *onderzoeksconditie* met 1 van de combinaties van de niveaus van taakinteresse en beloning. Het aantal observaties in 1 cel noemt men de *celfrequenties*.

Aangezien alle celfrequenties gelijk zijn aan elkaar, is de opzet **gebalanceerd**.

Onderstaande effectenplot toont de geobserveerde steekproefgemiddeldes binnen de 6 condities.



We observeren dat wanneer een geldbeloning toegekend wordt bij matig interessante en interessante taken, er een grotere gemiddelde score is. Dit is niet het geval bij een vervelende taak waar de gemiddelde score hetzelfde is bij het al dan niet toekennen van een beloning. Dit wijst op een mogelijke interactie tussen het al dan niet toekennen van een beloning en taakinteresse. We observeren ook dat het effect van taakinteresse groter wordt bij het toekennen van een geldbeloning. Op deze figuur kunnen we enkel aflezen of er aanwijzingen zijn voor positieve/negatieve effecten van de factoren en/of interacties. Via statistische toetsen gaan we na of deze effecten en interactie-effecten statistisch significant zijn.

Zowel de code die hoort bij de analyses (`motivatie.R`) als de data (`beloning.csv`) zijn terug te vinden op Ufora.

8.2 Variantie-analyse

8.2.1 Terminologie en werkwijze

Lineaire regressie met enkel nominale onafhankelijke variabelen wordt ook variantie-analyse genoemd.

Variantie-analyse is een verzamelnaam voor een geheel van methoden en technieken die nagaan of het gemiddelde van een uitkomst (minimaal vereiste meetniveau = intervalniveau) verschilt voor verschillende groepen van observaties. Variantie-analyse gaat na of (eventueel) gevonden verschillen tussen de gemiddelden gerelateerd zijn aan het verschil in groepen. De Engelse benaming voor deze techniek luidt Analysis of Variance en wordt vaak afgekort door **anova**.

Variantie-analyse is een procedure die sterk verweven is met het experimenteel onderzoek. De data die met een variantie-analyse geanalyseerd worden, kunnen echter zowel via een experimenteel onderzoek als via een observationele studie verzameld zijn. Bij niet-experimentele designs wordt dan gekeken naar een verschil tussen groepen die niet op experimentele basis samengesteld zijn.

Indien men in de analyse slechts één enkele factor (i.e. onafhankelijke variabele van nominaal niveau) beschouwt, spreekt men van **enkelvoudige** variantie-analyse of **eenwegs**variantie-analyse (Engels: 'oneway anova').

Indien men in de analyse rekening houdt met meerdere factoren (bvb. `beloning` en `taakinteresse`) spreekt men van **meervoudige** variantie-analyse of **meerwegs**variantie-analyse (Engels: 'multiway anova'), of kortweg variantie-analyse.

We kunnen de volgende types factoren onderscheiden:

- **tussensubject factor**: een factor waarbij *groepen* onderscheiden worden.
- **binnensubject factor**: een factor waarmee metingen bij eenzelfde subject onderscheiden kunnen worden (bvb. score wordt bij eenzelfde subject op 4 tijdsmomenten gemeten.) (Engels: 'repeated measures')

Hoewel we ons in deze cursus beperken tot between-subjects factoren, kunnen dergelijke technieken alle variaties in factoriële designs aan: zo is het mogelijk om within-subjects en between-subjects factoren in hetzelfde model te beschouwen. Dit komt in toekomstige cursussen aan bod.

Veronderstel dat er 2 factoren A en B zijn die uit respectievelijk I en J niveaus bestaan.

Aan de hand van een tweewegsvariantie-analyse gaat men na hoeveel van de variantie in de uitkomst Y verklaard wordt door de hoofdeffecten van de 2 factoren en hun interactie. Men beschouwt hierbij de volgende kwadratensommen:

- de kwadratensom die samenhangt met de hoofdeffecten van factor A,
- de kwadratensom die samenhangt met de hoofdeffecten van factor B,
- en de kwadratensom die de effecten i.v.m. de interactie van de factoren A en B groepeert.

De effecten worden getoetst aan de hand van F-toetsen (cfr. modelvergelijkingstoetsen).

Het aantal vrijheidsgraden dat bij de kwadratensommen hoort, is als volgt:

Bron	Kwadratensom	Vrijheidsgraden
Model	$\mathrm{SS}_{\mathrm{Model}}$ (*SSR*)	$\mathrm{df}_{\mathrm{Model}} = I \times J - 1$
Factor A	SS_{A}	$\mathrm{df}_{\mathrm{A}} = I - 1$
Factor B	SS_{B}	$\mathrm{df}_{\mathrm{B}} = J - 1$
Interactie $A \times B$	$\mathrm{SS}_{\mathrm{AB}}$	$\mathrm{df}_{\mathrm{AB}} = (I-1) \times (J-1)$
Error	$\mathrm{SS}_{\mathrm{Error}}$ (*SSE*)	$\mathrm{df}_{\mathrm{Error}} = n - I \times J$
Totaal	$\mathrm{SS}_{\mathrm{Totaal}}$ (*SST*)	$\mathrm{df}_{\mathrm{Totaal}} = n - 1$

met n het totaal aantal observaties. Dit is equivalent met het overeenkomstig aantal vrijheidsgraden bij de modelvergelijkingstoetsen aan de hand van hulpveranderlijken bij lineaire regressie.

Hoewel lineaire regressie en variantie-analyse technisch equivalent zijn, is de terminologie die bij variantie-analyse gehanteerd wordt vaak anders dan bij lineaire regressie, bvb. $\mathrm{SS}_{\mathrm{Tussen}}$ (*between*) i.p.v. SSR en $\mathrm{SS}_{\mathrm{Binnen}}$ (*within*) i.p.v. SSE. Het basisidee achter de F-toetsen is immers als volgt: men vergelijkt de variabiliteit van de uitkomst tussen de groepen met de variabiliteit binnen de groepen. Wanneer F veel groter is dan 1, betekent dit dat de variabiliteit tussen groepen te groot is om de hypothese van gelijke groepsgemiddelden te

ondersteunen. Verder spreekt men van sigma-restricties i.p.v. effect-codering en van GLM-restricties i.p.v. dummy-codering.

Bij een tweewegsvariantie-analyse gaat men doorgaans als volgt te werk:

1. Eerst wordt er getoetst of er een interactie bestaat tussen beide factoren.
2. Wanneer men kan besluiten dat er geen belangrijke / significante interacties aanwezig zijn, gaat men over tot het toetsen van de hoofdeffecten.

De assumptie van een constante residuele variantie komt in deze context overeen met gelijke varianties van de uitkomst (afhankelijke variabele) in de verschillende groepen die gevormd worden door de combinaties van de verschillende niveaus van de factoren. De assumptie van normaal verdeelde residuen komt overeen met een normaal verdeelde uitkomst in iedere groep.

Bij een **factorieel** design dat **gebalanceerd** is, kan men aantonen dat

$$\mathrm{SS}_{\mathrm{Model}} = \mathrm{SS}_{\mathrm{A}} + \mathrm{SS}_{\mathrm{B}} + \mathrm{SS}_{\mathrm{AB}}$$

Dit wordt ook een orthogonale decompositie genoemd: een decompositie waarbij de kwadratensommen van de componenten sommeren tot de totale kwadratensom en waarbij de som van de vrijheidsgraden van de componenten gelijk is aan het aantal vrijheidsgraden die horen bij de totale kwadratensom. In dit geval kunnnen de 3 effecten onafhankelijk van elkaar geschat worden (er is geen overlap tussen de effecten). Merk op dat we hier wel uitgaan van effect-codering. In dit geval zijn de Type I, Type II en Type III kwadratensommen equivalent.

Orthogonaliteit is een goede eigenschap maar komt in de regel enkel voor wanneer de predictoren (designmatrix) volledig zelf door de onderzoeker vastgelegd kunnen worden zoals in een experiment. Bij observationele data heeft men geen directe controle over de opzet in de designmatrix, wat vaak de bron is van moeilijkheden met interpretatie bij niet-experimentele data.

Hoewel veel (maar niet alle) experimenten ontworpen zijn met de intentie even grote groepen te creëren om op die manier een gebalanceerd design te bekomen, komt het in de praktijk voor dat deelnemers afhaken, data verloren gaan, etc. Bij complexe factoriële designs heeft een niet-gebalanceerde (niet-orthogonale) opzet tot gevolg dat de hoofd- en interactie-effecten niet langer onafhankelijk zijn. Het is dan niet meer mogelijk om de verklaarde kwadratensommen zonder meer op te splitsen over hoofd- en interactie-effecten. Dit betekent dat de verklaarde kwadratensom niet meer gelijk is aan de som van de kwadratensommen geassocieerd met de hoofd- en interactie-effecten. Er is geen unieke manier meer om de kwadratensom van een effect te bepalen en dus is er een verschil tussen de de Type I, Type II en Type III kwadratensommen.

8.2.2 Voorbeeld: motivatie

Beschouw het experiment rond motivatie waarbij gekeken wordt naar het effect van beloning en taakinteresse. Laat μ_{ij} de verwachte score voorstellen binnen groep i van factor A (`beloning`, $i = 1, 2$) en binnen groep j van factor B (`taakinteresse`, $j = 1, 2, 3$). We voeren de volgende notatie in voor de marginale gemiddelden:

- $\mu_{i.} = \frac{\sum_{j=1}^{J} \mu_{ij}}{J}$, de verwachte waarde voor de gemiddelde score binnen niveau i van factor A, over de niveaus van factor B heen
- $\mu_{.j} = \frac{\sum_{i=1}^{I} \mu_{ij}}{I}$, de verwachte waarde voor de gemiddelde score binnen niveau j van factor B, over de niveaus van factor A heen
- $\mu_{..} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}}{I \times J}$, het globaal gemiddelde

Via de 'dot-notatie' duiden we met een puntje aan over welke factor(en) het gemiddelde berekend wordt.

	Taakinteresse			
Beloning	vervelend ($j = 1$)	matig interessant ($j = 2$)	interessant ($j = 3$)	rijgemiddelde
geen ($i = 1$)	μ_{11}	μ_{12}	μ_{13}	$\mu_{1.}$
wel ($i = 2$)	μ_{21}	μ_{22}	μ_{23}	$\mu_{2.}$
kolomgemiddelde	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\mu_{..}$

De toetsen voor de hoofd- en interactie-effecten kunnen als volgt geschreven worden in functie van de cel- en marginale gemiddelden:

- $H_{01} : \mu_{1.} = \mu_{2.} = \mu_{..}$
  Deze hypothese stelt dat het hoofdeffect van factor A (`beloning`) nul is: de gemiddelde score is hetzelfde voor elk niveau van A.
- $H_{02} : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{..}$
  Deze hypothese stelt dat het hoofdeffect van factor B (`taakinteresse`) nul is: de gemiddelde score is hetzelfde voor elk niveau van B.
- H_{03} : Voor alle i, i' en alle $j, j' : \mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'}$
  Deze hypothese stelt dat er geen interactie is tussen beide factoren: het effect van factor A hangt niet af van het niveau van factor B en vice versa.

We bekijken eerst de parameterschattingen van het model; we maken hierbij gebruik van dummy-codering.

```
> fit1_beloning<-lm(score~beloning+taakinteresse+beloning:taakinteresse,data=motivatie)
> summary(fit1_beloning)

Call:
lm(formula = score ~ beloning + taakinteresse + beloning:taakinteresse,
data = motivatie)

Residuals:
Min     1Q Median     3Q    Max
-2     -1      0      1      2

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.000e+00  6.236e-01   3.207  0.00489 **
beloning2                 -8.158e-16  8.819e-01   0.000  1.00000
taakinteresse2             1.000e+00  8.819e-01   1.134  0.27172
taakinteresse3             2.000e+00  8.819e-01   2.268  0.03589 *
beloning2:taakinteresse2   2.000e+00  1.247e+00   1.604  0.12621
beloning2:taakinteresse3   4.000e+00  1.247e+00   3.207  0.00489 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.247 on 18 degrees of freedom
Multiple R-squared:  0.7879,Adjusted R-squared:  0.729
F-statistic: 13.37 on 5 and 18 DF,  p-value: 1.561e-05
```

Voor zowel beloning als taakinteresse is het eerste niveau het referentieniveau.

Op basis van bovenstaande parameterschattingen kunnen we afleiden dat

$$\begin{aligned}\hat{\mu}_{11} &= 2\\ \hat{\mu}_{23} &= 2+0+2+4=8\end{aligned}$$

etc.

We vinden bijgevolg volgende schattingen voor bovenstaande tabel:

Beloning	Taakinteresse vervelend ($j=1$)	matig interessant ($j=2$)	interessant ($j=3$)	rijgemiddelde
geen ($i=1$)	$\hat{\mu}_{11}=2$	$\hat{\mu}_{12}=3$	$\hat{\mu}_{13}=4$	$\hat{\mu}_{1.}=3$
wel ($i=2$)	$\hat{\mu}_{21}=2$	$\hat{\mu}_{22}=5$	$\hat{\mu}_{23}=8$	$\hat{\mu}_{2.}=5$
kolomgemiddelde	$\hat{\mu}_{.1}=2$	$\hat{\mu}_{.2}=4$	$\hat{\mu}_{.3}=6$	$\mu_{..}=4$

Merk op dat de schattingen voor de verwachte waarden binnen een cel gelijk zijn aan de overeenkomstige steekproefgemiddeldes. Dit is altijd het geval bij een volledig model (i.e. alle factoren en interacties in het model). Aangezien het design hier gebalanceerd is, zijn de geschatte marginale gemiddelden ook gelijk aan de overeenkomstige steekproefgemiddelden.

De predictie $\hat{Y}_{ijk}$ voor observatie k onder conditie (i,j) is gelijk aan het geschatte celgemiddelde dat correspondeert met conditie (i,j).

```
> predict(fit1_beloning)
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
2  2  2  2  3  3  3  3  4  4  4  4  2  2  2  2  5  5  5  5  8  8  8  8
```

We bekijken nu de resultaten van de toetsen voor het interactie-effect en de hoofdeffecten. We gaan hierbij op dezelfde manier te werk als voorheen (`Anova`, Type III kwadratensommen).

```
> fit1_beloning_test<-lm(score~beloning+taakinteresse+beloning:taakinteresse,data=motivatie,
                    contrasts=list(beloning=contr.sum,taakinteresse=contr.sum))
> Anova(fit1_beloning_test,type=3)
Anova Table (Type III tests)

Response: score
                        Sum Sq Df  F value    Pr(>F)
(Intercept)               384  1 246.8571  5.92e-12 ***
beloning                   24  1  15.4286 0.0009861 ***
taakinteresse              64  2  20.5714  2.24e-05 ***
beloning:taakinteresse     16  2   5.1429 0.0171139 *
Residuals                  28 18
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 2 en 18 vrijheidsgraden. De geobserveerde waarde is gelijk aan 5.14. De overeenkomstige p-waarde is gelijk aan 0.017. Dit betekent dat we op basis

van deze gegevens kunnen besluiten dat de interactie tussen beloning en taakinteresse significant is op het 5% significantieniveau (p-waarde kleiner dan 5%).

- Analoog kunnen we voor de toets voor het hoofdeffect van beloning zien dat de geobserveerde toetsingsgrootheid gelijk is aan 15.43. De overeenkomstige p-waarde wordt berekend op basis van de F-verdeling met 1 en 18 vrijheidsgraden en is gelijk aan 0.001. Het hoofdeffect van beloning (i.e. het gemiddeld effect over de niveaus van taakinteresse) is dus statistisch significant, het effect van beloning hangt wel af van het niveau van taakinteresse aangezien de interactie significant is.

- Analoog bekomen we voor de toets voor het hoofdeffect van taakinteresse een geobserveerde toetsingsgrootheid gelijk aan 20.57. De overeenkomstige p-waarde wordt berekend op basis van de F-verdeling met 2 en 18 vrijheidsgraden en is zeer klein. Het hoofdeffect van taakinteresse (i.e. het gemiddeld effect over de niveaus van beloning) is dus statistisch significant, het effect van taakinteresse hangt wel af van het niveau van beloning aangezien de interactie significant is.

- Aangezien het design gebalanceerd is, is de som van de kwadratensommen van de hoofd- en interactie-effecten gelijk aan de verklaarde kwadratensom van het model:

```
# Totale kwadratensom
> sst<-sum((motivatie$score-mean(motivatie$score))^2)
> sst
[1] 132
# Verklaarde kwadratensom
> sst-28
[1] 104
# Som van kwadratensommen van hoofd- en interactie-effecten
> 24+64+16
[1] 104
```

- Hier hebben we factoren 'beloning' (A) en 'taakinteresse' (B) met respectievelijk 2 en 3 niveaus ((2×3)-factorieel design).

  We vinden dat: $\text{SS}_\text{A} = 24$, $\text{SS}_\text{B} = 64$, $\text{SS}_\text{AB} = 16$, $\text{SS}_\text{Error} = 28$ en $\text{SS}_\text{Totaal} = 132$. Aangezien het design gebalanceerd is, geldt dat $\text{SS}_\text{Model} = \text{SS}_\text{A} + \text{SS}_\text{B} + \text{SS}_\text{AB} = 104$. We kunnen hieruit afleiden dat $R^2 = 104/132 = 0.788$. Dit kunnen we ook aflezen uit bovenstaande output (`Multiple R-squared`).

  Bij variantie-analyse wordt R^2 vaak de geschatte *eta-squared*, $\hat{\eta}^2$, genoemd.

$$\eta^2_\text{effect} = \frac{\sigma^2_\text{Model}}{\sigma^2_\text{Totaal}}$$

η^2 wordt geschat als

$$\hat{\eta}^2 = \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Totaal}}}$$

en is equivalent aan R^2.

- Voor de individuele effecten vinden we:

$$\begin{aligned} \hat{\eta}^2_A &= \text{SS}_{\text{A}}/\text{SS}_{\text{Totaal}} = 24/132 = 0.18 \\ \hat{\eta}^2_B &= \text{SS}_{\text{B}}/\text{SS}_{\text{Totaal}} = 64/132 = 0.48 \\ \hat{\eta}^2_{AB} &= \text{SS}_{\text{AB}}/\text{SS}_{\text{Totaal}} = 16/132 = 0.12 \end{aligned}$$

  Deze effecten zijn gelijk aan de eerder geziene sr^2_ℓ (kwadraat semi-partiële correlatie).

- Bij een partieel effect wordt gekeken naar de proportie van verklaarde variantie waarbij gecorrigeerd wordt voor de andere effecten:

$$\text{partiële } \hat{\eta}^2 = \frac{\text{SS}_{\text{effect}}}{\text{SS}_{\text{effect}} + \text{SS}_{\text{Error}}}$$

  bvb.:

$$\text{partiële } \hat{\eta}^2_A = \frac{24}{24 + 28} = 0.46$$

  wat betekent dat het hoofdeffect van A 46% van de residuele variantie verklaart na correctie voor het effect van B en het interactie-effect. Dit komt overeen met de eerder geziene pr^2_ℓ (kwadraat partiële correlatie).

  In de praktijk is het gebruikelijk om bij de analyse van factoriële designs $\hat{\eta}^2$ van de totale effecten te rapporteren en de partiële $\hat{\eta}^2$ voor de individuele effecten. Partiële effecten zijn onderling niet rechtstreeks vergelijkbaar omwille van de verschillende noemers.

  Analoog als voorheen kunnen de (semi)-partiële effecten als volgt opgevraagd worden in R:

```
> etaSquared(fit1_beloning_test,type=3,anova=TRUE)
                          eta.sq eta.sq.part SS df        MS         F            p
beloning                0.1818182   0.4615385 24  1 24.000000 15.428571 9.861125e-04
taakinteresse           0.4848485   0.6956522 64  2 32.000000 20.571429 2.240432e-05
beloning:taakinteresse 0.1212121   0.3636364 16  2  8.000000  5.142857 1.711387e-02
Residuals               0.2121212          NA 28 18  1.555556        NA           NA
```

8.2.3 Contrasten

Men kan mogelijks geïnteresseerd zijn in een verschil tussen specifieke celgemiddelden dat niet via H_{01}, H_{02} of H_{03} getoetst wordt.

- Wanneer er geen interactie is tussen taakinteresse en beloning, maar wel een hoofdeffect van taakinteresse, kunnen de gemiddeldes $\mu_{.1}$, $\mu_{.2}$ en $\mu_{.3}$ paarsgewijs met elkaar vergeleken worden.
- Wanneer er wel een interactie is tussen taakinteresse en beloning, kunnen de gemiddeldes paarsgewijs vergeleken worden binnen ieder niveau van beloning.

  We spreken van een **simpel hoofdeffect**: het effect van een onafhankelijke variabele binnen 1 niveau van de andere onafhankelijke variabele.

Men spreekt van contrasten. Contrasten kunnen ook aangewend worden om specifieke celgemiddelden met elkaar te vergelijken, niet louter om (simpele) hoofdeffecten nader te onderzoeken. Wees echter voorzichtig hierbij: zorg dat deze contrasten betekenisvol zijn en dat je er zeker van bent dat je dergelijke vergelijkingen tussen celgemiddelden wenst te toetsen.

Contrasten kunnen getoetst worden via modelvergelijkingstoetsen zoals gezien in sectie 4.1. Hierbij worden lineaire restricties opgelegd aan de parameters en wordt het model met restricties vergeleken met het model zonder restricties.

Een geheel van lineaire restricties m.b.t. de parameters vormt een algemeen lineaire hypothese (ALH) in regressie. In matrixnotatie is een ALH uit te drukken als

$$\mathbf{L}\boldsymbol{\beta} = \mathbf{c}.$$

Om contrasten te toetsen, moeten we dus een L-matrix opstellen die de lineaire hypothese weergeeft.

Voorbeeld

Stel model A bevat 4 predictoren, in totaal dienen dus 5 regressiecoëfficiënten geschat te worden: β_0, β_1, β_2, β_3 en β_4. Een set bijkomende restricties is bvb. als volgt:

$$\begin{aligned} \beta_1 &= 0 \\ 2\beta_2 &= \beta_1 + \beta_3 \text{ of nog: } \beta_1 - 2\beta_2 + \beta_3 = 0 \\ \beta_2 &= \beta_3 \text{ of nog: } \beta_2 - \beta_3 = 0 \end{aligned}$$

In bovenstaand voorbeeld is:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Elke rij in de L-matrix stelt een specifieke restrictie voor. Afhankelijk van het teken en de waarde voor de regressieparameters in de restricitie worden de rijen van de L-matrix opgesteld. $\boldsymbol{\beta}$ is de vector die de regressieparameters bevat. $\mathbf{c}$ bevat voor elke restrictie de uiteindelijke veronderstelde waarde. Veronderstel dat men in het voorbeeld naar motivatie wenst te toetsen of de gemiddelde score bij een interessante taak zonder geldbeloning gelijk is aan de gemiddelde score bij een vervelende taak met geldbeloning, dan is de hypothese:

$$\mu_{13} = \mu_{21}$$

Dit is equivalent met

$$\mu_{13} - \mu_{21} = 0 \textit{ (hypothese 1)}$$

Bovenstaande hypothese wordt een contrast genoemd: een verschil tussen celgemiddeldes.

Veronderstel dat men wenst te toetsen of de gemiddelde score bij een matig interessante taak zonder geldbeloning gelijk is aan de gemiddelde score bij een vervelende taak, ongeacht beloning, dan is de hypothese:

$$\mu_{12} = \mu_{.1}$$

Dit is equivalent met

$$\mu_{12} - \mu_{.1} = 0$$

Verder is

$$\mu_{.1} = \frac{\mu_{11} + \mu_{21}}{2} = \frac{1}{2}\mu_{11} + \frac{1}{2}\mu_{21}$$

en dus kan de hypothese als volgt als een contrast geschreven worden:

$$\mu_{12} - \frac{1}{2}\mu_{11} - \frac{1}{2}\mu_{21} = 0 \textit{ (hypothese 2)}$$

Indien we bovenstaande hypotheses wensen te toetsen, moeten we het contrast eerst herschrijven in termen van de parameters van het model zodat we de L-matrix kunnen opstellen.

In het bovenstaande model voor motivatie (dummy-codering) hebben we 1 hulpveranderlijke voor beloning (x_b) en 2 hulpveranderlijken voor taakinteresse (x_{t_2}, x_{t_3}). Voor de interactie hebben we 2 hulpveranderlijken, namelijk $x_b \times x_{t_2}$ en $x_b \times x_{t_3}$.

Het model kan dan als volgt geschreven worden:

$$Y_\ell = \beta_0 + \beta_1 x_{\ell b} + \beta_2 x_{\ell t_2} + \beta_3 x_{\ell t_3} + \beta_4 x_{\ell b} x_{\ell t_2} + \beta_5 x_{\ell b} x_{\ell t_3} + \varepsilon_\ell \tag{4}$$

waarbij ℓ de index voor individu ℓ is. Voor beide factoren is het eerste niveau het referentieniveau. Bijgevolg is $x_b = 1$ bij `beloning=2` en 0 bij `beloning=1`. $x_{t_2} = 1$ voor `taakinteresse=2` en 0 elders; $x_{t_3} = 1$ voor `taakinteresse=3` en 0 elders.

We schrijven nu de hypotheses in functie van de parameters van bovenstaand model.

Hypothese 1:

$$\begin{aligned} \mu_{21} &= \beta_0 + \beta_1 \\ \mu_{13} &= \beta_0 + \beta_3 \\ \Rightarrow \mu_{13} - \mu_{21} &= \beta_3 - \beta_1 \end{aligned}$$

Hypothese 2:

$$\begin{aligned} \mu_{11} &= \beta_0 \\ \mu_{21} &= \beta_0 + \beta_1 \\ &\Rightarrow \quad \mu_{.1} = \beta_0 + \frac{1}{2}\beta_1 \\ \mu_{12} &= \beta_0 + \beta_2 \\ \Rightarrow \mu_{12} - \mu_{.1} &= \beta_2 - \frac{1}{2}\beta_1 \end{aligned}$$

Beide hypotheses kunnen geschreven worden als $\mathbf{L}\boldsymbol{\beta} = 0$ waarbij

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{bmatrix}'$$

Hypothese 1:

$$\begin{array}{rcc} & & \begin{array}{cccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{array} \\ \boldsymbol{L} & = & \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

Hypothese 2:

$$\begin{array}{rcc} & & \begin{array}{cccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{array} \\ \boldsymbol{L} & = & \begin{bmatrix} 0 & -1/2 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Beide hypotheses kunnen ook tegelijkertijd getoetst worden d.m.v. een set contrasten, dit betekent dat de hypothese stelt dat $\mu_{13} = \mu_{21}$ én $\mu_{12} = \mu_{.1}$:

$$\boldsymbol{L} = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & -1/2 & 1 & 0 & 0 & 0 \end{bmatrix}$$

In R kunnen dergelijke hypotheses getoetst worden aan de hand van `lht` (package `car`). We moeten hierbij het model zonder restricties en de L-matrix meegeven. Deze laatste is opgesteld op basis van het model zonder restricties; als we in dat model het referentieniveau zouden wijzigen van de nominale variabelen (bvb. laatste i.p.v. eerste) moet de L-matrix opnieuw opgesteld worden! Bij het commando `lht` geven we ook mee dat we een F-toets willen uitvoeren om beide modellen te vergelijken.

```
> L1<-c(0,-1,0,1,0,0)
> L2<-c(0,-0.5,1,0,0,0)
> L<-rbind(L1,L2)
> L
   [,1] [,2] [,3] [,4] [,5] [,6]
L1    0 -1.0    0    1    0    0
L2    0 -0.5    1    0    0    0
> lht(fit1_beloning,L,test="F")
Linear hypothesis test

Hypothesis:
- beloning2  + taakinteresse3 = 0
- 0.5 beloning2  + taakinteresse2 = 0

Model 1: restricted model
Model 2: score ~ beloning + taakinteresse + beloning:taakinteresse

Res.Df  RSS  Df   Sum of Sq  F    Pr(>F)
1     20 36.727
2     18 28.000  2    8.7273 2.8052  0.087
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De output is heel gelijklopend aan de output die we verkrijgen bij modelvergelijkingen a.d.h.v. het commando `anova`.

Model 1 is het model met lineaire restricties en model 2 het model zonder lineaire restricties (volledige model).

Het aantal vrijheidsgraden dat overeenstemt met de kwadratensom voor het contrast is gelijk aan 2, dit is het aantal lineair onafhankelijke rijen in de L-matrix.

De toetsingsgrootheid volgt onder de hypothese dat $\mu_{21} = \mu_{13}$ én $\mu_{12} = \mu_{.1}$ een F-verdeling met 2 en 18 vrijheidsgraden. De geobserveerde toetsingsgrootheid is hier gelijk aan

$(8.7273/2)/(28/18) = 2.81$. Hiermee komt een p-waarde van 0.087 overeen. Dit betekent dat de hypothese niet verworpen kan worden op het 5% significantieniveau (p-waarde groter dan 5%). Indien we de nulhypothese wel zouden verwerpen, zouden we besluiten dat er evidentie is dat minstens één van de contrasten verschilt van 0.

8.3 Covariantie-analyse

Lineaire regressie met een combinatie van nominale predictoren en predictoren gemeten op minstens intervalniveau wordt ook covariantie-analyse genoemd.

Ook deze term is nauw verbonden met het experimenteel onderzoek. In de praktijk is het niet altijd mogelijk om de onderzoekseenheden volledig at random toe te wijzen aan de verschillende condities van een experiment. Dit betekent dat er mogelijks niet-bedoelde effecten m.b.t. de afhankelijke variabele een systematische invloed uitoefenen.

In een dergelijk geval kan men via de statistische weg corrigeren voor de niet-bedoelde effecten. De implementatie van het principe in de context van variantie-analyse geeft aanleiding tot wat men covariantie-analyse (**ancova**) noemt. De *covariaten* zijn predictoren van intervalniveau en de *factoren* zijn predictoren van nominaal niveau.

Opnieuw geldt hier dat data afkomstig van zowel experimentele als niet-experimentele designs geanalyseerd kunnen worden via covariantie-analyse (equivalent met lineaire regressie).

In sectie 4.4.3 beschouwden we het voorbeeld rond pijneducatie (sectie 1.2.2). Aangezien participanten in de studie niet at random toegewezen zijn aan de verschillende condities, bekeken we het effect van conditie, na statistische controle voor het effect van leeftijd en de graad van depressie. Dit is dus een voorbeeld van covariantie-analyse waarbij er 2 covariaten van intervalniveau zijn.

9 Referenties

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26*, 1295-1303.

Baron, R.M., & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1185.

Faraway, J.J. (2002) *Practical Regression and Anova using R.*

Kutner, M., Nachtscheim, C., Neter, J., Li, W. (2004). *Applied Linear Statistical Models, 5th ediдtion.* McGraw-Hill.

MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104.

MacKinnon, D.P., Lockwood, C.M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*, 99128.

Mosely, G.L., 2004, Evidence for a direct relationship between cognitive and psychological change during an education intervention in people with chronic low back pain. *European Journal of Pain, 8*, 39-45

Preacher, K.J., Hayes, & A.F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717-731.

Wong, P. P., Monette, G., & Weiner, N. I. (2001) Mathematical models of cognitive recovery. *Brain Injury, 15*, 519530.

Het multivariaat lineair model

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

1 Inleiding

Onderzoekers zijn vaak geïnteresseerd in het effect van onafhankelijke variabelen of predictoren op verschillende afhankelijke variabelen of uitkomsten.

Voorbeeld
Onderzoek naar het effect van alcohol op

- het aantal gemaakte fouten
- de snelheid waarmee een taak uitgevoerd wordt

Het meest efficiënte is uiteraard om de verschillende uitkomsten in eenzelfde onderzoek te beschouwen en dus worden deze uitkomsten bij dezelfde participanten gemeten.

- Het is mogelijk om aparte toetsen voor elke uitkomst uit te voeren maar dit betekent dat verschillende hypotheses simultaan getoetst worden.

  Een correctie voor het aantal toetsen is noodzakelijk om een inflatie van de kans op een type I fout (α) te vermijden.

  Stel dat er $q = 10$ uitkomsten zijn wat betekent dat er $q = 10$ univariate toetsen uitgevoerd worden, telkens op het 5% significantieniveau. In dat geval zal de kans op een Type I fout beduidend groter worden dan $\alpha = .05$. In het geval de nulhypothese waar is voor de $q = 10$ toetsen en de toetsen onafhankelijk zijn, krijgen we:

$$\begin{aligned} P(\text{tenminste 1 Type I fout}) &= 1 - P(\text{geen enkele Type I fout}) \\ &= 1 - (0.95)^{10} \\ &= 0.4012631 \end{aligned}$$

  In het geval de uitkomsten gecorreleerd zijn (dan zijn de toetsen gecorreleerd), zal deze kans ergens tussen 0.05 en 0.40 liggen.

  De Bonferroni correctie bestaat erin dat men bij het uitvoeren van q toetsen, iedere toets afzonderlijk op significantieniveau α/q uitvoert. Op die manier zal de kans op tenminste 1 Type I fout hoogstens gelijk zijn aan α. Wanneer q groot is, betekent dit dat de correctie kan leiding tot een substantieel verlies aan statistische power.

- Een ander nadeel van univariate toetsen is dat deze geen rekening houden met de samenhang tussen de uitkomsten.

Multivariate analyses bieden een aantal voordelen:

- De kans op een type I fout wordt niet opgeblazen.
- De correlaties tussen uitkomsten worden mee in rekening gebracht.
- De power van multivariate toetsen is in het algemeen hoger, vooral wanneer de verschillende uitkomsten sterk gecorreleerd zijn.

In het algemeen gaat men als volgt te werk bij een multivariate analyse:

- Eerst wordt een multivariate analyse uitgevoerd op het geheel van alle uitkomsten. Op die manier toetst men de nulhypothese dat de set van predictoren in het model geen invloed hebben op de afhankelijke variabelen.
- Als de nulhypothese in de vorige stap verworpen wordt, kan men via univariate analyses nagaan op welke afhankelijke variabelen de predictoren een invloed hebben.
- Voor iedere afhankelijke variabele die in de vorige stap geselecteerd werd, kan men gaan kijken wat de invloed van de verschillende predictoren zijn.

In dit hoofdstuk beschouwen we **multivariate lineaire regressie**, dit is de multivariate uitbreiding van univariate lineaire regressie. In het multivariate geval wordt het effect van predictoren op meerdere uitkomsten gemodelleerd. We veronderstellen hierbij dat deze uitkomsten van **minstens intervalniveau** zijn.

2 Voorbeeld

2.1 Data

We beschouwen data uit een onderzoek waarbij men geïnteresseerd is in de samenhang tussen academische vaardigheden en een aantal psychologische constructen.

De dataset bestaat uit observaties bij $n = 600$ studenten. Voor iedere student werden gegevens voor 6 variabelen verzameld. De uitkomsten zijn de volgende psychologische constructen:

- `locus of control`
- `self-concept`
- `motivation`

De uiteindelijke scores voor deze variabelen zijn z-scores en mogen als van intervalniveau beschouwd worden.

De predictoren zijn variabelen die academische vaardigheden (gestandaardiseerde testscores) weergeven:

- `reading`
- `writing`
- `science`

De scores op deze variabelen worden als van intervalniveau beschouwd.

Ook het opleidingsniveau van iedere student (1: laag, 2: gemiddeld of 3: hoog) wordt geregistreerd (`program`). Dit is een nominale variabele die bestaat uit 3 niveaus.

Zowel de code die hoort bij de analyses (`academisch.R`) als de data (`mvreg.csv`) zijn terug te vinden op Minerva.

Hieronder wordt een deel van de data getoond (eerste 6 rijen):

```
  locus.of.control self.concept  reading  writing  science   motivation program
1       -1.1439546    0.7226413 37.40555 39.03284 33.53282  0.368973076       2
2        0.5041339    0.1113640 52.76078 51.99504 65.22504  0.520318508       2
3        1.6285460    0.6299338 59.77192 54.65165 64.60450  0.436838150       2
4        0.3680964   -0.1385281 42.85432 41.12136 48.49381 -0.004323991       3
5       -0.2801896   -0.4522264 54.75628 49.94721 50.38166  1.256924033       2
6        1.0667162    0.6229352 56.59180 58.27413 61.51582  1.355576158       2
```

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 20 en 25.

2.2 Onderzoeksvragen

In dit hoofdstuk wensen we de volgende vier onderzoeksvragen te beantwoorden:

(1) Zijn er effecten van respectievelijk `reading`, `writing`, `science` en het opleidingsniveau op de 3 psychologische variabelen?

(2) Is er een verschil in gemiddelde uitkomst bij opleidingsniveau 2 (`program=2`) versus opleidingsniveau 1 (`program=1`) bij elk van de 3 psychologische variabelen?

(3) Is het effect van `writing` op `locus of control` gelijk aan het effect van `writing` op `self-concept`?

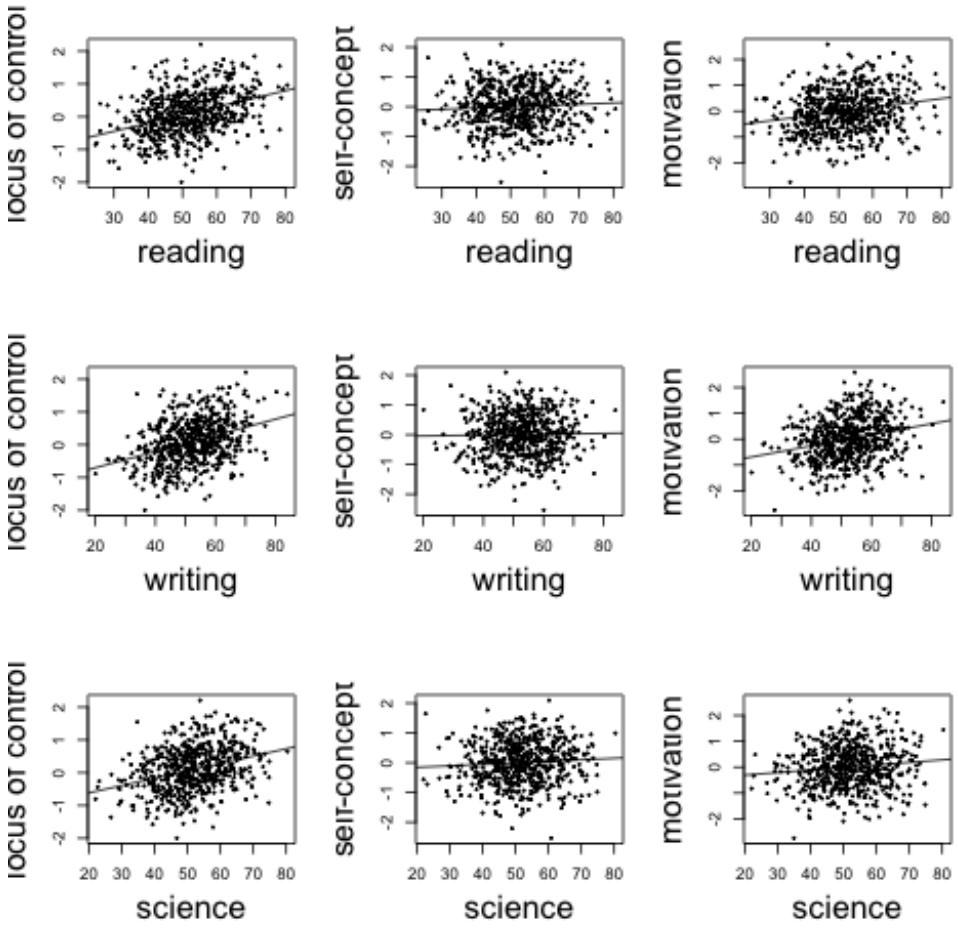
(4) Zijn de effecten van `reading`, `writing` en `science` op `locus of control` gelijk?

Bij deze onderzoeksvragen beschouwen we de academische variabelen en opleidingsniveau als onafhankelijke variabelen (predictoren) en de psychologische variabelen als de afhankelijke variabelen (uitkomst). Bij die definiëring worden geen expliciete veronderstellingen gemaakt m.b.t. eventuele causale relaties.

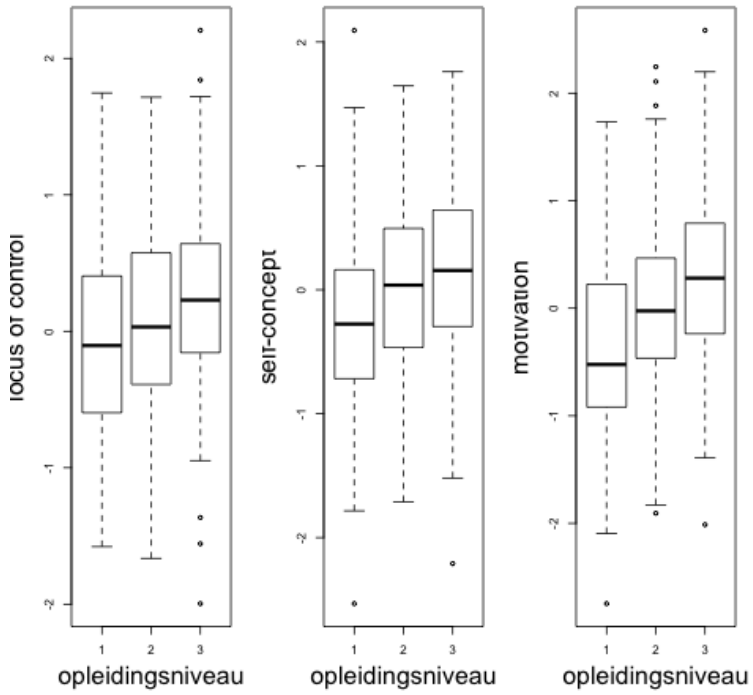
2.3 Data-exploratie

We starten eerst met een visuele exploratie van de data.

De assocaties tussen de verschillende continue variabelen kunnen we onderzoeken aan de hand van een scatterplot. Hieronder zijn telkens de best passende regressierechten toegevoegd. Dit biedt inzicht in de (lineaire) associatie tussen predictor en uitkomst.



Om de associatie tussen opleidingsniveau en de verschillende continue psychologische variabelen na te gaan, maken we gebruik van boxplots.



2.4 Notatie

We voeren de volgende notatie in:

- Y_{i1} score locus of control van deelnemer i ($i = 1, \ldots, n$ met $n = 600$)
- Y_{i2} score van self-concept van deelnemer i
- Y_{i3} score van motivation van deelnemer i
- x_{i1} score van reading van deelnemer i
- x_{i2} score van writing van deelnemer i
- x_{i3} score van science van deelnemer i

Aangezien opleidingsniveau nominaal is en uit drie niveaus bestaat, zullen we bij het opnemen van deze variabele als predictor gebruik maken twee hulpveranderlijken X_4 en X_5. We maken gebruik van dummy-codering met `program=3` als referentieniveau.

$x_{i4} = 1$ als deelnemer i in opleidingsniveau 2 zit (`program=2`), anders is $x_{i4} = 0$

$x_{i5} = 1$ als deelnemer i in opleidingsniveau 1 zit (`program=1`), anders is $x_{i5} = 0$

```
contrasts(academ$program)
  2 1
3 0 0
2 1 0
1 0 1
```

Beschouw de volgende lineaire regressiemodellen voor de drie uitkomsten:

$$
\begin{aligned}
Y_{i1} &= \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + \beta_{31}x_{i3} + \beta_{41}x_{i4} + \beta_{51}x_{i5} + \varepsilon_{i1} \\
Y_{i2} &= \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2} + \beta_{32}x_{i3} + \beta_{42}x_{i4} + \beta_{52}x_{i5} + \varepsilon_{i2} \\
Y_{i3} &= \beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2} + \beta_{33}x_{i3} + \beta_{43}x_{i4} + \beta_{53}x_{i5} + \varepsilon_{i3}
\end{aligned}
$$

met $E(\varepsilon_{i1}) = E(\varepsilon_{i2}) = E(\varepsilon_{i3}) = 0$. Hierbij stelt $\beta_{k\ell}$ het effect van onafhankelijke variabele k ($k = 1, \ldots, 5$) op uitkomst ℓ ($\ell = 1, 2, 3$) voor.

- De nulhypotheses (voor elke predictor) die horen bij onderzoeksvraag 1 kunnen in functie van de modelparameters als volgt geschreven worden:

  H_0: $\beta_{11} = \beta_{12} = \beta_{13} = 0$
  H_0: $\beta_{21} = \beta_{22} = \beta_{23} = 0$
  H_0: $\beta_{31} = \beta_{32} = \beta_{33} = 0$
  H_0: $\beta_{41} = \beta_{42} = \beta_{43} = \beta_{51} = \beta_{52} = \beta_{53} = 0$

- De nulhypothese die hoort bij onderzoeksvraag 2 kan in functie van de modelparameters als volgt geschreven worden:

  H_0: $\beta_{41} - \beta_{51} = \beta_{42} - \beta_{52} = \beta_{43} - \beta_{53} = 0$

- De nulhypothese die hoort bij onderzoeksvraag 3 kan in functie van de modelparameters als volgt geschreven worden:

  H_0: $\beta_{21} = \beta_{22}$

- De nulhypothese die hoort bij onderzoeksvraag 4 kan in functie van de modelparameters als volgt geschreven worden:

  H_0: $\beta_{11} = \beta_{21} = \beta_{31}$

Merk hierbij op dat het onderling vergelijken van effecten van verschillende predictoren op eenzelfde uitkomst slechts in beperkte contexten mogelijk is. In deze studie wordt gewerkt met gestandardiseerde testscores, wat vergelijking wel zinvol maakt. Concreet betekent dit dat een toename van een eenheid bij de drie academische vaardigheden vergelijkbaar is en dat de effecten van deze variabelen bijgevolg met elkaar vergeleken kunnen worden.

2.5 Univariate aanpak

Bij een univariate aanpak voor de analyse kunnen we naar de effecten van iedere predictor voor iedere uitkomst afzonderlijk kijken.

Via lineaire regressie kunnen we bvb. voor `locus of control` de volgende nulhypotheses toetsen:

H_0: $\beta_{11} = 0$

H_0: $\beta_{21} = 0$

H_0: $\beta_{31} = 0$

H_0: $\beta_{41} = \beta_{51} = 0$

De resultaten zijn als volgt:

```
> modeluni<-lm(locus.of.control~reading+writing+science+program,data=academ)
> modeluni_effect<-lm(locus.of.control~reading+writing+science+program,
                                  contrasts=c(program=contr.sum),data=academ)
>
> summary(modeluni)

Call:
lm(formula = locus.of.control ~ reading + writing + science +
program, data = academ)

Residuals:
Min      1Q  Median      3Q     Max
-1.9560 -0.3889 -0.0219  0.3725  1.9039
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.373094   0.162593  -8.445 2.34e-16 ***
reading      0.012505   0.003718   3.363 0.000819 ***
writing      0.012145   0.003391   3.581 0.000370 ***
science      0.005761   0.003641   1.582 0.114109
program2    -0.123875   0.057607  -2.150 0.031931 *
program1    -0.251671   0.068470  -3.676 0.000259 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.607 on 594 degrees of freedom
Multiple R-squared:  0.1868,Adjusted R-squared:  0.1799
F-statistic: 27.28 on 5 and 594 DF,  p-value: < 2.2e-16

> Anova(modeluni_effect,type=3)
Anova Table (Type III tests)

Response: locus.of.control
            Sum Sq  Df F value    Pr(>F)
(Intercept)  34.529   1 93.7161 < 2.2e-16 ***
reading       4.168   1 11.3128 0.0008193 ***
writing       4.725   1 12.8248 0.0003700 ***
science       0.922   1  2.5037 0.1141093
program       5.030   2  6.8255 0.0011730 **
Residuals   218.856 594
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We vinden de volgende resultaten voor de verschillende predictoren

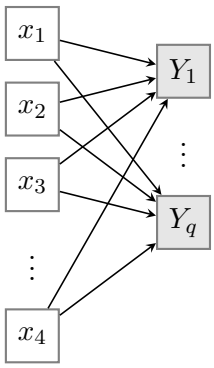
- `reading`: $\hat{\beta}_{11} = 0.013, F(1, 594) = 11.312, p = .001$
- `writing`: $\hat{\beta}_{21} = 0.012, F(1, 594) = 12.825, p < .001$
- `science`: $\hat{\beta}_{31} = 0.006, F(1, 594) = 2.504, p = 0.114$
- `program`: $\hat{\beta}_{41} = -0.252, \hat{\beta}_{51} = -0.124, F(2, 594) = 6.826, p = .001$

Enkel het effect van `science` op `locus of control` is niet statistisch significant op het 5% significantieniveau.

Analoog zouden we deze effecten op de andere uitkomsten kunnen bekijken. Zoals aangehaald leidt het toetsen van het effect van een predictor op elke uitkomst afzonderlijk tot een verhoogde kans op een Type I fout. Bovendien houden we op die manier ook geen rekening met de correlaties tussen de drie uitkomsten.

Via een multivariate lineaire regressie kunnen we simultaan het effect van een predictor op alle uitkomsten toetsen.

3 Multivariate lineaire regressie



- $\mathbf{Y} = [Y_1, \ldots, Y_q]$: q afhankelijke variabelen die conceptueel samenhoren
- $\mathbf{X} = [x_1, \ldots, x_p]$: p predictoren
- multivariate toetsen: houden rekening met de correlatiestructuur tussen de afhankelijke variabelen

3.1 De structuur van het model

Veronderstel dat we het effect van p predictoren op q uitkomsten wensen te modelleren aan de hand van multivariate lineaire regressie. We beschikken hiervoor over een steekproef die bestaat uit n observaties.

Algemeen noteren we het multivariaat lineair model als volgt: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Dit is de beknopte matrixnotatie voor:

$$\begin{bmatrix} y_{11} & y_{12} & \dots & y_{1q} \\ y_{21} & y_{22} & \dots & y_{2q} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ y_{n1} & y_{n2} & \dots & y_{nq} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0q} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1q} \\ \vdots & \vdots & & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pq} \end{bmatrix}$$
$$+ \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1q} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2q} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{nq} \end{bmatrix}$$

- $\mathbf{Y}$ is een respons matrix van orde $n \times q$. Elke rij van $\mathbf{Y}$ correspondeert met de q scores van 1 individu.
$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1q} \\ y_{21} & y_{22} & \dots & y_{2q} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ y_{n1} & y_{n2} & \dots & y_{nq} \end{bmatrix}$$
- $\mathbf{X}$ is de *model matrix* van orde $n \times (p+1)$ (inclusief het intercept). Deze matrix is van dezelfde orde als in het univariaat lineair model (1 uitkomst).
- $\boldsymbol{\beta}$ is een $(p+1) \times q$ matrix van regressiecoëfficiënten. We hebben een aparte kolom van β's voor elke kolom van $\mathbf{Y}$.
- $\boldsymbol{\varepsilon}$ is een matrix van dezelfde orde als $\mathbf{Y}$ en bevat de random fouttermen voor elk individu, voor elke afhankelijke variabele.

Ieder individu i $(i = 1, ..., n)$ heeft dus een vector van scores:

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iq} \end{bmatrix}$$

en een vector van fouttermen:

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iq} \end{bmatrix}.$$

3.2 Stochastische assumpties

We leunen op de volgende assumpties voor het multivariaat lineair model:

1. de basisassumptie dat het model volledig en lineair is:

   $$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

   Hierbij zijn zowel ϵ als $\mathbf{0}$ $n \times q$ matrices. Deze assumptie impliceert ook:

   $$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\beta.$$

2. de homogeniteitsassumptie:

   $\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma} \qquad i = 1, \ldots, n.$

   Hierbij is $\boldsymbol{\Sigma}$ een $q \times q$ variantie-covariantiematrix. Er worden geen restricties opgelegd aan de structuur van de covariantiematrix, men zegt dat de covariantiematrix *unstructured* is.

3. de assumptie van onafhankelijke individuen:

   $\mathrm{Cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) = \mathbf{0}$ voor alle $i \neq j$.

3.3 Parameterschattingen: kleinste kwadratenschatters

Het gefit multivariaat lineair model is $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{e}$ met $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{XB}$ de *matrix* met fouttermen. De kleinste kwadratenmethode "minimaliseert":

$$\boldsymbol{E} = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{XB})'(\boldsymbol{Y} - \boldsymbol{XB}).$$

De oplossing hiervoor is analoog aan het univariate geval. $\mathbf{B}$ wordt als volgt geschat:

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

De oplossing $\boldsymbol{B}$ is een matrix van orde $(p+1) \times q$ (er zijn p predictoren en q uitkomsten). Elke kolom bevat de schatters voor de regressieparameters die horen bij een welbepaald uitkomst.

Merk hierbij op dat de waarden in de kolommen van $\boldsymbol{B}$ dezelfde zijn als mochten we ze apart hebben geschat voor elke kolom van $\boldsymbol{Y}$ afzonderlijk (in een univariaat model).

Een onvertekende schatter voor $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-p-1} = \frac{\boldsymbol{E}}{n-p-1}.$$

4 Toetsing

4.1 Multivariate toetsen

Bij het toetsen van de effecten van predictoren in een multivariaat model veronderstellen we dat de q uitkomsten een multivariate normale verdeling volgen:

$$\boldsymbol{Y} \sim N_q(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

4.1.1 Multivariate normale verdeling

De multivariate normale verdeling is een uitbreiding van de univariate normale verdeling naar meerdere dimensies.

Veronderstel dat de variabelen Y_1 en Y_2 bivariaat normaal verdeeld zijn (i.e. multivariaat normaal met $q = 2$):

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right).$$

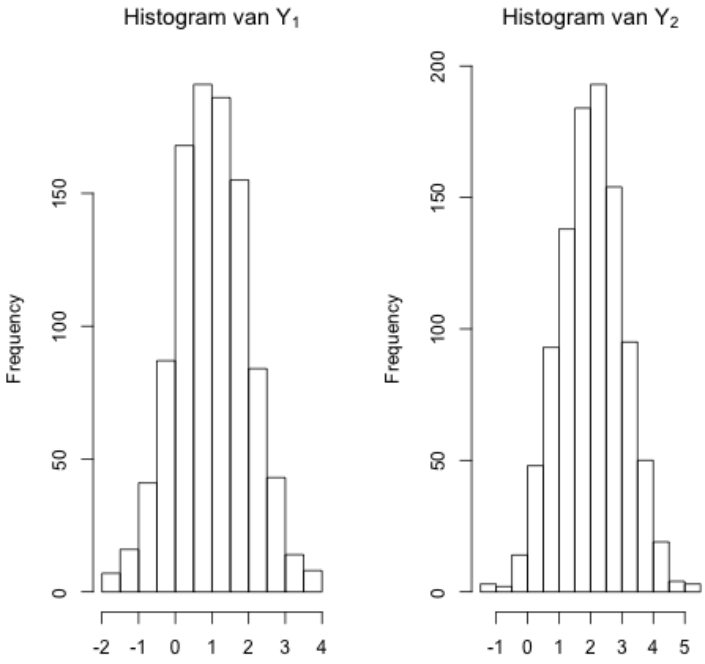
Hieruit volgt dat $E(Y_1) = \mu_1$, $\text{Var}(Y_1) = \sigma_1^2$, $E(Y_2) = \mu_2$ en $\text{Var}(Y_2) = \sigma_2^2$. Verder geldt dat $\text{Cov}(Y_1, Y_2) = \sigma_{12}$. Een multivariate verdeling legt niet enkel de verdeling van de individuele variabelen vast maar ook de samenhang tussen de variabelen.

Een set van variabelen is multivariaat normaal verdeeld als elke lineaire combinatie van deze variabelen een univariate normale verdeling volgt.

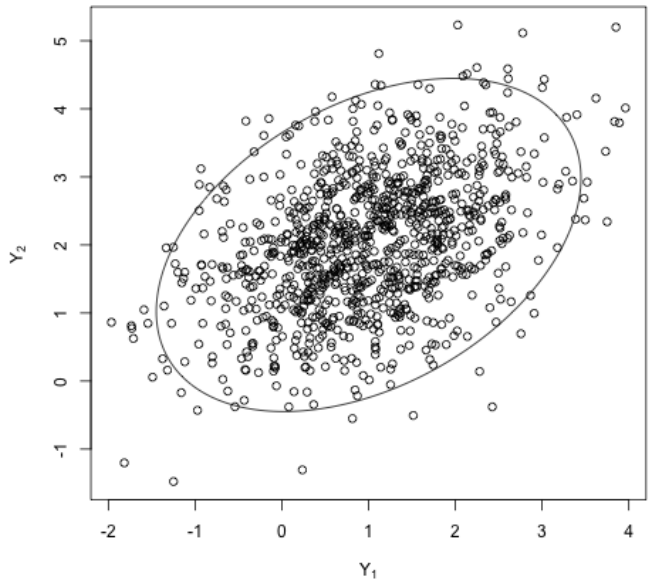
Hieruit volgt dat wanneer een set van variabelen een multivariate normale verdeling volgt dat de individuele variabelen univariaat normaal verdeeld zijn maar het omgekeerde is niet noodzakelijk waar.

Veronderstel dat we 1000 observaties genereren van twee variabelen Y_1 en Y_2 die bivariaat normaal verdeeld zijn met $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 1$ en $\sigma_{12} = 0.4$.

De histogrammen op basis van de 1000 observaties voor beide variabelen zien er als volgt uit:



Het geobserveerde spreidingsdiagram van Y_2 t.ov. Y_1 is als volgt:



De contouren van de ellips zijn zo bepaald dat de ellips 95% van de observaties bevat voor een bivariate normale verdeling met bovenstaande parameters.

De code voor het genereren van de data en de figuren staat ter beschikking op Minerva (`MultivariaatNormaal.R`, ter informatie).

4.1.2 Modelvergelijkingen

Analoog aan toetsen bij univariate lineaire regressie kunnen modelvergelijkingen gebruikt worden om geneste modellen met elkaar te vergelijken.

Veronderstel bvb. dat we een multivariaat model met 10 predictoren (model A) wensen te vergelijken met een model met 5 predictoren (model B, bevat een subset van de 10 predictoren). In dat geval toetsen we de nulhypothese die stelt dat geen enkele predictor uit model A die we weglaten bij model B een invloed heeft op minstens één van de uitkomsten.

Laat M_F het volledig model voorstellen (model A) en M_R het gereduceerd model (model B). Onder de nulhypothese stellen we dat het gereduceerd model evenwaardig is met het volledig model (kortom, de extra predictoren zijn overbodig):

$$H_0 : M_F = M_R \quad \text{met} \quad R \subset F.$$

De geschatte uitkomsten op basis van beide modellen zijn:

$$\begin{aligned}\hat{\mathbf{Y}}_F &= \mathbf{X}_F\mathbf{B}_F \\ \hat{\mathbf{Y}}_R &= \mathbf{X}_R\mathbf{B}_R.\end{aligned}$$

De E en H matrices

In het univariaat geval is SSE de kwadratensom van de residuen. Laat SSE_F de foutkwadratensom voorstellen voor het volledig model (model A) en SSE_R de foutkwadratensom voor het gereduceerd model (model B). In een modelvergelijkingstoets worden beide modellen met elkaar vergeleken op basis van $\text{SSH} = \text{SSE}_R - \text{SSE}_F$.

De multivariate tegenhanger voor de SSE_F term is de *Error Sum of Squares and Cross Products* (E-SSCP) matrix $\mathbf{E}$:

$$\mathbf{E} = \mathbf{E_F} = (\mathbf{Y} - \hat{\mathbf{Y}}_\mathbf{F})'(\mathbf{Y} - \hat{\mathbf{Y}}_\mathbf{F}).$$

De multivariate tegenhanger voor de SSE_R term is

$$\mathbf{E_R} = (\mathbf{Y} - \hat{\mathbf{Y}}_\mathbf{R})'(\mathbf{Y} - \hat{\mathbf{Y}}_\mathbf{R}).$$

Een multivariate tegenhanger voor de $\text{SSH} = (\text{SSE}_R - \text{SSE}_F)$ term is de *Hypothesis Sum of Squares and Cross Products* (H-SSCP) matrix $\mathbf{H}$:

$$\mathbf{H} = \mathbf{E_R} - \mathbf{E_F}.$$

Multivariate toetsingsgrootheden

Er bestaan verschillende multivariate toetsingsgrootheden die gebaseerd zijn op $\mathbf{E}$ en $\mathbf{H}$:

- Wilks' Lambda
- Pillai's trace
- Hotelling-Lawley trace
- Roy's largest root

In de sociale wetenschappen rapporteert men zo goed als altijd enkel Wilks' Lambda:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} \qquad (0 < \Lambda < 1).$$

Hierbij stelt $|\mathbf{D}|$ de determinant van een matrix $\mathbf{D}$ voor.

Deze maat Λ kan ruwweg geïnterpreteerd worden als de hoeveelheid variatie in de uitkomsten die niet verklaard wordt door de predictoren die getoetst worden. Λ varieert tussen 0 (alle verklaarde variatie in de uitkomsten is te wijten aan de predictoren die getoetst worden) en 1 (de predictoren verklaren niets van de variatie in de uitkomsten).

In de meeste gevallen is de exacte verdeling van de multivariate statistieken onder de nulhypothese niet gekend, maar benaderen ze na transformatie een F-verdeling.

De berekening van het aantal vrijheidsgraden is niet triviaal. We geven de berekingen hier ter informatie; je hoeft dit niet zelf te kunnen narekenen voor de voorbeelden die aan bod komen. Laat s het aantal vrijheidsgraden voorstellen die overeenkomt met de te toetsen effecten (i.e. het verschil in aantal geschatte regressieparameters tussen het volledig en het gereduceerd model). Definieer verder:

$$\begin{aligned} r &= n - p - 1 - \frac{q - s + 1}{2} \\ u &= \frac{qs - 2}{4} \\ t &= \begin{cases} \sqrt{\frac{q^2 s^2 - 4}{q^2 + s^2 - 5}} & \text{als } q^2 + s^2 - 5 > 0 \\ 0 & \text{anders} \end{cases} \end{aligned}$$

De toetsingrootheid op basis van Wilks' Lambda wordt dan berekend als $F = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \times \frac{rt-2u}{qs}$.

Onder H_0 geldt dat deze toetsingsgrootheid benaderend een F-verdeling volgt met qs en $rt - 2u$ vrijheidsgraden.

In de sociale wetenschappen wordt vaak onderstaande strategie gehanteerd voor het toetsen van het effect van predictoren in het multivariate geval:

- Rapporteren welke multivariate test gehanteerd werd (bvb. *All multivariate tests reported in this paper are based on Wilks' Lambda*).
- Enkel bij een statistisch significante multivariate toets wordt er naar de univariate toetsen gekeken. In dat geval is er immers evidentie voor het feit dat minstens één predictor een effect heeft op minstens één uitkomst.

  Het is dus niet toegelaten om statistisch significante univariate toetsen te vermelden indien de corresponderende multivariate toets niet significant is.
- Als de multivariate toets statistisch significante resultaten oplevert, worden deze verder geïnterpreteerd aan de hand van de univariate regressiecoëfficiënten.

Merk op dat deze strategie niet vrij van kritiek is. Deze strategie waarbij eerst een multivariate toets uitgevoerd wordt, laat toe om de kans op een type I fout te controleren en rekening te houden met de correlatie tussen de uitkomsten. Vaak is men echter niet geïnteresseerd in een omnibus toets (waarbij hier effecten voor alle uitkomsten simultaan bekeken worden) maar in specifieke contrasten (zie sectie 4.2).

4.1.3 Voorbeeld

We hernemen het voorbeeld over het onderzoek bij studenten naar de samenhang tussen academische vaardigheden en een aantal psychologische constructen (zie sectie 2).

In R kunnen we voor multivariate lineaire regressie opnieuw gebruik maken van het commando `lm` om het model te definiëren en van het commando `anova` om modelvergelijkingstoetsen uit te voeren. Hierbij dient gespecifieerd te worden dat de toets gebaseerd moet zijn op Wilks' Lambda (`test="Wilks"`).

We moeten eerst een matrix maken waarbij elke kolom een uitkomst voorstelt. Deze matrix wordt dan als afhankelijke variabele in het lineair regressiemodel gebruikt. Op die manier wordt een multivariaat lineair regressiemodel gefit.

```
> Y <- as.matrix(academ[,c("locus.of.control","self.concept","motivation")])
> modelmulti <- lm(Y~reading+writing+science+program,data=academ)
```

Wanneer we het effect van `science` op de drie uitkomsten simultaan willen toetsen, kunnen we gebruik maken van `anova` om een modelvergelijkingstoets uit te voeren. De nulhypothese H_0 stelt dat $\beta_{31} = \beta_{32} = \beta_{33} = 0$.

```
> modelmulti_S <- lm(Y~reading+writing+program,data=academ)
> anova(modelmulti_S,modelmulti,test="Wilks")
Analysis of Variance Table

Model 1: Y ~ reading + writing + program
Model 2: Y ~ reading + writing + science + program
   Res.Df Df Gen.var. Wilks approx F num Df den Df  Pr(>F)
1    595     0.45377
2    594 -1  0.45201 0.98341   3.3299      3    592 0.01931 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We lezen af dat Wilks' $\Lambda = 0.983$ (indien gewenst kunnen de andere toetsingsgrootheden ook op een analoge manier opgevraagd worden). De resultaten voor de bijhorende toets zijn $F(3, 592) = 3.330, p = .019$. De vrijheidsgraden hiervoor worden berekend zoals aangegeven op pagina 19 (dit hoef je niet zelf na te kunnen rekenen). We kunnen H_0 verwerpen op het 5% significantieniveau en besluiten dat `science` op minstens 1 van de uitkomsten een effect heeft (minstens 1 van de β's is verschillend van 0).

We kunnen dit in meer detail bekijken via de univariate analyses. Deze krijgen we voor de drie uitkomsten via `summary` van het gefitte multivariate model.

```
> summary(modelmulti)
Response locus.of.control :

Call:
lm(formula = locus.of.control ~ reading + writing + science +
program, data = academ)

Residuals:
   Min      1Q  Median      3Q     Max
-1.9560 -0.3889 -0.0219  0.3725  1.9039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.373094   0.162593  -8.445 2.34e-16 ***
reading      0.012505   0.003718   3.363 0.000819 ***
writing      0.012145   0.003391   3.581 0.000370 ***
science      0.005761   0.003641   1.582 0.114109
program2    -0.123875   0.057607  -2.150 0.031931 *
program1    -0.251671   0.068470  -3.676 0.000259 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.607 on 594 degrees of freedom
Multiple R-squared:  0.1868,Adjusted R-squared:  0.1799
F-statistic: 27.28 on 5 and 594 DF,  p-value: < 2.2e-16


Response self.concept :

Call:
```

```
lm(formula = self.concept ~ reading + writing + science + program,
data = academ)

Residuals:
   Min       1Q   Median       3Q      Max
-2.38183 -0.46594  0.00604  0.46063  2.28836

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.051018   0.184577   0.276   0.7823
reading      0.001308   0.004220   0.310   0.7568
writing     -0.004293   0.003850  -1.115   0.2652
science      0.005306   0.004133   1.284   0.1998
program2    -0.146876   0.065396  -2.246   0.0251 *
program1    -0.423359   0.077728  -5.447 7.52e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6891 on 594 degrees of freedom
Multiple R-squared:  0.05404,Adjusted R-squared:  0.04607
F-statistic: 6.786 on 5 and 594 DF,  p-value: 3.629e-06


Response motivation :

Call:
lm(formula = motivation ~ reading + writing + science + program,
data = academ)

Residuals:
    Min       1Q   Median       3Q      Max
-2.31821 -0.50736 -0.03076  0.51596  2.33499

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.691146   0.203952  -3.389 0.000749 ***
reading      0.009674   0.004664   2.074 0.038481 *
writing      0.017535   0.004254   4.122 4.29e-05 ***
science     -0.009001   0.004567  -1.971 0.049209 *
```

```
program2    -0.259367   0.072261  -3.589 0.000359 ***
program1    -0.619696   0.085887  -7.215 1.65e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7614 on 594 degrees of freedom
Multiple R-squared:   0.15,Adjusted R-squared:  0.1428
F-statistic: 20.96 on 5 and 594 DF,  p-value: < 2.2e-16
```

We stellen vast dat er enkel bij de uitkomst `motivation` een statistisch significant effect is van `science` ($p = 0.049$). Merk op dat de p-waarde hier een stuk groter is dan de p-waarde van de multivariate toets, onthoud dat de multivariate toets vaak meer power heeft dan de univariate toetsen.

We kunnen op een analoge manier de effecten van de overige predictoren toetsen. Om de effecten van alle afzonderlijke predictoren te toetsen, kunnen we ook gebruik van van het commando `Anova` (R-package `car`). Merk net zoals bij univariate lineaire regressie op dat we gebruik maken van Type III toetsen en dat we hiervoor eerst overgaan op effect-codering voor de nominale predictoren (in dit geval is dit `program`).

```
> modelmulti_effect <- lm(Y~reading+writing+science+program,
+                         contrasts=c(program=contr.sum),data=academ)
> Anova(modelmulti_effect,type=3,test="Wilks")

Type III MANOVA Tests: Wilks test statistic
             Df test stat approx F num Df den Df    Pr(>F)
(Intercept)  1   0.84297   36.759      3    592 < 2.2e-16 ***
reading      1   0.97643    4.764      3    592  0.002727 **
writing      1   0.94739   10.957      3    592 5.186e-07 ***
science      1   0.98341    3.330      3    592  0.019305 *
program      2   0.89144   11.671      6   1184 9.806e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In de eerste kolom `Df` lezen we af hoeveel variabelen coderen voor de predictor, bij de continue predictoren is dit telkens 1 en bij `program` is dit 2 aangezien er gebruik gemaakt wordt van 2 hulpveranderlijken.

We lezen af dat de multivariate toets voor elke predictor statistisch significant is op het 5% significantieniveau.

Voor **onderzoeksvraag 1** kunnen we dus voor elke predictor de nulhypothese verwerpen. Voor de schattingen en interpretatie van de parameters kunnen we kijken naar de resultaten van de univariate lineaire regressies, alsook voor het toetsen van de effecten (aangezien de multivariate toets telkens statistisch significant is).

4.2 Algemeen lineaire hypotheses

Onderzoeksvraag 1 in het voorbeeld over het onderzoek bij studenten kon beantwoord worden aan de hand van modelvergelijkingstoetsen. Voor onderzoeksvragen 2, 3 en 4 gaan we een stap verder en moeten we specifieke contrasten van regressieparameters toetsen. Dit kan aan de hand van algemeen lineaire hypotheses.

4.2.1 Contrasten

We werken nog steeds met het volgende multivariaat lineair model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In het univariate geval ($q = 1$), hebben we in sectie 8.2.3 van het hoofdstuk rond lineaire regressie gezien hoe toetsen voor specifieke contrasten van effecten van predictoren uitgevoerd kunnen worden aan de hand van het toetsen van een algemene lineaire hypothese (ALH):

$$\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}.$$

Op die manier kunnen lineaire restricties opgelegd worden aan de parameters van het model en kan het model met restricties vergeleken worden met het model zonder restricties. Door gebruik te maken van de ALH kunnen we specifieke onderzoeksvragen met betrekking tot onze β-parameters beantwoorden. In het multivariate geval ($q > 1$) kunnen we echter ook combinaties van uitkomsten beschouwen bij het toetsen en bvb. de effecten van een predictor vergelijken over uitkomsten. Een algemene lineaire hypothese in het multivariate geval wordt bijgevolg:

$$\boldsymbol{L}\boldsymbol{\beta}\boldsymbol{M} = \boldsymbol{K}.$$

> $\boldsymbol{L}$: L-matrix (cfr. univariate lineaire regressie), bevat contrasten voor regressiecoëfficiënten.
>
> $\boldsymbol{M}$: transformatiematrix (M-matrix), geeft combinatie van uitkomsten weer.

Vaak is $\boldsymbol{M}$ eenheidsmatrix (dit wil zeggen dat alle uitkomsten afzonderlijk beschouwd worden, de toetsen houden wel rekening met de correlaties tussen uitkomsten) en $\boldsymbol{K}$ een nulmatrix.

Net zoals bij univariate lineaire regressie kunnen dergelijke hypotheses in R getoetst worden aan de hand van `lht` (package `car`). We moeten hierbij het model zonder restricties, de L-matrix, de M-matrix en K meegeven.

De uiteindelijke toets is opnieuw een multivariate toets en we maken gebruik van Wilks' Lambda. Bij een statistisch significante multivariate toets kan men kijken naar de resultaten van de univariate toetsen.

4.2.2 Voorbeeld

We hernemen het voorbeeld over het onderzoek bij studenten naar de samenhang tussen academische vaardigheden en een aantal psychologische constructen (zie sectie 2).

Voor dit specifieke voorbeeld is $n = 600$, $q = 3$ en $p = 5$ (er zijn 4 predictoren maar voor opleidingsniveau worden 2 hulpveranderlijken gebruikt).

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ . & . & . \\ . & . & . \\ . & . & . \\ y_{600,1} & y_{600,2} & y_{600,3} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{15} \\ 1 & x_{21} & \dots & x_{25} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & x_{600,1} & \dots & x_{600,5} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \\ \beta_{41} & \beta_{42} & \beta_{43} \\ \beta_{51} & \beta_{52} & \beta_{53} \end{bmatrix}$$

$$+ \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ . & . & . \\ . & . & . \\ . & . & . \\ \varepsilon_{600,1} & \varepsilon_{600,2} & \varepsilon_{600,3} \end{bmatrix}$$

Onderzoeksvraag 2

Is er een verschil in gemiddelde uitkomst bij opleidingsniveau 2 (`program=2`) versus opleidingsniveau 1 (`program=1`) bij elk van de 3 psychologische variabelen?

$\Rightarrow H_0$: $\beta_{41} - \beta_{51} = \beta_{42} - \beta_{52} = \beta_{43} - \beta_{53} = 0$

Deze onderzoeksvraag gaat over een contrast tussen 2 regressieparameters, voor elke uitkomst afzonderlijk.

- Stel $\boldsymbol{L} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$

$$\boldsymbol{L\beta} = \begin{pmatrix} \beta_{41} - \beta_{51} & \beta_{42} - \beta_{52} & \beta_{43} - \beta_{53} \end{pmatrix}$$

- Stel $\boldsymbol{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ en $\boldsymbol{K} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$

$$\begin{aligned} \boldsymbol{L\beta M} &= \boldsymbol{K} \\ \begin{pmatrix} \beta_{41} - \beta_{51} & \beta_{42} - \beta_{52} & \beta_{43} - \beta_{53} \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

```
> ### L-matrix
> L <- c(0, 0, 0, 0, 1, -1)
>
> ### M-matrix
> M <- matrix(
+   c(1, 0, 0,
+     0, 1, 0,
+     0, 0, 1),
+   nrow = 3,
+   byrow = TRUE)
>
> ### K
> K <- c(0, 0, 0)
>
> ### We slaan ALH in object op om meer gedetailleerde informatie krijgen
> ### via print()
> ### verbose = TRUE geeft ook meer uitgebreide informatie over het geschatte contrast
>
> onderzoeksvraag2 <- lht(modelmulti,
+                         hypothesis.matrix = L,
+                         P = M,
+                         rhs = K,
+                         verbose = TRUE)

Hypothesis matrix:
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    0    1   -1
```

```
Right-hand-side matrix:
[1] 0 0 0

Estimated linear function (hypothesis.matrix %*% coef - rhs):
[1] 0.1277951 0.2764834 0.3603294

> print(onderzoeksvraag2)

Response transformation matrix:
                  [,1] [,2] [,3]
locus.of.control    1    0    0
self.concept        0    1    0
motivation          0    0    1

Sum of squares and products for the hypothesis:
         [,1]     [,2]      [,3]
[1,] 1.471132 3.182780  4.147985
[2,] 3.182780 6.885913  8.974126
[3,] 4.147985 8.974126 11.695607

Sum of squares and products for error:
        [,1]      [,2]      [,3]
[1,] 218.85624  34.14870  35.93761
[2,]  34.14870 282.04029  77.83401
[3,]  35.93761  77.83401 344.36143

Multivariate Tests:
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1 0.0467631 9.680617      3    592 3.0287e-06 ***
Wilks             1 0.9532369 9.680617      3    592 3.0287e-06 ***
Hotelling-Lawley  1 0.0490572 9.680617      3    592 3.0287e-06 ***
Roy               1 0.0490572 9.680617      3    592 3.0287e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Wilks' $\Lambda = 0.953, F(3, 592) = 9.681, p < .001$. We verwerpen bijgevolg H_0 op het 5% significantieniveau.

Het aantal vrijheidsgraden in de kolom `Df` is gelijk aan het aantal rijen van de L-matrix.

We kunnen nu naar deze contrasten gaan kijken binnen de univariate modellen.

```
> model1<-lm(locus.of.control~reading+writing+science+program,data=academ)
> model2<-lm(self.concept~reading+writing+science+program,data=academ)
> model3<-lm(motivation~reading+writing+science+program,data=academ)
>
> onderzoeksvraag2_1<-lht(model1,L,verbose=TRUE)

Hypothesis matrix:
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    0    1   -1

Right-hand-side vector:
[1] 0

Estimated linear function (hypothesis.matrix %*% coef - rhs)
[1] 0.1277951


Estimated variance of linear function
[1] 0.004090243

> print(onderzoeksvraag2_1)
Linear hypothesis test

Hypothesis:
program2 - program1 = 0

Model 1: restricted model
Model 2: locus.of.control ~ reading + writing + science + program

    Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    595 220.33
2    594 218.86  1    1.4711 3.9928 0.04615 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> onderzoeksvraag2_2<-lht(model2,L,verbose=TRUE)
```

```
Hypothesis matrix:
    [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    0    1   -1

Right-hand-side vector:
[1] 0

Estimated linear function (hypothesis.matrix %*% coef - rhs)
[1] 0.2764834


Estimated variance of linear function
[1] 0.005271101

> print(onderzoeksvraag2_2)
Linear hypothesis test

Hypothesis:
program2 - program1 = 0

Model 1: restricted model
Model 2: self.concept ~ reading + writing + science + program

    Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    595 288.93
2    594 282.04  1    6.8859 14.502 0.0001545 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> onderzoeksvraag2_3<-lht(model3,L,verbose=TRUE)

Hypothesis matrix:
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    0    1   -1

Right-hand-side vector:
[1] 0

Estimated linear function (hypothesis.matrix %*% coef - rhs)
```

```
[1] 0.3603294


Estimated variance of linear function
[1] 0.006435831

> print(onderzoeksvraag2_3)
Linear hypothesis test

Hypothesis:
program2 - program1 = 0

Model 1: restricted model
Model 2: motivation ~ reading + writing + science + program

    Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    595 356.06
2    594 344.36  1    11.696 20.174 8.498e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De univariate testen tonen een statistisch significant verschil tussen opleidingsniveau 2 en 1 voor alle uitkomsten: `locus of control` ($\hat{\beta}_{41} - \hat{\beta}_{51} = 0.128, p = .046$), `self-concept` ($\hat{\beta}_{42} - \hat{\beta}_{52} = 0.276, p < .001$) en `motivation` ($\hat{\beta}_{43} - \hat{\beta}_{53} = 0.360, p < .001$).

Onderzoeksvraag 3

Is het effect van `writing` op `locus of control` gelijk aan het effect van `writing` op `self-concept`?

$\Rightarrow$ H_0: $\beta_{21} - \beta_{22} = 0$

- Stel $\boldsymbol{L} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$

$$\boldsymbol{L\beta} = \begin{pmatrix} \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}$$

- Stel $\boldsymbol{M} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ en $\boldsymbol{K} = \begin{pmatrix} 0 \end{pmatrix}$

$$\boldsymbol{L\beta M} = \boldsymbol{K}$$

$$\begin{pmatrix} \beta_{21} - \beta_{22} \end{pmatrix} = \begin{pmatrix} 0 \end{pmatrix}$$

```
> L <- c(0, 0, 1, 0, 0, 0)
> M <- matrix(
+   c(1, -1, 0),
+   nrow = 3)
> K <- matrix(0)
> onderzoeksvraag3 <- lht(modelmulti,
+                         hypothesis.matrix = L,
+                         P = M,
+                         rhs = K,
+                         verbose = TRUE)

Hypothesis matrix:
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    1    0    0    0

Right-hand-side matrix:
     [,1]
[1,]    0

Estimated linear function (hypothesis.matrix %*% coef - rhs):
[1] 0.01643848

> print(onderzoeksvraag3)

Response transformation matrix:
                  [,1]
locus.of.control    1
self.concept       -1
motivation          0

Sum of squares and products for the hypothesis:
     [,1]
[1,] 8.656629

Sum of squares and products for error:
    [,1]
[1,] 432.5991
```

```
Multivariate Tests:
                  Df test stat approx F num Df den Df     Pr(>F)
Pillai             1 0.0196182 11.88638      1    594 0.00060546 ***
Wilks              1 0.9803818 11.88638      1    594 0.00060546 ***
Hotelling-Lawley   1 0.0200107 11.88638      1    594 0.00060546 ***
Roy                1 0.0200107 11.88638      1    594 0.00060546 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

H_0 wordt verworpen: $F(1, 594) = 11.886$ ($p < 0.001$).

Onderzoeksvraag 4

Zijn de effecten van `reading`, `writing` en `science` op `locus of control` gelijk?

$\Rightarrow H_0$: $\beta_{11} - \beta_{21} = \beta_{11} - \beta_{31} = 0$

- Stel $\boldsymbol{L} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \end{pmatrix}$

$$\boldsymbol{L\beta} = \begin{pmatrix} \beta_{11} - \beta_{21} & \beta_{12} - \beta_{22} & \beta_{13} - \beta_{23} \\ \beta_{11} - \beta_{31} & \beta_{12} - \beta_{32} & \beta_{13} - \beta_{33} \end{pmatrix}$$

- Stel $\boldsymbol{M} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ en $\boldsymbol{K} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$\begin{aligned} \boldsymbol{L\beta M} &= \boldsymbol{K} \\ \begin{pmatrix} \beta_{11} - \beta_{21} \\ \beta_{11} - \beta_{31} \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

```
> L1 <- matrix(
+   c(0, 1, -1, 0, 0, 0,
+     0, 1, 0, -1, 0, 0),
+   nrow = 2,
+   byrow = TRUE)
> M <- matrix(c(1, 0, 0),
+             nrow = 3)
> K <- matrix(c(0, 0),
```

```
+             nrow = 2)
> onderzoeksvraag4_1 <- lht(modelmulti,
+                           hypothesis.matrix = L1,
+                           P = M,
+                           rhs = K,
+                           verbose = TRUE)

Hypothesis matrix:
    [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    1   -1    0    0    0
[2,]    0    1    0   -1    0    0

Right-hand-side matrix:
     [,1]
[1,]    0
[2,]    0

Estimated linear function (hypothesis.matrix %*% coef - rhs):
[1] 0.0003595715 0.0067431421

> print(onderzoeksvraag4_1)

Response transformation matrix:
[,1]
locus.of.control    1
self.concept        0
motivation          0

Sum of squares and products for the hypothesis:
[,1]
[1,] 0.5945669

Sum of squares and products for error:
[,1]
[1,] 218.8562

Multivariate Tests:
                 Df test stat  approx F num Df den Df  Pr(>F)
Pillai            2 0.0027093 0.8068602      2    594 0.44675
```

```
Wilks             2 0.9972907 0.8068602      2    594 0.44675
Hotelling-Lawley  2 0.0027167 0.8068602      2    594 0.44675
Roy               2 0.0027167 0.8068602      2    594 0.44675
```

We vinden geen evidentie tegen H_0 aangezien $F(2, 594) = 0.807(p = .447)$.

Onderzoeksvraag 4

We bekijken nu een alternatieve formulering voor de onderzoeksvraag:
$\Rightarrow H_0$: $\beta_{11} - \beta_{21} = \beta_{21} - \beta_{31} = 0$

- Stel $\boldsymbol{L} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}$

$$\boldsymbol{L\beta} = \begin{pmatrix} \beta_{11} - \beta_{21} & \beta_{12} - \beta_{22} & \beta_{13} - \beta_{23} \\ \beta_{21} - \beta_{31} & \beta_{22} - \beta_{32} & \beta_{23} - \beta_{33} \end{pmatrix}$$

- Stel $\boldsymbol{M} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ en $\boldsymbol{K} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$\begin{aligned} \boldsymbol{L\beta M} &= \boldsymbol{K} \\ \begin{pmatrix} \beta_{11} - \beta_{21} \\ \beta_{21} - \beta_{31} \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

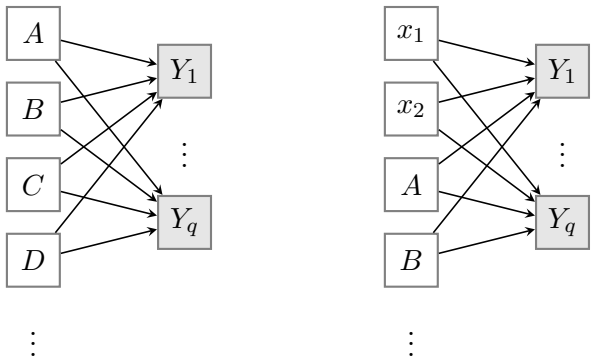
```
> L2 <- matrix(
+   c(0, 1, -1, 0, 0, 0,
+     0, 0, 1, -1, 0, 0),
+   nrow = 2,
+   byrow = TRUE)
>
>
> M <- matrix(c(1, 0, 0),
+             nrow = 3)
>
> K <- matrix(c(0, 0),
+             nrow = 2)
>
> onderzoeksvraag4_2 <- lht(modelmulti,
```

```
+                           hypothesis.matrix = L2,
+                           P = M,
+                           rhs = K,
+                           verbose = TRUE)

Hypothesis matrix:
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    1   -1    0    0    0
[2,]    0    0    1   -1    0    0

Right-hand-side matrix:
      [,1]
[1,]    0
[2,]    0

Estimated linear function (hypothesis.matrix %*% coef - rhs):
[1] 0.0003595715 0.0063835706

> print(onderzoeksvraag4_2)

Response transformation matrix:
                   [,1]
locus.of.control    1
self.concept        0
motivation          0

Sum of squares and products for the hypothesis:
         [,1]
[1,] 0.5945669

Sum of squares and products for error:
      [,1]
[1,] 218.8562

Multivariate Tests:
                 Df test stat  approx F num Df den Df  Pr(>F)
Pillai            2 0.0027093 0.8068602      2    594 0.44675
Wilks             2 0.9972907 0.8068602      2    594 0.44675
Hotelling-Lawley  2 0.0027167 0.8068602      2    594 0.44675
```

```
Roy                 2 0.0027167 0.8068602      2    594 0.44675
```

We krijgen identieke resultaten voor de toets wanneer de nulhypothese op een alternatieve manier geformuleerd wordt.

5 MANOVA en MANCOVA



5.1 Multivariate (co)variantie-analyse

Multivariate variantie-analyse (MANOVA, Multivariate ANalysis Of VAriance) is de multivariate uitbreiding van anova. Dit betekent dat alle nominale onafhankelijke variabelen van nominaal niveau zijn. Hierbij toetst men dus in essentie een gelijkheid van gemiddelden tussen groepen maar voor meerdere uitkomsten ($q > 1$).

Net zoals het geval is bij variantie-analyse, werd historisch veel belang gehecht aan MANOVA maar technisch is dit een bijzonder geval van het multivariaat lineair regressiemodel.

Analoog is multivariate covariantie-analyse (ANCOVA, Multivariate ANalysis Of COVAriance) de multivariate uitbreiding van ancova, waarbij er een combinatie is van predictoren van minstens intervalniveau en predictoren van nominaal niveau.

5.2 Voorbeeld

Als we in het voorbeeld over het onderzoek bij studenten naar de samenhang tussen academische vaardigheden en een aantal psychologische constructen (zie sectie 2) enkel kijken

naar het effect van opleidingsniveau, dan kunnen de onderzoeksvraag en de hypotheses m.b.t. het effect van deze variabele ook als volgt uitgedrukt worden:

- Zijn de gemiddelde locus of control, self-concept en motivation gelijk over de 3 opleidingsniveaus?
- Formeel kunnen we dit schrijven als:

$$H_0 : \begin{cases} \mu_{11} = \mu_{21} = \mu_{31} \\ \mu_{12} = \mu_{22} = \mu_{32} \\ \mu_{13} = \mu_{23} = \mu_{33} \end{cases}$$

  waarbij μ_{11} gemiddelde locus of control voor opleidingsniveau 1, μ_{21} gemiddelde locus of control voor opleidingsniveau 2, ...

  Deze gemiddelden worden soms celgemiddelden genoemd.

Deze hypothese kan getoetst worden via multivariaat lineaire regressie met locus of control, self-concept en motivation als afhankelijke variabelen en opleidingsniveau als factor. Aangezien de hypotheses hier uitgedrukt kunnen worden in termen van celgemiddelden, spreekt men soms van MANOVA. Indien ook gecontroleerd wordt voor continue predictoren (zoals in voorgaande analyses), spreekt men van MANCOVA.

Het Veralgemeend Lineair Model

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

1 Situering

Bij het univariaat algemeen lineair model hebben we het effect van een set van predictoren (onafhankelijke variabelen) op 1 uitkomst (afhankelijke variabele) van minstens intervalniveau gemodelleerd.

Dergelijke modellen kunnen echter niet gebruikt worden indien de uitkomst bvb.

- categorisch is,
- een aantal, een frequentie voorstelt,
- een proportie voorstelt.

In deze gevallen is er niet voldaan aan de assumpties van het algemeen lineair model (zoals lineariteit, normaal verdeelde residuen en een constante variantie van de residuen).

Logistische regressie laat wel toe om het effect van predictoren op een categorische uitkomst te modelleren.

Het logistische regressiemodel en het algemeen lineair model maken deel uit van de grotere familie van het **veralgemeend lineair model**.

2 Componenten

Veronderstel dat we het effect van p predictoren X_1, X_2, ..., X_p op 1 uitkomst Y wensen te modelleren.

We beschikken hierbij over n onafhankelijke observaties.

De klasse van veralgemeend lineaire modellen kan dan als volgt beschreven worden:

1. Y_1, Y_2, ..., Y_n zijn n onafhankelijk uitkomsten die een verdeling volgen die behoort tot de exponentiële familie van kansverdelingen.

   De verwachte waarde is $E(Y_i) = \mu_i$.

2. Er wordt een lineaire predictor op basis van de predictoren gebruikt:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

3. Een **link-functie** linkt de lineaire predictor aan de verwachtingswaarde μ_i:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

Merk opnieuw op dat Y een kansvariabele is terwijl de predictoren als constanten beschouwd worden.

Voorbeelden

- Bij het algemeen lineair model is de link-functie gelijk aan de identiteitsfunctie:

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

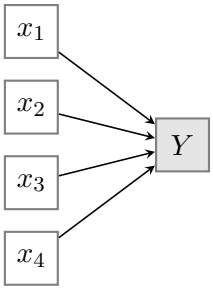
- Bij binaire logistische regressie is de uitkomst binair ($Y_i = 0$ of $Y_i = 1$). μ_i is de proportie waarin $Y_i = 1$ en wordt als π_i genoteerd. De link-functie is de logit-transformatie:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

3 Bijzondere gevallen

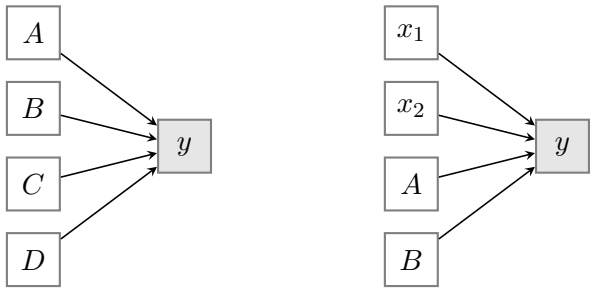
3.1 Het algemeen lineair model

Univariaat regressiemodel



Zowel Y als x zijn numerieke variabelen van minstens intervalniveau.

Anova en Ancova

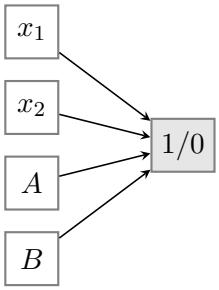


- A, B, C en D zijn *factoren*: categorische predictoren van nominaal niveau
- x_1, x_2 zijn *covariaten*: numerieke predictoren van minstens intervalniveau

3.2 Logistische regressie

De uitkomst is een categorische variabele (nominaal of ordinaal).

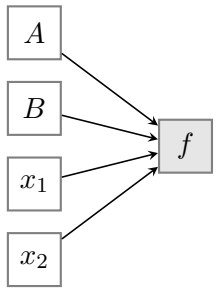
In deze cursus beschouwen we enkel binaire logistische regressie waarbij de uitkomst nominaal is en uit 2 niveaus bestaat (binaire uitkomst).



Een alternatief voor logistische regressie is **probit regressie**.

3.3 Poisson regressie

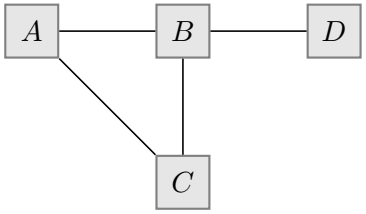
De uitkomst is een frequentie ('count') die een poisson verdeling volgt.



- $f = 0, 1, 2, \ldots$ is een frequentie

3.4 Loglineaire analyse

Dit is een speciaal geval van poisson regressie. Hierbij worden de celfrequenties gemodelleerd van meerwegs-tabellen. De bedoeling is om de associaties tussen verschillende nominale variabelen in kaart te brengen.



- alle variabelen zijn categorisch (nominaal of ordinaal)
- geen onderscheid tussen afhankelijke en onafhankelijke variabelen
- analyse naar de interrelaties tussen de variabelen

Logistische regressie

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

1 Inleiding

Bij het algemeen lineair model hebben we het verband bestudeerd tussen een afhankelijke variabele (uitkomst) van minstens intervalniveau en zowel continue als categorische onafhankelijke variabelen (predictoren).

Veronderstel nu dat de uitkomst Y nominaal is en slechts twee mogelijke waarden aanneemt: bvb. succes of falen, leven of sterven, aanvaardbaar of niet, etc. Deze twee mogelijke uitkomsten kunnen als 0 en 1 gecodeerd worden. Meestal wordt ervoor gekozen om het voorvallen van een bepaalde gebeurtenis, het 'succes' (ziekte, sterven, aandoening) te coderen als $Y = 1$. Bijgevolg betekent $Y = 0$ dat de gebeurtenis zich niet voordoet ('falen').

Het gemiddelde van de uitkomst is de proportie waarin $Y = 1$, en wordt de succeskans $\pi = P(Y = 1)$ genoemd.

Bij logistische regressie wordt de succeskans π gemodelleerd in functie van de onafhankelijke variabelen.

Voorbeeld

Wat is de kans dat een patiënt een ziekte overleeft als

- hij/zij in goede conditie is,
- hij/zij een bepaald type medicatie krijgt en
- hij/zij in een welbepaalde leeftijdscategorie zit?

In deze cursus beschouwen we **binaire logistische regressie**: Y is nominaal en bestaat uit twee niveaus.

Daarnaast bestaat er ook **multinomiale logistische regressie**, hierbij bestaat de nominale variabele Y uit meer dan twee niveaus. Bij **ordinale logistische regressie** is Y van ordinaal niveau.

De data die met logistische regressie geanalyseerd worden, kunnen zowel via experimentele als observationele studies verzameld zijn.

We werken met het softwarepakket R. Alle data en R-code voor de voorbeelden uit de cursus worden ter beschikking gesteld (zie verder).

2 Voorbeelden

2.1 Programmeertaak

We beschouwen data afkomstig van een kleinschalige studie uitgevoerd binnen een bedrijf (Kutner, Nachtscheim, Neter, & Li ; 2004).

Men gaat het effect na van ervaring met programmeren op het succesvol uitvoeren van een complexe taak binnen een bepaalde tijdspanne. Er nemen 25 personen deel aan de studie. De participanten hebben verschillende niveaus van ervaring, gemeten in aantal maanden.

Zowel de code die hoort bij de analyses (`programmeertaak.R`) als de data (`programmeertaak.csv`) zijn terug te vinden op Minerva.

De dataset bevat de volgende variabelen:

taak Indicator voor het al dan niet succesvol uitvoeren van de taak

`taak`=1 als de taak succesvol uitgevoerd wordt binnen de voorziene tijd

`taak`=0 als de taak niet succesvol uitgevoerd wordt binnen de voorziene tijd

ervaring Ervaring in aantal maanden

Hieronder wordt een deel van de data getoond (eerste 6 rijen):

```
> head(programmeer)
   taak ervaring
1    0       14
2    0       29
3    0        6
4    1       25
5    1       18
6    0        4
```

`taak` is de binaire afhankelijke variabele. Aan de hand van logistische regressie kunnen we de kans dat de taak succesvol uitgevoerd wordt binnen de voorziene tijd modelleren in functie van de `ervaring`.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 16, 26 en 29.

2.2 Moraliteit bij proeven met dieren

We beschouwen data uit een studie uit de volgende paper:

Wuensch, K. L., & Poteat, G. M. (1998). Evaluating the morality of animal research: Effects of ethical ideology, gender, and purpose. *Journal of Social Behavior and Personality, 13, 139-150.*

De data staan op de website van de eerste auteur. De code die hoort bij de analyses (`moraliteit.R`) is ter beschikking gesteld op Minerva en toont ook hoe de data ingelezen kunnen worden.

In deze studie wordt aan 315 participanten (psychologiestudenten aan een universiteit in de Verenigde Staten) gevraagd om te doen alsof ze in een commissie zetelen die een klacht behandelt tegen een (fictief!) onderzoek met dieren, uitgevoerd aan de universiteit. De klacht bevat een eenvoudige maar eerder emotionele beschrijving over het onderzoek met dieren. Het gaat om onderzoek waarbij katten een canule ingeplant krijgen in de hersenen waarlangs verschillende chemische producten toegediend kunnen worden. Na afloop van het onderzoek wordt een dissectie op de hersenen uitgevoerd. In de klacht wordt gesuggereerd dat het onderzoek ook via computersimulaties gedaan kan worden.

De participanten krijgen ook de verdediging van de onderzoeker te lezen waarin gesteld wordt dat de dieren op geen enkel moment pijn voelen, dat er geen andere mogelijkheid is om het onderzoek uit te voeren en waarin de voordelen van het onderzoek benadrukt worden.

De participanten worden toegewezen aan één van volgende vijf condities die verschillen in het vermelde doel van het onderzoek met dieren:

- het testen van chemische stoffen voor het ontwikkelen van haarverzorgingsproducten (*kosmetisch doel*)
- het contrasteren van twee hypothesen omtrent de werking van een bepaalde nucleus in de hersenen (*theorie*)
- het testen van een synthetisch groeihormoon dat potentieel kan leiden tot een grotere vleesproductie (*vleesconsumptie*)
- het zoeken naar een medicijn voor een hersenziekte die katten bedreigt (*dierengeneeskunde*)
- het evalueren van een potentieel geneesmiddel voor een bepaalde hersenziekte bij jong volwassenen (*medisch*)

Na het lezen van het materiaal wordt aan de participanten gevraagd om te beslissen of ze het onderzoek al dan niet zouden stopzetten.

Verder verstrekken de participanten ook enkele demografische gegevens en vullen ze vragenlijsten in om idealisme en relativisme te meten.

De onderzoekers zijn geïnteresseerd in het effect van `conditie`, `idealisme` en `relativisme` op het al dan niet laten stopzetten van het onderzoek met dieren.

Wij beschouwen de volgende variabelen uit de dataset:

decision Indicator voor de beslissing die genomen wordt (`stop` of `continue`)

idealism Idealisme, wordt als van intervalniveau verondersteld

De score voor idealisme wordt bekomen op basis van de antwoorden op 10 specifieke items (elk gemeten op een 9-punten schaal) van de Ethics Position Questionnaire.

Een voorbeeld van een dergelijk item is:

> A person should make certain that their actions never intentionally harm another even to a small degree. (1=completely disagree, 9=completely agree)

relatvsm Relativisme, wordt als van intervalniveau verondersteld

De score voor relativisme wordt bekomen op basis van de antwoorden op 10 specifieke items (elk gemeten op een 9-punten schaal) van de Ethics Position Questionnaire.

Een voorbeeld van een dergelijk item is:

> What is ethical varies from one situation and society to another. (1=completely disagree, 9=completely agree)

gender Gender (`Female` of `Male`)

conditie Voorgestelde doel van onderzoek met dieren (`cosmetic`, `theory`, `meat`, `veterin` of `medicine`)

Hieronder wordt een deel van deze data getoond (eerste 6 rijen):

```
> head(moraliteit[,c(1,2,3,4,12)])
  decision idealism relatvsm gender conditie
1     stop      8.2      5.1 Female cosmetic
2 continue      6.8      5.3   Male cosmetic
3 continue      8.2      6.0 Female cosmetic
4     stop      7.4      6.2 Female cosmetic
5 continue      1.7      3.1 Female cosmetic
6 continue      5.6      7.7   Male cosmetic
```

`decision` is de binaire afhankelijke variabele. Aan de hand van logistische regressie kunnen we de kans dat beslist wordt om het onderzoek met dieren stop te laten zetten, modelleren in functie van de `conditie`, `idealism`, en `relatvsm`, waarbij gecorrigeerd wordt voor `gender`.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 14, 20, 28, 30 en 32.

3 Enkelvoudige logistische regressie

We wensen de relatie tussen een uitkomst Y en 1 predictor X te modelleren.

Veronderstel dat we beschikken over n onafhankelijke observaties ($i = 1, \ldots, n$).

Y_i is de score (0 of 1) voor de uitkomst voor observatie i, x_i is de score voor de predictor.

Bij het voorbeeld rond de programmeertaak (zie sectie 2.1) is de uitkomst of de taak al dan niet succesvol werd uitgevoerd binnen de voorziene tijd, de predictor is het aantal maanden ervaring. Men wenst het effect van ervaring op de uitkomst te kennen. Men beschikt hierbij over $n = 25$ onafhankelijke observaties.

3.1 Bernoulli verdeling en odds

Y_i is een kansvariabele en volgt een Bernoulli verdeling:

$$\begin{aligned} \mathrm{E}(Y_i) &= P(Y_i = 1) = \pi_i \\ \mathrm{Var}(Y_i) &= \pi_i(1 - \pi_i) \end{aligned}$$

De odds dat $Y_i = 1$ is gelijk aan

$$\mathrm{ODDS} = \frac{\pi_i}{1 - \pi_i}.$$

π_i is een kans is en dus geldt dat $0 \leq \pi_i \leq 1$.

Een odds kleiner dan 1 betekent dat de succeskans π_i kleiner is dan $1/2$. Omgekeerd komt een odds groter dan 1 overeen met een succeskans π_i groter dan $1/2$.

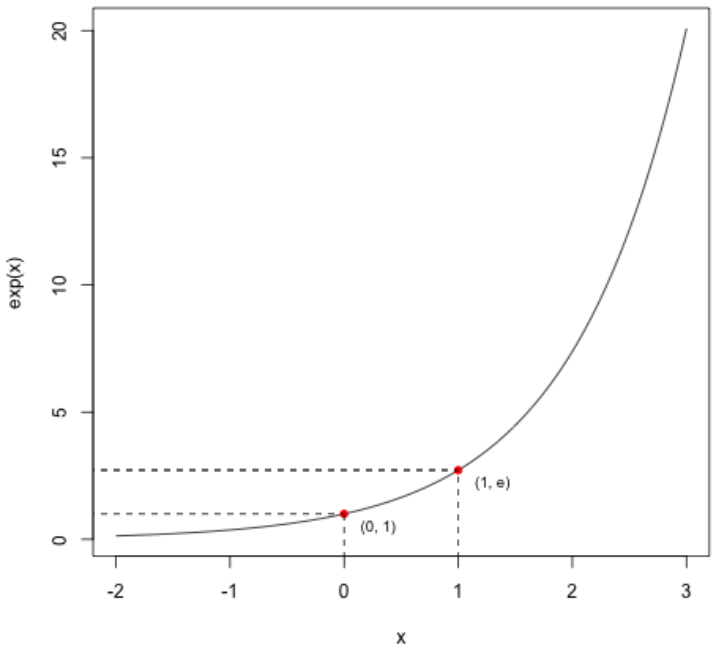
Voorbeeld

Veronderstel dat we via een survey te weten komen dat 20% van de werknemers in een groot bedrijf op regelmatige basis bijscholing volgt, dan is de odds dat een werknemer regelmatig bijscholing volgt $0.20/0.80 = 0.25 = 1/4$, men zegt vaak dat de odds 1 tegen 4 is. Analoog is de odds op geen regelmatige bijscholing 4 tegen 1.

3.2 Exponentiële en logaritmische functies

$e^x = \exp(x)$ stelt de exponentiële functie voor. e is het grondtal ($e = 2.71828\ldots$) zodat $\exp(x)$ gelijk is aan zijn eigen afgeleide.

De functie ziet er als volgt uit:



- Er geldt dat $\exp(0) = 1$ en $\exp(1) = e$.
- We zien dat $\exp(x)$ steeds positief is ($\exp(x) > 0$) en toeneemt als x toeneemt.
- Enkele rekenregels:

$$\begin{aligned} \exp(a+b) &= \exp(a) \times \exp(b) \\ \exp(a-b) &= \frac{\exp(a)}{\exp(b)} \end{aligned}$$

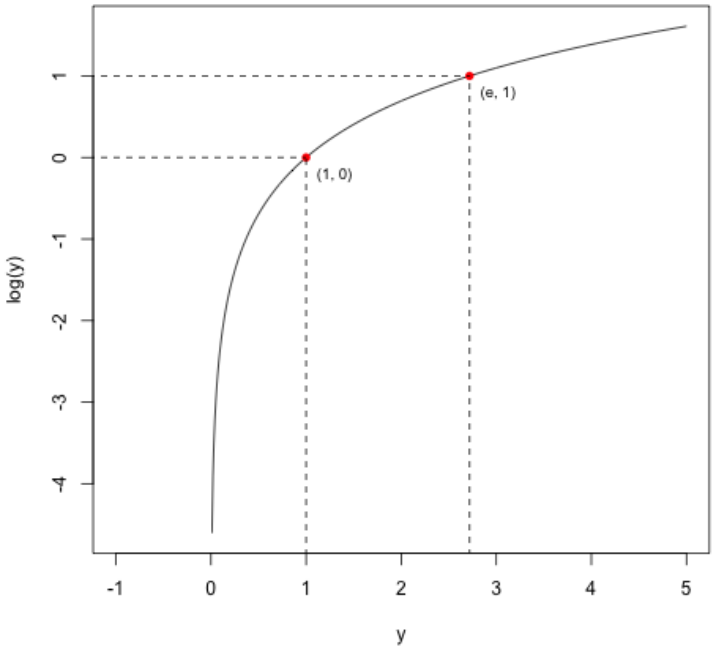
De exponentiële functie is beschikbaar in (statistische) software en in de meeste wetenschappelijke rekenmachines.

De inverse van de exponentiële functie is de $\log_e$-transformatie (ln) of kortweg log-transformatie. Dit betekent dat

$$\log(y) = x \Leftrightarrow y = \exp(x)$$

Merk op dat de logaritmische functie slechts gedefinieerd is voor $y > 0$ aangezien er altijd geldt dat $\exp(x) > 0$.

De functie ziet er als volgt uit:



- $\log(y)$ kan elke waarde tussen $-\infty$ en $+\infty$ aannemen.
- $\log(1) = 0$ en $\log(e) = 1$
- Enkele rekenregels:

$$\begin{aligned} \log(a \times b) &= \log(a) + \log(b) \\ \log\left(\frac{a}{b}\right) &= \log(a) - \log(b) \end{aligned}$$

Logaritmische functies zijn ook beschikbaar in (statistische) software en in de meeste wetenschappelijke rekenmachines. Ga wel steeds na welk grondtal gebruikt wordt. Voor de logaritmische functie met grondtal e wordt soms de notatie ln gebruikt terwijl log dan de logaritmische functie met grondtal 10 aanduidt.

In deze cursus beschouwen we de logaritmische functie met grondtal e (i.e. de natuurlijke logaritme) en duiden deze aan met log.

3.3 Het model voor logistische regressie

Het doel is om π_i te modelleren in functie van x_i.

Bij het voorbeeld rond de programmeertaak (sectie 2.1) betekent dit het modelleren van de kans om de taak succesvol uit te voeren in functie van ervaring.

Bij enkelvoudige lineaire regressie wordt de verwachte waarde van de uitkomst Y als volgt gemodelleerd in functie van de predictor:

$$E(Y_i|x_i) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, n.$$

Wanneer Y_i echter binair is, moeten we ervoor zorgen dat $0 \leq \pi_i \leq 1$ terwijl de lineaire functie $\beta_0 + \beta_1 x_i$ niet beperkt is tot de waarden tussen 0 en 1. Om een zinvol model te bekomen, maakt men bij logistische regressie gebruik van de $\log_e$-transformatie (ln) of kortweg de log-transformatie van de odds dat $Y_i = 1$, namelijk:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, n. \tag{1}$$

Net als bij lineaire regressie, is Y_i een kansvariabele en wordt x_i als een constante beschouwd.

Aangezien de odds elke positieve waarde kan aannemen, kan de logaritme van de odds eender welke waarde aannemen (zowel positief als negatief).

Een negatieve waarde van de logaritme stemt overeen met een odds die kleiner is dan 1 en dus met een succeskans π_i kleiner dan $1/2$.

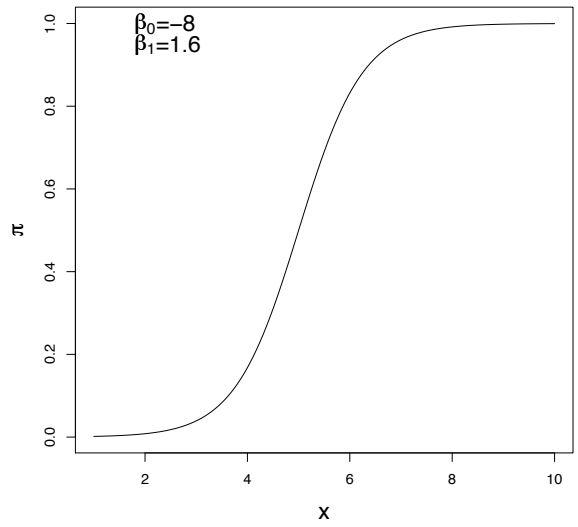
Een positieve waarde van de logaritme stemt overeen met een odds groter dan 1 en dus een π_i groter dan $1/2$.

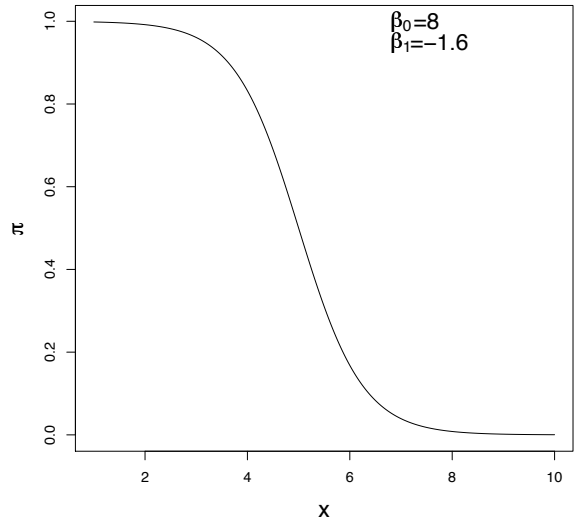
De logaritme van de odds wordt vaak kortweg de **log-odds** of **logit** genoemd.

Uit (1) volgt dat de succeskans π_i gelijk is aan

$$\pi_i = P(Y_i = 1|x_i) = E(Y_i|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad i = 1, \ldots, n. \tag{2}$$

De onderstaande figuren tonen 2 voorbeelden van deze functie:





De functie benadert in realistische gevallen de vorm van een (omgekeerde) S en is bijna lineair in het middenstuk. Het teken van β_1 bepaalt of π stijgt of daalt bij een verhoging in x.

3.4 Interpretatie parameters

We gaan na wat er gebeurt indien de predictor X met 1 eenheid stijgt. We doen dit door de odds dat $Y_i = 1$ te bekijken voor x_i (ODDS_1) en voor $x_i + 1$ (ODDS_2). Er geldt dat

$$\begin{aligned} \text{ODDS}_1 &= \frac{\pi_i(x_i)}{1-\pi_i(x_i)} = \exp(\beta_0 + \beta_1 x_i) \\ \text{ODDS}_2 &= \frac{\pi_i(x_i+1)}{1-\pi_i(x_i+1)} = \exp(\beta_0 + \beta_1(x_i+1)) = \exp(\beta_0 + \beta_1 x_i) \times \exp(\beta_1) \end{aligned}$$

We zien dus dat

$$\frac{\text{ODDS}_2}{\text{ODDS}_1} = \exp(\beta_1)$$

Met andere woorden: iedere toename van de predictor X met 1 eenheid zorgt voor een verandering met factor $\exp(\beta_1)$ van de odds dat $Y_i = 1$. Het is dus eerder de factor $\exp(\beta_1)$ dan de parameter β_1 zelf die interpreteerbaar is in een logistisch regressiemodel.

Een verhouding (ratio) van 2 odds, is een **odds ratio**.

Merk op dat

- $\exp(\beta_1) = 1$ indien $\beta_1 = 0$
- $\exp(\beta_1) > 1$ indien $\beta_1 > 0$
- $\exp(\beta_1) < 1$ indien $\beta_1 < 0$

$\beta_1 = 0$ geeft dus aan dat de odds dat $Y_i = 1$ constant blijft (en dus de kans dat $Y_i = 1$) indien de predictor verandert, $\beta_1 > 0$ geeft aan dat de odds (en bijgevolg de kans) groter wordt indien de predictor toeneemt en $\beta_1 < 0$ betekent dat de odds (en bijgevolg de kans) kleiner wordt indien de predictor toeneemt.

$\exp(\beta_0)$ is de odds dat $Y_i = 1$ voor $x_i = 0$.

4 Meervoudige logistische regressie

4.1 Het model

Het enkelvoudig logistisch regressiemodel in (2) kan makkelijk uitgebreid worden naar het geval waarbij er meerdere onafhankelijke variabelen / predictoren zijn.

In de studie rond moraliteit bij proeven met dieren (zie sectie 2.2), is de uitkomst of al dan niet beslist wordt om het onderzoek stop te laten zetten. Men wenst het effect van `conditie`, `idealism`, en `relatvsm` op de uitkomst te kennen. Men beschikt hierbij over $n = 315$ onafhankelijke observaties.

Met p predictoren X_1, X_2, ..., X_p wordt het logistisch regressiemodel:

$$\begin{aligned} \pi_i &= P(Y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ip}) = E(Y_i|x_{i1}, x_{i2}, \ldots, x_{ip}) \\ &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})} \quad i = 1, \ldots, n. \end{aligned} \tag{3}$$

x_{i1}, x_{i2}, ..., x_{ip} stellen de waarden voor de p predictoren bij observatie i voor en worden als constanten beschouwd.

Net zoals bij model (2), heeft deze functie een (omgekeerde) S-vorm.

De onafhankelijke variabelen kunnen p verschillende predictoren zijn of sommige kunnen, net zoals bij lineaire regressie, interactie-effecten voorstellen.

Merk op dat, analoog aan (1) geldt:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad i = 1, \ldots, n. \tag{4}$$

4.2 Interpretatie parameters

We gaan opnieuw na wat er gebeurt met de odds dat $Y_i = 1$ wanneer predictor X_ℓ met 1 eenheid toeneemt terwijl de overige predictoren constant blijven:

$$\begin{aligned} \text{ODDS}_1 &= \frac{\pi_i(x_{i1}, x_{i2}, \ldots, x_{i\ell}, \ldots, x_{ip})}{1 - \pi_i(x_{i1}, x_{i2}, \ldots, x_{i\ell}, \ldots, x_{ip})} \\ &= \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_\ell x_{i\ell} + \ldots + \beta_p x_{ip}) \\ \text{ODDS}_2 &= \frac{\pi_i(x_{i1}, x_{i2}, \ldots, x_{i\ell} + 1, \ldots, x_{ip})}{1 - \pi_i(x_{i1}, x_{i2}, \ldots, x_{i\ell} + 1, \ldots, x_{ip})} \\ &= \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_\ell (x_{i\ell} + 1) + \ldots + \beta_p x_{ip}) \\ &= \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_\ell x_{i\ell} + \ldots + \beta_p x_{ip}) \times \exp(\beta_\ell) \end{aligned}$$

We zien dus opnieuw dat

$$\frac{\text{ODDS}_2}{\text{ODDS}_1} = \exp(\beta_\ell)$$

Als alle andere predictoren constant blijven, verandert de odds dat $Y_i = 1$ met factor $\exp(\beta_\ell)$ met iedere toename van de predictor X_ℓ ($\ell = 1, \ldots, p$) met 1 eenheid.

De odds dat $Y_i = 1$ is gelijk aan $\exp(\beta_0)$ indien alle predictoren 0 zijn.

5 Nominale predictoren

Net als bij het algemeen lineair model kunnen ook nominale variabelen de rol van predictor vervullen. Opnieuw geldt dat we een nominale predictor met I niveaus moeten hercoderen tot $I - 1$ nieuwe hulpveranderlijken die we vervolgens in het regressiemodel kunnen stoppen.

- **Dummy-codering**: de bijhorende regressiecoëfficiënten geven het verschil aan t.o.v. een referentiecategorie.
- **Effect-codering**: de bijhorende regressiecoëfficiënten geven het verschil aan t.o.v. het gemiddelde over de niveaus van de variabele.

In de studie rond moraliteit bij proeven met dieren (zie sectie 2.2) is `conditie` een nominale variabele met 5 niveaus. Het standaard coderingsschema dat gehanteerd wordt in R is dummy-codering.

```
> class(moraliteit$conditie)
[1] "factor"
> contrasts(moraliteit$conditie)
           cosmetic theory meat veterin
medicine        0      0    0       0
cosmetic        1      0    0       0
theory          0      1    0       0
meat            0      0    1       0
veterin         0      0    0       1
```

In dit coderingsschema fungeert `conditie=medicine` als referentiecategorie.

Ook `gender` is een nominale variabele.

```
> class(moraliteit$gender)
[1] "factor"
> contrasts(moraliteit$gender)
         Male
Female    0
Male      1
```

In bovenstaand coderingsschema wordt dummy-codering gehanteerd waarbij `gender=Female` het referentieniveau is.

6 Parameterschattingen

Om de optimale parameterschattingen B_0, B_1, ..., B_p voor de logistische regressie te vinden, wordt het principe van de maximale aannemelijkheid (maximum likelihood) gehanteerd. Dit betekent dat de waarden voor B_0, B_1, ..., B_p gekozen worden die het meest aannemelijk zijn, gegeven de data in de steekproef.

De methode houdt in dat een 'likelihood function' (waarschijnlijkheidsfunctie) of kortweg likelihood wordt opgesteld die een functie is van de te schatten parameter(s) en de data en dat deze gemaximaliseerd wordt in functie van de parameters.

Er moet een stelsel van niet-lineaire vergelijkingen opgelost worden. Hiervoor worden efficiënte numerieke iteratieve technieken gebruikt (Fisher scoring). Dergelijke berekeningen zijn vrij complex maar de resultaten kunnen op een vrij eenvoudige manier m.b.v. statistische software bekomen worden.

Naast de schatters voor de parameters wordt ook andere nuttige informatie berekend, zoals de schattingen voor de variabiliteit van deze schatters en de **model deviance**.

De model deviance geeft aan hoe goed het logistisch model erin slaagt om de invloed van de predictoren op de uitkomst weer te geven (cfr. fout kwadratensom bij het algemeen lineair model). Anders gezegd: de model deviance drukt uit hoe goed de succeskansen π_i benaderd kunnen worden door een logistische transformatie. Hoe kleiner de deviance, hoe beter de fit van het model.

Bij binaire logistische regressie is de deviance gelijk aan -2 keer de logaritme van de likelihood van de data onder het model, of kortweg -2 keer de log likelihood (LL):

$$\text{Deviance} = -2 \times LL$$

Eenmaal de schatters gevonden zijn, kunnen deze waarden in (3) ingevuld worden om de

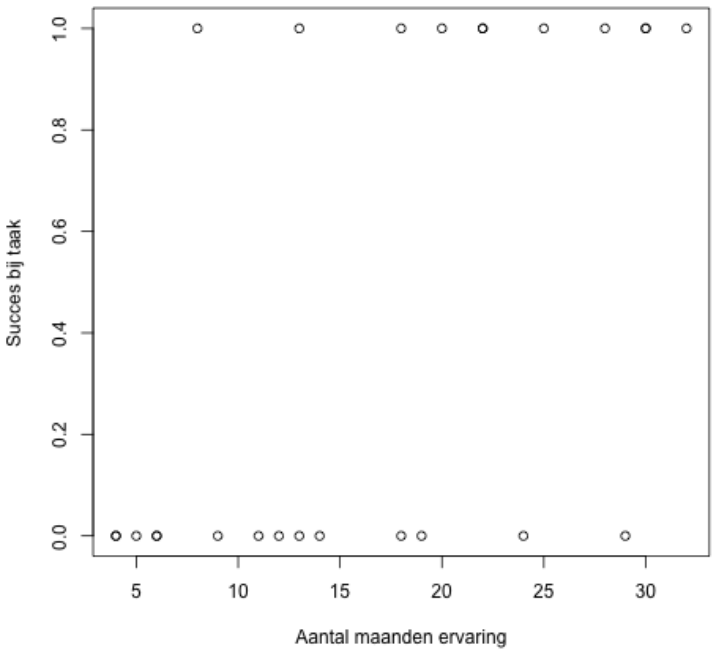
gefitte waarden te bekomen:

$$\hat{\pi}_i = \frac{\exp(B_0 + B_1 x_{i1} + B_2 x_{i2} + \ldots + B_p x_{ip})}{1 + \exp(B_0 + B_1 x_{i1} + B_2 x_{i2} + \ldots + B_p x_{ip})} \qquad i = 1, \ldots, n. \tag{5}$$

6.1 Voorbeeld: programmeertaak

We beschouwen de data van de studie met de programmeertaak (zie sectie 2.1).

Als we een spreidingsdiagram maken van het al dan niet succesvol uitvoeren van de taak in functie van het aantal maanden ervaring, krijgen we:



Deze figuur is niet erg informatief omdat de uitkomst binair is. Het zou interessanter zijn om de kans op een succesvolle taak in functie van de ervaring te tekenen.

Een binaire logistische regressie-analyse kan in R uitgevoerd worden aan de hand van het commando `glm` (Generlized Linear Model) waarbij aangegeven wordt dat de uitkomst Y

binomiaal is (i.e. Bernouilli verdeeld bij binaire uitkomst) en dat de logit van de verwachte uitkomst gemodelleerd wordt. In dit geval is de logit de *link-functie*, dit is de functie die de verwachte waarde van de uitkomst Y_i linkt aan de lineaire predictor $\beta_0 + \beta_1 x_i$.

De output is als volgt:

```
> model_programmeer <- glm(taak ~ ervaring,
+                family = binomial(link = "logit"),data=programmeer)
> summary(model_programmeer)

Call:
glm(formula = taak ~ ervaring, family = binomial(link = "logit"),
data = programmeer)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
             Estimate  Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
ervaring     0.16149    0.06498   2.485   0.0129 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 25.425  on 23  degrees of freedom
AIC: 29.425

Number of Fisher Scoring iterations: 4
```

Bij de kolom onder `Estimate`, kunnen we aflezen dat $b_0 = -3.05970$ en $b_1 = 0.16149$. In R krijg je niet standaard de exponent van de geschatte coëfficiënten. Die kunnen als volgt bekomen worden:

```
> exp(-3.05970)
[1] 0.04690176
```

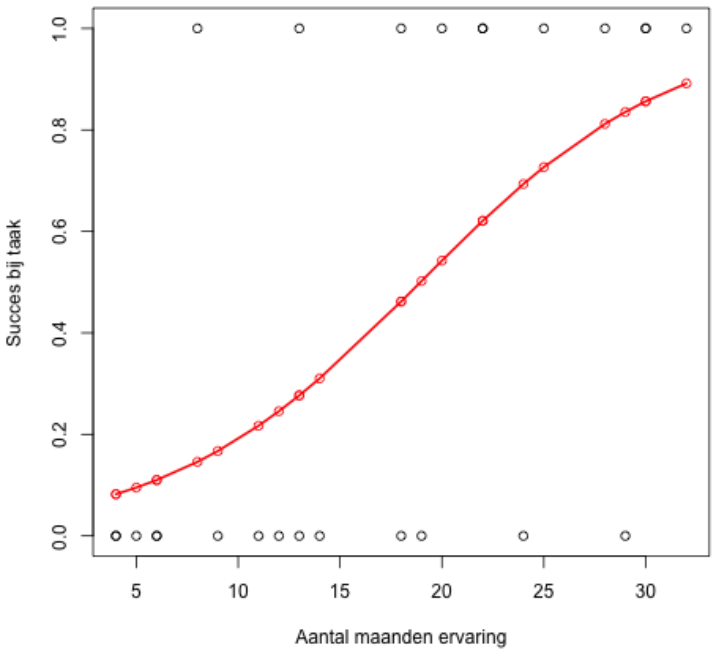
```
> exp(0.16149)
[1] 1.175261
> # Het is handiger om rechtstreeks de exponent te nemen van de  geschatte parameters
> # die opgeslaan zijn in het model-object in R:
> exp_coefs <- exp(model_programmeer$coef)
> exp_coefs
(Intercept)  ervaring
0.04690196  1.17525591
```

We vinden dus dat $\exp(b_0) = 0.047$ en $\exp(b_1) = 1.18$.

Dit betekent dat we schatten dat, indien de ervaring met 1 maand toeneemt, de odds dat de taak succesvol uitgevoerd wordt, toeneemt met factor 1.18. Of nog: we schatten dat de odds toeneemt met 18% voor iedere bijkomende maand ervaring.

De geschatte odds bij geen ervaring is gelijk aan $0.047 < 1$, wat betekent dat de geschatte kans om de taak goed uit te voeren zonder ervaring, kleiner is dan 50%.

De geschatte kans op een succesvolle taak in functie van ervaring, kan bepaald worden zoals in uitdrukking (5). Dit is de volle (rode) lijn die we toevoegen aan het spreidingsdiagram:



We zien dat de curve inderdaad een S-vorm heeft en dat de geschatte kans toeneemt indien de ervaring toeneemt.

De model deviance voor dit model lezen we af bij `Residual deviance` en is hier gelijk aan 25.425. We kunnen dit ook als volgt bekomen:

```
> summary(model_programmeer)$deviance
[1] 25.42457
```

Verschillende pseudomaten voor R^2 zoals bij het algemeen lineair model kunnen ook berekend worden. Hoewel deze maten niet geïnterpreteerd kunnen worden als de proportie van de variantie in de uitkomst die verklaard wordt door de predictoren, geven ze wel de sterkte weer van het verband tussen uitkomst en predictoren (maat tussen 0 en 1). In onderstaande output staan twee vaak gebruikte pseudomaten voor R^2 bij logistische regressie:

```
> # Cox & Snell
> R_CoSn
```

```
[1] 0.2987401
> # Nagelkerke (aanpassing Cox & Snell)
> R_Nag
[1] 0.4002599
```

De exacte berekeningen hiervoor zijn terug te vinden in de R-code op Minerva maar beschouwen we hier niet (hoeft niet gekend te zijn).

6.2 Voorbeeld: moraliteit bij proeven met dieren

We beschouwen de data van de studie rond moraliteit bij proeven met dieren (zie sectie 2.2).

We fitten eerst het logistisch regressiemodel waarbij we dummy-codering hanteren voor `conditie` en `gender` (zie de restrictieschema's in sectie 5).

Aangezien de uitkomst Y bij logistische regressie binair ($Y = 1$ of $Y = 0$) is, kan het commando `glm` gebruikt worden met een numerieke binaire variabele als uitkomst (zoals het geval is bij het voorbeeld rond de programmeertaak) of kan de uitkomst een `factor` zijn. In het laatste geval zal in R een dummy-variabele aangemaakt worden die de waarden 0 of 1 aanneemt. Onthou dat R altijd de kans modelleert dat $Y = 1$; kijk dus steeds goed na hoe de uitkomst gecodeerd is.

In dit voorbeeld vinden we:

```
> class(moraliteit$decision)
[1] "factor"
> contrasts(moraliteit$decision)
          continue
stop             0
continue         1
```

We zien dus dat R de kans om het onderzoek te laten doorgaan zal modelleren. We veranderen dit hier zodat de kans om het onderzoek te laten stopzetten gemodelleerd wordt:

```
> moraliteit$decision <- relevel(moraliteit$decision,ref="continue")
> contrasts(moraliteit$decision)
          stop
continue     0
stop         1
```

Via het commando `relevel` geven we aan dat we het referentieniveau $Y = 0$ wensen te veranderen. Dit is equivalent met wat we voorheen deden:

```
> moraliteit$decision <- factor(moraliteit$decision,levels=c("continue","stop"))
```

Wanneer we de uitkomst modelleren in functie van `conditie`, `idealism`, `relatvsm` en `gender`, krijgen we:

```
> model_moraliteit<-glm(decision~conditie+idealism+relatvsm+gender,
                        family = binomial(link = "logit"),data=moraliteit)
> summary(model_moraliteit)

Call:
glm(formula = decision ~ conditie + idealism + relatvsm + gender,
family = binomial(link = "logit"), data = moraliteit)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.3350  -0.9402   0.4645   0.8266   2.1564

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.2789     1.0330  -2.206  0.02737 *
conditiecosmetic   0.7095     0.4202   1.688  0.09136 .
conditietheory     1.1596     0.4278   2.710  0.00672 **
conditiemeat       0.8659     0.4244   2.040  0.04130 *
conditieveterin    0.5423     0.4098   1.323  0.18572
idealism           0.7012     0.1139   6.156 7.48e-10 ***
relatvsm          -0.3264     0.1267  -2.576  0.01000 *
genderMale        -1.2551     0.2766  -4.537 5.70e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 425.57  on 314  degrees of freedom
Residual deviance: 338.06  on 307  degrees of freedom
AIC: 354.06
```

```
Number of Fisher Scoring iterations: 4

> exp_coefs_moraliteit1 <- exp(model_moraliteit$coef)
> exp_coefs_moraliteit1
(Intercept) conditiecosmetic   conditietheory     conditiemeat  conditieveterin
0.1023991        2.0329474        3.1887299        2.3772438        1.7200166

 idealism          relatvsm        genderMale
 2.0160882        0.7214972     0.2850518
```

- $b_0 = -2.2789$, $\exp(b_0) = 0.1023991$. Dit betekent dat de geschatte odds op laten stopzetten gelijk is 0.10 bij vrouwen in de conditie met een voorgesteld medisch doel en een score van 0 op `idealism` en `relatvsm`.

- Bij de predictor `conditie` worden de verschillende condities vergeleken met conditie `medicine`.

  We zien dat de geschatte coëfficiënt bij `cosmetic` gelijk is aan $b_1 = 0.7095$ en $\exp(b_1) = 2.0329474$. Bijgevolg is de geschatte odds op laten stopzetten een factor 2.03 groter (i.e. ongeveer 2 maal zo groot) in de conditie waarbij het voorgestelde doel kosmetisch is t.o.v. de conditie met een voorgesteld medisch doel, als alle overige predictoren constant blijven.

  We zien dat de geschatte coëfficiënt bij `veterin` gelijk is aan $b_4 = 0.5423$ en $\exp(b_4) = 1.7200166$. Bijgevolg is de geschatte odds op laten stopzetten een factor 1.72 groter (i.e. een toename van 72%) in de conditie waarbij het voorgestelde doel dierengeneeskunde is t.o.v. de conditie met een voorgesteld medisch doel, als alle overige predictoren constant blijven.

  De overige coëfficiënten bij de verschillende condities kunnen op een analoge manier geïnterpreteerd worden.

- De geschatte coëfficiënt bij `idealism` is $b_5 = 0.7012$ en $\exp(b_5) = 2.0160882$. Binnen eenzelfde conditie en wanneer gender en relativisme constant blijven, schatten we dat de odds op laten stopzetten een factor 2.02 groter wordt (i.e. ongeveer verdubbelt) wanneer de score voor idealisme met 1 eenheid toeneemt.

- De geschatte coëfficiënt bij `relatvsm` is $b_6 = -0.3264$ en $\exp(b_6) = 0.7214972$. Binnen eenzelfde conditie en wanneer gender en idealisme constant blijven, schatten we dat de odds op laten stopzetten een factor 0.72 kleiner wordt (i.e. een afname van 28%) wanneer de score voor relativisme met 1 eenheid toeneemt.

- De geschatte coëfficiënt bij `gender` is $b_7 = -1.2551$ en $\exp(b_7) = 0.2850518$. Hier worden mannen vergeleken met vrouwen (referentiecategorie).

  Binnen eenzelfde conditie en bij een constante score voor idealisme en relativisme, schatten we dat de odds op laten stopzetten een factor 0.29 kleiner is (i.e. een afname van 71%) bij mannen dan bij vrouwen.

We fitten nu het logistisch regressiemodel waarbij we effect-codering hanteren voor `conditie` en `gender`.

```
> conditie_eff<-moraliteit$conditie
> contrasts(conditie_eff)<-contr.sum
> contrasts(conditie_eff)
          [,1] [,2] [,3] [,4]
medicine    1    0    0    0
cosmetic    0    1    0    0
theory      0    0    1    0
meat        0    0    0    1
veterin    -1   -1   -1   -1

> gender_eff<-moraliteit$gender
> contrasts(gender_eff)<-contr.sum
> contrasts(gender_eff)
        [,1]
Female    1
Male     -1

> model_moraliteit_effect<-glm(decision~conditie_eff+idealism+relatvsm+gender_eff,
+                              family = binomial(link = "logit"),data=moraliteit)
> summary(model_moraliteit_effect)

Call:
glm(formula = decision ~ conditie_eff + idealism + relatvsm +
gender_eff, family = binomial(link = "logit"), data = moraliteit)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.3350  -0.9402   0.4645   0.8266   2.1564

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.25094    0.99005  -2.274   0.0230 *
conditie_eff1 -0.65548    0.26438  -2.479   0.0132 *
conditie_eff2  0.05401    0.26728   0.202   0.8399
conditie_eff3  0.50415    0.27362   1.843   0.0654 .
conditie_eff4  0.21046    0.27219   0.773   0.4394
idealism       0.70116    0.11390   6.156 7.48e-10 ***
relatvsm      -0.32643    0.12673  -2.576   0.0100 *
gender_eff1    0.62754    0.13831   4.537 5.70e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 425.57  on 314  degrees of freedom
Residual deviance: 338.06  on 307  degrees of freedom
AIC: 354.06

Number of Fisher Scoring iterations: 4

> exp_coefs_moraliteit2 <- exp(model_moraliteit_effect$coef)
> exp_coefs_moraliteit2
(Intercept) conditie_eff1 conditie_eff2 conditie_eff3 conditie_eff4
0.1052999     0.5191943     1.0554948     1.6555706     1.2342515

 idealism      relatvsm      gender_eff1
 2.0160882     0.7214972     1.8730013
```

We zien dat de schattingen voor de effecten van `idealism` en `relatvsm` hetzelfde blijven; voor `conditie` en `gender` zijn deze gewijzigd aangezien die variabelen op een andere manier gecodeerd zijn. Bij effect-codering wordt er vergeleken met het gemiddelde over alle niveaus van de variabele.

- Bij de predictor `conditie` worden de verschillende condities vergeleken met het gemiddelde over condities.

  We zien dat de geschatte coëfficiënt bij `cond_eff1` gelijk is aan $b_1 = -0.65548$ en $\exp(b_1) = 0.5191943$. In het bovenstaand restrictieschema zien we dat deze parameter het verschil bekijkt tussen `conditie=medicine` met het gemiddelde over condities. Bijgevolg is de geschatte odds op laten stopzetten een factor 0.52 kleiner in de conditie waarbij het

voorgestelde doel medisch is in vergelijking met het gemiddelde over alle condities heen, als alle overige predictoren constant blijven.

We zien dat de geschatte coëfficiënt bij `conditie_eff4` gelijk is aan $b_4 = 0.21046$ en $\exp(b_4) = 1.2342515$. Bijgevolg is de geschatte odds op laten stopzetten een factor 1.23 groter (i.e. een toename van 23%) in de conditie waarbij het voorgestelde doel vleesconsumptie is in vergelijking met het gemiddelde over alle condities heen, als alle overige predictoren constant blijven.

De regressiecoëfficiënt voor `conditie=veterin` ontbreekt in de output; op basis van de output kunnen we niet rechtstreeks de vergelijking tussen `conditie=veterin` en het gemiddelde over condities heen aflezen gezien de hulpveranderlijken voor deze categorie allemaal op -1 gezet zijn. Bij effect-codering weten we echter dat som van de regressiecoëfficiënten die horen bij de verschillende niveaus gelijk is aan 0. Er geldt dus dat de geschatte parameter voor `conditie=veterin` gelijk is aan $0 - (-0.65548 + 0.05401 + 0.50415 + 0.21046) = -0.11314$. We vinden dus dat de geschatte odds op laten stopzetten een factor $\exp(-0.11314) = 0.89$ kleiner is in de conditie waarbij het voorgestelde doel dierengeneeskunde is in vergelijking met het gemiddelde over alle condities heen, als alle overige predictoren constant blijven.

- Bij de predictor `gender` worden de verschillende condities vergeleken met het gemiddelde over condities.

  De geschatte coëfficiënt is $b_7 = 0.62754$ en $\exp(b_7) = 1.8730013$. Hier worden vrouwen vergeleken met het gemiddelde over mannen en vrouwen.

  Binnen eenzelfde conditie en bij een constante score voor idealisme en relativisme, schatten we dat de odds op laten stopzetten een factor 1.87 groter is bij vrouwen in vergelijking met het gemiddelde over mannen en vrouwen heen.

  We kunnen afleiden dat de parameter die hoort bij `gender=Man` gelijk is aan $0 - 0.62754 = -0.62754$; bijgevolg schatten we dat de odds op laten stopzetten een factor $\exp(-0.62754) = 0.53$ kleiner is bij mannen in vergelijking met het gemiddelde, als alle overige predictoren constant blijven.

Beide modellen stellen hetzelfde logistisch regressiemodel voor; enkel de parameterisatie is gewijzigd waardoor de parameters een andere betekenis krijgen. We lezen af dat de deviance van het model gelijk is aan 338.06.

7 Toetsing

7.1 Toets voor 1 predictor: Wald toets

Wanneer we wensen te toetsen of het effect van een predictor statistisch significant is, kunnen we gebruik maken van een Wald toets. Dit is de tegenhanger van de F-toets bij het algemeen lineair model.

De nulhypothese H_0 die getoetst wordt, stelt dat de predictor geen invloed heeft op de uitkomst. Onder H_0 volgt de Wald toetsingsgrootheid een χ^2-verdeling met *df* vrijheidsgraden, waarbij *df* gelijk is aan het aantal parameters dat getoetst wordt.

We verwerpen H_0 indien de Wald statistiek groot wordt. Laat w^* de geobserveerde waarde voor de Wald statistiek voorstellen. De p-waarde is de kans om rechts van w^* te liggen onder de χ^2_{df}-verdeling:

$$p - \text{waarde} = P(X^2 \geq w^*) \text{ waarbij } X^2 \sim \chi^2_{df}$$

We verwerpen H_0 op het $\alpha \times 100\%$ significantieniveau indien de p-waarde kleiner is dan α.

7.1.1 Predictor van intervalniveau

Wanneer we wensen te toetsen of predictor ℓ (van minstens intervalniveau) uit het model weggelaten kan worden, toetsen we $H_0 : \beta_\ell = 0$ tegenover $H_1 : \beta_\ell \neq 0$ of equivalent: $H_0 : \exp(\beta_\ell) = 1$ tegenover $H_1 : \exp(\beta_\ell) \neq 1$.

Indien S_{B_ℓ} de standaardfout van de schatter B_ℓ voor β_ℓ voorstelt dan is de Wald toetsingsgrootheid in dit geval gelijk aan

$$W = \left(\frac{B_\ell}{S_{B_\ell}}\right)^2$$

Deze toetsingsgrootheid volgt onder H_0 een χ^2-verdeling met 1 vrijheidsgraad.

We hernemen het voorbeeld rond de programmeertaak (zie sectie 2.1). Het resultaat voor de Wald toets voor het effect van ervaring kan bekomen worden via het commando `Anova` (R-package `car`). Net zoals bij lineaire regressie geven we aan dat we met type III toetsen werken. Verder moeten we ook specifiëren dat we de resultaten voor de Wald toets wensen (`test="Wald"`).

```
> Anova(model_programmeer,type=3,test = "Wald")
```

```
Analysis of Deviance Table (Type III tests)

Response: taak
            Df  Chisq   Pr(>Chisq)
(Intercept)  1 5.9029    0.01512 *
ervaring     1 6.1760    0.01295 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Bij `Df` lezen we af dat het aantal vrijheidsgraden gelijk is aan 1. De geobserveerde Wald statistiek die hoort bij ervaring, lezen we af bij `Chisq` en is gelijk aan 6.176. De bijhorende p-waarde is gelijk aan 0.01295. Dit betekent dat de nulhypothese dat er geen verband is tussen ervaring en het uitvoeren van de taak verworpen wordt op het 5% significantieniveau (p-waarde kleiner dan 0.05).

We hernemen de output die we bekwamen via de `summary` van het gefitte logistisch regressiemodel:

```
> summary(model_programmeer)

Call:
glm(formula = taak ~ ervaring, family = binomial(link = "logit"),
data = programmeer)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
        Estimate  Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
ervaring     0.16149    0.06498   2.485   0.0129 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 25.425  on 23  degrees of freedom
AIC: 29.425
```

```
Number of Fisher Scoring iterations: 4
```

Aangezien het aantal vrijheidsgraden *df* (i.e. aantal te toetsen parameters) hier gelijk is aan 1, kunnen we het resultaat van de Wald toets ook in deze output aflezen (cfr. gelijkheid t-toets en F-toets bij lineaire regressie).

Op basis van de schatting voor het effect van `ervaring` (waarde bij `Estimate`) en de standaardfout (waarde bij `Std. Error`), kunnen we afleiden dat de Wald toetsingsgrootheid gelijk is aan $(0.16149/0.06498)^2 = 6.176$. De waarde bij `z value` is gelijk aan de schatting gedeeld door de standaardfout (0.16149/0.06498) en is dus gelijk aan de vierkantswortel van de Wald statistiek. Bij `Pr(>|z|)` lezen we de overeenkomstige p-waarde af.

7.1.2 Predictor van nominaal niveau

In het geval van een categorische predictor met I niveaus brengt de Wald toets alle $I-1$ hulpveranderlijken die coderen voor de predictor, mee in rekening. Met $I-1$ hulpveranderlijken volgt deze toetsingsgrootheid onder de nulhypothese dat de predictor geen effect heeft, een χ^2-verdeling met $I-1$ vrijheidsgraden.

We hernemen het voorbeeld rond moraliteit bij proeven met dieren (zie sectie 2.2). Om de effecten van de afzonderlijke predictoren te toetsen aan de hand van de Wald toets, maken we gebruik van `Anova`. Net zoals bij lineaire regressie, maken we bij de predictoren van nominaal niveau gebruik van effect-codering.

```
> Anova(model_moraliteit_effect,type=3,test="Wald")
Analysis of Deviance Table (Type III tests)

Response: decision
              Df   Chisq Pr(>Chisq)
(Intercept)    1  5.1691    0.02299 *
conditie_eff   4  8.1823    0.08512 .
idealism       1 37.8920  7.477e-10 ***
relatvsm       1  6.6342    0.01000 *
gender_eff     1 20.5863  5.700e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We observeren dat enkel het effect van `conditie` niet statistisch significant is op het 5%

significantieniveau (p-waarde is gelijk aan $0.085 > 0.05$). We zien dat de Wald toetsingsgrootheid voor `conditie` gelijk is aan 8.1823. Het aantal vrijheidsgraden is 4 aangezien er 4 parameters getoetst worden (er zijn 4 hulpveranderlijken).

7.2 Modelvergelijkingen: Likelihood Ratio toets

Om twee modellen A en B met elkaar te vergelijken waarbij model B genest is in model A, kunnen we gebruik maken van de Likelihood Ratio statistiek LR:

$$\begin{aligned} LR &= \text{Deviance Model } B - \text{Deviance Model } A \\ &= -2LL_B - (-2LL_A) \\ &= 2LL_A - 2LL_B \end{aligned}$$

Onder de nulhypothese dat beide modellen niet van elkaar verschillen, volgt deze toetsingsgrootheid een χ^2-verdeling met $df_B - df_A$ vrijheidsgraden. Net zoals bij het algemeen lineair model is $df_B - df_A$ het verschil van het aantal geschatte parameters onder model A en het aantal geschatte parameters onder model B.

Modelvergelijkingstoetsen voor logistische regressiemodellen kunnen in R uitgevoerd worden aan de hand van het commando `anova`. Hierbij dient gespecifieerd te worden dat het om een Likelihood Ratio toets gaat (`test="LRT"`).

Een toets voor 1 predictor is een speciaal geval van een modelvergelijkingstoets waarbij het model met de predictor vergeleken wordt met het model zonder de predictor. Merk echter op dat de Likelihood Ratio toets en de Wald toets niet equivalent zijn. De Wald toets kan op een relatief simpele manier berekend worden op basis van de schattingen van de regressieparameters en de overeenkomstige standaardfouten terwijl bij de Likelihood Ratio toets de likelihoods van de modellen die vergeleken worden, berekend moeten worden. In de praktijk geeft men de voorkeur aan de Likelihood Ratio toetsen tenzij dit computationeel moeilijk wordt. In grote steekproeven maakt het meestal niet veel uit welke toets gebruikt wordt, aangezien ze dan vaak tot dezelfde resultaten leiden. De resultaten voor modelvergelijkingstoetsen voor het effect van de afzonderlijke predictoren kunnen ook bekomen worden via het commando `Anova` met `test="LR"`.

7.2.1 Voorbeeld: programmeertaak

Herneem het voorbeeld rond de programmeertaak (zie sectie 2.1). Het effect van ervaring kan als volgt getoetst worden via het commando `anova`:

```
> nullmod <- glm(taak ~ 1,
+                family = binomial(link = "logit"),data=programmeer)
> anova(nullmod,model_programmeer,test="LRT")
Analysis of Deviance Table

Model 1: taak ~ 1
Model 2: taak ~ ervaring
Resid. Df Resid. Dev   Df Deviance Pr(>Chi)
1        24     34.296
2        23     25.425  1   8.8719 0.002896 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Hierbij wordt het model met `ervaring` vergeleken met het model zonder `ervaring` (dit is het nulmodel aangezien er geen andere predictoren in het model zitten). Bij `Df` lezen we af dat het aantal vrijheidsgraden gelijk is aan 1 (aantal te toetsen parameters). We stellen vast dat de geobserveerde toetsingsgrootheid gelijk is aan 8.8719 (`Deviance`). De bijhorende p-waarde is gelijk aan 0.0029. We besluiten bijgevolg dat het model met ervaring significant beter is dan het model zonder ervaring. Merk op dat de resultaten inderdaad anders zijn dan bij de Wald toets voor ervaring, we komen echter wel tot dezelfde conclusie. Dit resultaat kan ook bekomen worden via het commando `Anova` met `test=`"`LR`":

```
> Anova(model_programmeer,type=3,test = "LR")
Analysis of Deviance Table (Type III tests)

Response: taak
         LR Chisq Df Pr(>Chisq)
ervaring   8.8719  1   0.002896 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid lezen we hier af bij `LR Chisq`.

7.2.2 Voorbeeld: moraliteit bij proeven met dieren

Herneem het voorbeeld rond moraliteit bij proeven met dieren (zie sectie 2.2).

Via het commando `Anova` met `test=`"`LR`" kunnen we ook de resultaten voor het toetsen van de

effecten van de predictoren aan hand van de Likelihood Ratio toets bekomen. We doen dit op basis van het model waarbij effect-codering voor de nominale predictoren gebruikt wordt.

```
> Anova(model_moraliteit_effect,type=3,test="LR")
Analysis of Deviance Table (Type III tests)

Response: decision
             LR Chisq Df Pr(>Chisq)
conditie_eff    8.443  4   0.076630 .
idealism       46.340  1  9.943e-12 ***
relatvsm        6.911  1   0.008567 **
gender_eff     21.546  1  3.454e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De resultaten voor de verschillende predictoren zijn niet identiek aan deze bekomen aan de hand van de Wald toets; de conclusies blijven wel hetzelfde.

Als we het model met en zonder `idealism` en `relatvsm` willen vergelijken (en dus het effect van beide predictoren simultaan willen toetsen), kunnen we gebruik maken van `anova`:

```
> model_moraliteit0<-glm(decision~conditie+gender,
+                        family = binomial(link = "logit"),data=moraliteit)
> model_moraliteit<-glm(decision~conditie+idealism+relatvsm+gender,
+                        family = binomial(link = "logit"),data=moraliteit)
> anova(model_moraliteit0,model_moraliteit,test="LRT")
Analysis of Deviance Table

Model 1: decision ~ conditie + gender
Model 2: decision ~ conditie + idealism + relatvsm + gender
     Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       309     389.78
2       307     338.06  2   51.722 5.87e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid is gelijk aan 51.722 met een p-waarde die heel klein is. We verwerpen bijgevolg de nulhypothese die stelt dat beide predictoren geen effect hebben op de uitkomst en kiezen het model dat beide predictoren wel bevat. Het aantal vrijheidsgraden is hier gelijk aan 2, aangezien we 2 parameters toetsen.

8 Interactie (moderatie)

Net zoals bij het lineair regressiemodel kunnen interactiermen toegevoegd worden in het logistisch regressiemodel. Op die manier kan gemodeleerd worden dat het effect van een predictor varieert naargelang het niveau van een andere predictor.

Veronderstel dat er 2 predictoren X_1 en X_2 van intervalniveau zijn en veronderstel volgend logistisch regressiemodel met interactieterm:

$$\begin{aligned} \pi_i &= P(Y_i = 1|x_{i1}, x_{i2}) = E(Y_i|x_{i1}, x_{i2}) \\ &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2})} \quad i = 1, \ldots, n. \end{aligned} \tag{6}$$

We gaan na wat er gebeurt indien de predictor X_1 met 1 eenheid stijgt terwijl X_2 constant blijft. We doen dit door de odds dat $Y_i = 1$ te bekijken voor x_{i1} (ODDS_1) en voor $x_{i1} + 1$ (ODDS_2). Er geldt dat

$$\begin{aligned} \text{ODDS}_1 &= \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}) \\ \text{ODDS}_2 &= \exp(\beta_0 + \beta_1 (x_{i1} + 1) + \beta_2 x_{i2} + \beta_3 (x_{i1} + 1) x_{i2}) \end{aligned}$$

Er kan aangetoond worden dat

$$\frac{\text{ODDS}_2}{\text{ODDS}_1} = \exp(\beta_1 + \beta_3 x_{i2}) = \exp(\beta_1) \exp(\beta_3 x_{i2})$$

Het effect van X_1 wijzigt naargelang het niveau van X_2.

- Wanneer $X_2 = 0$, stelt $\exp(\beta_1)$ de factor voor waarmee de odds wijzigt indien X_1 met 1 eenheid toeneemt.
- De odds ratio $\frac{\text{ODDS}_2}{\text{ODDS}_1}$ (effect van 1 eenheid toename in X_1) wijzigt met factor $\exp(\beta_3)$ als X_2 met 1 eenheid toeneemt.

We kunnen interacties dus op een analoge manier interpreteren als bij lineaire regressie.

Herneem het voorbeeld rond moraliteit bij proeven met dieren (zie sectie 2.2).

Ter illustratie voegen we een interactierm tussen `idealism` en `gender` toe. We hanteren de dummy-codering voor de nominale predictoren.

```
> model_moraliteit_int<-glm(decision~conditie+idealism+relatvsm+gender+idealism:gender,
```

```
+                        family = binomial(link = "logit"),data=moraliteit)
> summary(model_moraliteit_int)

Call:
glm(formula = decision ~ conditie + idealism + relatvsm + gender +
idealism:gender, family = binomial(link = "logit"), data = moraliteit)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.4326  -0.9311   0.4333   0.8244   2.3093

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.9837     1.1800  -2.528  0.01146 *
conditiecosmetic      0.7118     0.4245   1.677  0.09357 .
conditietheory        1.1974     0.4316   2.775  0.00553 **
conditiemeat          0.8362     0.4269   1.959  0.05013 .
conditieveterin       0.5505     0.4151   1.326  0.18478
idealism              0.8204     0.1481   5.540 3.02e-08 ***
relatvsm             -0.3337     0.1269  -2.630  0.00853 **
genderMale            0.8016     1.4953   0.536  0.59192
idealism:genderMale  -0.3243     0.2328  -1.393  0.16367
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 425.57  on 314  degrees of freedom
Residual deviance: 336.14  on 306  degrees of freedom
AIC: 354.14

Number of Fisher Scoring iterations: 4
```

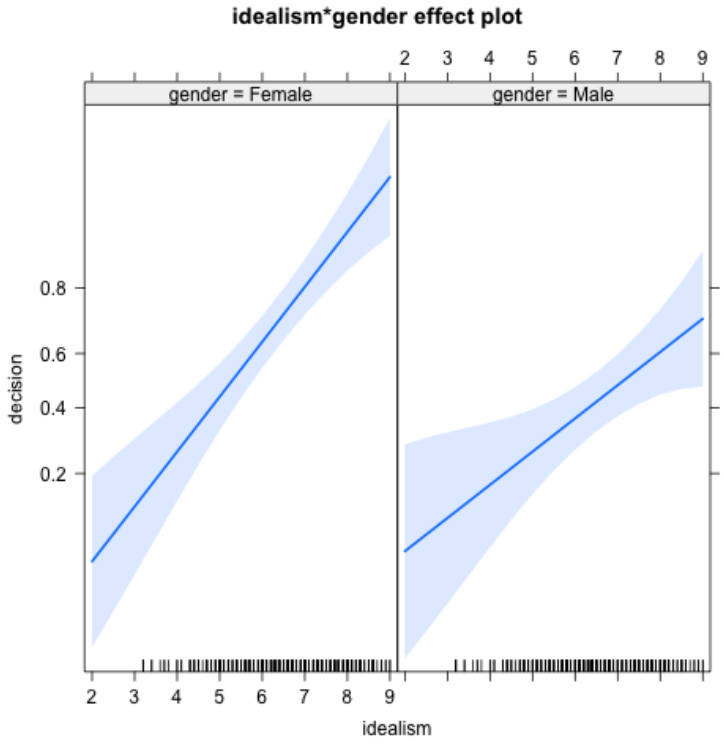
We zien dat de geschatte coëfficiënt voor het hoofdeffect van `idealism` gelijk is aan 0.8204. Wegens het interactie-effect met `gender` stelt dit hoofdeffect het effect van `idealism` voor bij het referentieniveau van gender, namelijk vrouwen. Als alle overige predictoren constant blijven, zien we dus een positief effect van `idealism` bij vrouwen. Wanneer de score op `idealism` bij vrouwen met 1 eenheid toeneemt terwijl alle overige predictoren constant blijven, schatten we dat de odds op laten stopzetten met factor $\exp(0.8204) = 2.27$ toeneemt.

Dit effect is kleiner voor mannen, de geschatte parameter voor het interactie-effect is gelijk aan -0.3243 wat betekent dat het geschatte effect van `idealism` op de logit-schaal bij mannen gelijk is aan 0.8204-0.3243=0.4961. Wanneer de score op `idealism` bij mannen met 1 eenheid toeneemt terwijl alle overige predictoren constant blijven, schatten we dat de odds op laten stopzetten met factor $\exp(0.4961) = 1.64$ toeneemt.

De factor waarmee de odds wijzigt indien `idealism` met 1 eenheid toeneemt is dus inderdaad een factor $\exp(-0.3243) = 0.723$ kleiner bij mannen dan bij vrouwen ($0.723 \times 2.27 = 1.64$).

Het interactie-effect kan ook weergegeven worden aan de hand van de functie `effect` in het R package `effects`.

```
> library(effects)
> effect("idealism:gender",model_moraliteit_int)

idealism*gender effect
           gender
idealism     Female      Male
2          0.06289544 0.0725350
4          0.25719659 0.1741804
5          0.44023081 0.2572646
7          0.80226405 0.4829732
9          0.95440264 0.7158498
```



Merk op dat de Y-as op bovenstaande figuur de geschatte kans dat $Y = 1$ voorstelt. Let echter op de schaal: deze is niet proportioneel waardoor het effect van `idealism` lineair lijkt.

Via een modelvergelijkingstoets (Likelihood Ratio toets) gaan we na of het interactie-effect statistisch significant is.

```
> anova(model_moraliteit,model_moraliteit2,test="LRT")
Analysis of Deviance Table

Model 1: decision ~ conditie + idealism + relatvsm + gender
Model 2: decision ~ conditie + idealism + relatvsm + gender + idealism:gender
   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       307     338.06
2       306     336.14  1   1.9174   0.1661
```

We vinden dat $p = 0.166$ en bijgevolg is het interactie-effect niet statistisch significant op het 5% significantieniveau.

9 Referenties

Kutner, M., Nachtscheim, C., Neter, J., Li, W. (2004). *Applied Linear Statistical Models, 5th edidtion*. McGraw-Hill.

Moore, D.S., McCabe, G.P. (2005). *Introduction to the practice of statistics*. W.H. Freeman & Co Ltd.

Meta-analyse

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

1 Situering

Replicatie is belangrijk voor de wetenschap en het opbouwen van kennis. Er kan een onderscheid gemaakt worden tussen de volgende soorten replicatie (Thompson, 1996):

- Interne replicatie: uitgevoerd door oorspronkelijke onderzoekers

  Statistische resampling en cross-validatie

- Externe replicatie: uitgevoerd door andere onderzoekers

  Dit houdt in dat andere steekproeven onderzocht worden op andere tijdstippen of plaatsen. Veel replicatiestudies zijn zogenaamde *constructreplicaties* waarbij de nieuwe onderzoeker het geschikte design, de metingen en de data-analyse selecteert om de generaliseerbaarheid van het originele resultaat te testen. Hierbij wordt een strikte imitatie van de originele studie vermeden (dit in tegenstelling tot *operationele replicaties*).

  Een potentieel probleem hierbij is dat variatie in methodes, metingen of steekproeven geassocieerd kunnen zijn met het fenomeen dat bestudeerd wordt.

Bij een meta-analyse worden resultaten van verschillende studies geïntegreerd om onderliggende populatie-effecten te bestuderen. Het hoofddoel is om systematisch bij te houden op welke facetten studies verschillen zodat tussen-studie variatie verklaard kan worden.

In de literatuur wordt geijverd voor het feit dat men bij een data-analyse niet enkel gebruik dient te maken van statistische significantie (zie hoofdstuk 'Statistische en praktische significantie') maar ook van effectgroottes en betrouwbaarheidsintervallen in individuele studies en van meta-analyses voor het integreren van resultaten van verscheidene studies. Omwille van de belangrijke bijdrage van verschillende wetenschappers bij het ontwikkelen van geschikte methodes, worden meta-analyses tegenwoordig veel frequenter aangewend om populatie-effecten te onderzoeken.

2 Wat is een meta-analyse?

Meta-analyse is een statistische techniek om het gemiddelde en de variantie van onderliggende populatie-effecten te schatten op basis van een **set empirische studies** die dezelfde onderzoeksvraag proberen te beantwoorden.

Voorbeelden

- Is een cognitieve gedragstherapie effectief bij het behandelen van angst bij kinderen en adolescenten?
- Hebben ooggetuigen vertekende herinneringen van gebeurtenissen?
- Verschilt temperament volgens geslacht?
- Hoe kunnen zwangere vrouwen geholpen worden bij het stoppen met roken?
- ...

Dergelijke voorbeelden illustreren de diversiteit van vragen die psychologen onderzoeken om menselijk gedrag te bestuderen.

- De antwoorden op deze vragen kunnen verschillen van studie tot studie.
- Replicatie is een manier om om te gaan met problemen die gecreëerd worden door meetfouten.
- Verschillende onderzoeken rond hetzelfde onderwerp laten toe om dezelfde of gelijkaardige vragen te beantwoorden via het gebruik van meta-analyse.
- Praktische relevantie voor theorievorming: een meta-analyse kan duidelijkheid scheppen over het feit of verbanden al dan niet situatief bepaald zijn. De gecompliceerde theorievorming in de psychologie kan hierdoor vermoedelijk afgeslankt worden.

  Dit maakt eenduidige richtlijnen voor de praktijk mogelijk.

Meta-analyse betreft de **integratie** en **cumulatie** van onderzoeksbevindingen. De effectgroottes die in iedere studie berekend worden op basis van ruwe scores, worden nu gezien als de data die geanalyseerd moeten worden. Het gaat dus om een *analyse van analyses*. Zoals gezien in het hoofdstuk 'Statistische en praktische significantie', zijn er verschillende manieren om een effectgrootte te kwantificeren: de pearson correlatiecoëfficiënt (r), een gestandaardiseerd gemiddeld verschil (d), odds ratio's, relatieve risico's,.... We concentreren ons hier op d en r en we hebben reeds gezien dat d in r omgezet kan worden en vice versa.

- Samenhangen: correlaties
  Notatie: r (steekproef) of ρ (populatie)
- Effecten van ingrepen (manipulaties)
  Notatie: d (steekproef) of δ (populatie)

Empirisch stelt men vaak vast dat verschillende studies resulteren in verschillende waarden voor de effectgroottes. In de psychologie is het eerder regel dan uitzondering dat verschillende onderzoeken die peilen naar hetzelfde verband tot wisselende vaststellingen komen.

Mogelijke redenen hiervoor zijn:

- Statistische artefacten
  - Steekproeffouten
  - Betrouwbaarheidsfouten
  - Beperkte variatie in scores
- Wisselende constructen

  Voorbeeld

  Succes, prestatie etc. kunnen verschillende betekenissen hebben.
- Variatie in operationalisatie

  Voorbeeld

  Stress wordt in de ene studie gemeten aan de hand van een vragenlijst en in een andere studie via een gestructureerd interview.
- Werking van derde variabelen: *moderatoren*

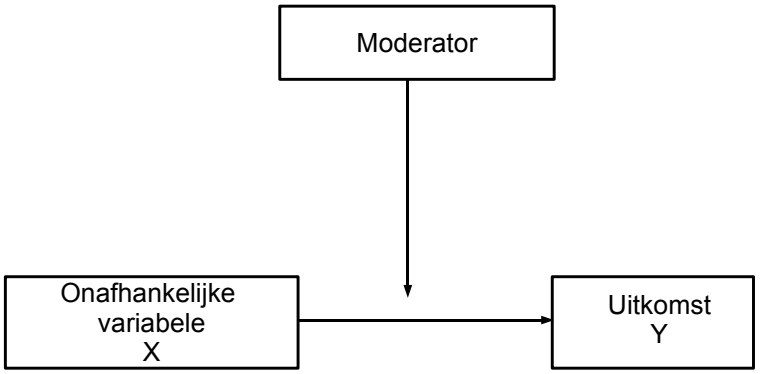
  Het is niet ongewoon dat ook wanneer de hiervoor genoemde mogelijkheden niet van toepassing zijn, er toch nog een aanzienlijke variatie in de gevonden verbanden/effecten blijft bestaan. Deze variatie wordt typisch toegeschreven aan situatiekenmerken die van studie tot studie verschillen.

  Voorbeelden

  - De relatie tussen toewijding en tevredenheid wordt gerelateerd aan verschillen in bedrijfscultuur.
  - Een cognitieve gedragstherapie blijkt een groter effect te hebben bij adolescenten met eetstoornissen als de therapie ook groepstherapie inhoudt dan wanneer er geen groepstherapie is.

Dergelijke situatiekenmerken worden als **moderator variabelen** betiteld, m.a.w. het variërend resultaat wordt toegeschreven aan de verschillende 'waarden' van de moderator in de verschillende studies.

Moderatie = interactie



Het effect van X op Y hangt af van het niveau van de moderator. De grootte of het teken van de samenhang tussen X en Y varieert naargelang de waarden van de moderator.

3 Doelstellingen

Bij een meta-analyse wordt methodologie op een even strikte manier toegepast bij het bespreken van wetenschappelijke literatuur als bij empirisch onderzoek.

De doelstellingen zijn:

- het schatten van de **ware (gemiddelde) samenhang**, ρ ($\overline{\rho}$) of het ware (gemiddeld) effect, δ ($\overline{\delta}$), na correctie voor statistische artefacten

- het schatten van de **variantie** van de ware samenhang, σ^2_ρ, of de variantie van het ware effect, σ^2_δ, na correctie voor statistische artefacten

Het algemeen schema dat hierbij gehanteerd wordt, is het volgende:

- de geschatte effectgroottes van de individuele studies beschouwen,
- deze schattingen converteren naar eenzelfde meeteenheid (bvb. d omzetten in r)
- en een **gewogen** gemiddelde effectgrootte berekenen met bijhorende standaardfout.

Op basis van deze geschatte gecombineerde effectgrootte en zijn precisie kunnen betrouwbaarheidsintervallen voor de ware (gemiddelde) effectgrootte geconstrueerd worden.

- Op die manier kan nagegaan worden hoe groot een effect is en of dit effect al dan niet statistisch significant is.
- De gewichten die aan de resultaten van een welbepaalde studie gegeven worden zijn vaak gebaseerd op de steekproefgrootte (of meer bepaald op de standaardfout van de studie-specifieke effectgrootte) aangezien dit de accuraatheid m.b.t. de steekproeftrekking reflecteert.
- De similariteit van effectgroottes tussen studies kan nagegaan worden d.m.v. een toets voor homogeniteit (lage power!) of via het schatten van σ^2_ρ (σ^2_δ).
  Als $\sigma^2_\rho > 0$ ($\sigma^2_\delta > 0$), stelt men zich de vraag of deze variantie toegeschreven dient te worden aan de werking van moderatoren.

  De variatie in de gevonden verbanden moet eerst uitgezuiverd worden voor niet-inhoudelijke artefacten (steekproeffouten, betrouwbaarheidsfouten, beperkte variatie in scores). Als blijkt dat de resterende variatie verwaarloosbaar is t.o.v. de oorspronkelijke, is er geen enkele reden om deze variatie toe te schrijven aan de werking van moderatoren.

4 Voorbeeld 1: Evaluative Conditioning in Humans

Evaluative Conditioning in Humans: A Meta-Analysis
Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010)
Psychological Bulletin, *136*, 390-421

Evaluatieve conditionering (EC) verwijst naar een verandering in de valentie van een prikkel (stimulus) omwille van de koppeling met een andere positieve of negatieve prikkel.

Het samen voorkomen van een neutrale prikkel A met een affectief geladen prikkel B zorgt ervoor dat A een evaluatieve betekenis krijgt die congruent is met de waarde van Y.

A: geconditioneerde stimulus ('conditioned stimulus' CS)

B: ongeconditioneerde stimulus ('unconditioned stimulus' US)

- EC is een vorm van Pavloviaans leren.
- EC meet enkel de verandering in evaluatieve respons op de CS, i.e. de verandering in het 'aangenaam vinden' van de stimulus.

Open vragen / kwesties:

- Generaliseerbaarheid van EC?
- Onderscheid met andere vormen van leren?
- Aard van theoretische mechanismes?

Immers: veel tegenstrijdige bevindingen in literatuur

→ meta-analyse

Drie doelstellingen van deze studie:

1. Schatten van een globaal EC effect over een brede range van studies
2. Heterogeniteit in onderzoek rond EC bestuderen
   - Kwalitatief: coderen van een set van procedurale variaties
   - Kwantitatief: schatten van heterogeniteit
3. Indien heterogeniteit: onderzoek van moderatoren
   Variëren EC effecten volgens procedurale karakteristieken?
   Tabel 2 geeft een overzicht van potentiële moderatoren.
   Het coderen gebeurt door 2 verschillende personen; hierbij wordt de intercodeurbetrouwbaarheid berekend.

5 Verschillende stappen bij een meta-analyse

Stap 1 Literatuuronderzoek

Stap 2 Inclusiecriteria bepalen

Stap 3 Effectgroottes berekenen

Zie ook hoofdstuk 'Statistische en praktische significantie'

Stap 4 Meta-analyse uitvoeren

Herinner:

- Het hoofddoel is het schatten van populatie-effecten via het combineren van effectgroottes.
- Meer specifiek: er wordt een **gewogen** gemiddelde beschouwd.
- Het 'gewicht' is typisch een functie van de steekproefgrootte.
  Logisch: hoe groter een steekproef, hoe preciezer de schatter voor de effectgrootte.
- Daarnaast worden ook betrouwbaarheidsintervallen geconstrueerd.

In essentie zijn er 2 manieren om een meta-analyse te conceptualiseren (basismodellen): het **fixed effects model** en het **random effects model**.

Fixed effects

Hierbij veronderstelt men dat de studies in de meta-analyse een steekproef zijn uit een populatie met een vaste (maar ongekende) effectgrootte. De effectgrootte op populatieniveau is gelijk voor alle studies in de meta-analyse.

Random effects

Hier gaat men er van uit dat effectgroottes in de populatie variëren van steekproef tot steekproef. Dit betekent dat elke studie in de meta-analyse afkomstig is van een populatie met een andere effectgrootte dan elke andere studie in de meta-analyse. We kunnen hier op de volgende manier over nadenken: de populatie-effectgroottes zijn als het ware een steekproef van een universum van mogelijke effecten, een 'superpopulatie' (Becker, 1996; Hedges, 1992).

Fixed of random?

Het kiezen van een geschikt model hangt af van

- de assumpties over de populaties waarvan de studies een steekproef zijn,
- het type van conclusie die de onderzoeker beoogt.

In veel gevallen is het niet redelijk om te veronderstellen dat empirische data voldoen aan de assumptie van vaste parameters (Barrick & Mount, 1991).

Wat betreft het type conclusie die de onderzoeker wenst te trekken (Hedges & Vevea, 1998):

- Fixed effects modellen zijn enkel geschikt voor conclusies die gelden voor de studies die in de meta-analyse opgenomen zijn (conditionele inferentie).
- Random effects modellen laten besluitvorming toe die verder reikt dan enkel de studies die opgenomen zijn in de analyse (niet-conditionele inferentie).

In de gedragswetenschappen wenst men meestal verder te generaliseren dan de opgenomen studies en in dat opzicht zijn random effects modellen meer geschikt. Ondanks deze argumentering vonden Hunter en Schmidt (2000) in de journal 'Psychological Bulletin' (toen recentelijke) 21 meta-analytische studies waarbij een fixed model gebruikt werd tegenover geen enkele die een random effects model gebruikte.

Het theoretische gevolg daarvan is dat de toetsen gebaseerd op de gecombineerde effectgroottes de kans op een type I fout niet controleren. Dit betekent hier concreet dat de kans op een type I fout groter is dan het vooropgestelde significantieniveau α.

Een toets voor de homogeniteit van effectgroottes kan aangewend worden om na te gaan of de populatie-effectgroottes fixed of random zijn. Wanneer men op basis van deze toets de nulhypothese van homogene effectgroottes niet kan verwerpen, zou men kunnen besluiten dat de effectgroottes van de steekproeven min of meer equivalent zijn en dat de hypothese van homogene effectgroottes aannemelijk is.

Een heel belangrijke opmerking hierbij is dat deze toetsen misleidend kunnen zijn aangezien er in de literatuur al vaak geargumenteerd is dat hun power om variatie in effectgroottes te detecteren, laag is (zie bvb. Hedges & Pigott, 2002).

Notatie

Veronderstel dat we een meta-analyse op basis van correlatiecoëfficiënten uitvoeren. We beschouwen k studies in de meta-analyse ($i = 1, \ldots, k$):

- r_i: geobserveerde (gemeten) samenhang/correlatie in studie i

- ρ_i: ware samenhang/correlatie in studie i (populatieniveau)
- n_i: het aantal observaties in studie i

We kunnen schrijven dat:

$$r_i = \rho_i + e_i \qquad (i = 1, \ldots, k)$$

waarbij e_i de steekproeffout voorstelt. Steekproeven leiden immers steeds tot steekproeffouten en daarom wijkt het studieresultaat r_i af van de populatiewaarde ρ_i. De verwachting van de foutcomponent e_i, $E(e_i)$, wordt gelijk aan 0 verondersteld. De variantie van e_i wordt genoteerd als $\sigma^2_{e_i}$. In wat volgt maken we ook de veronderstelling dat e_i normaal verdeeld is. De variantie σ^2_e stelt het volgende gewogen gemiddelde voor:

$$\sigma^2_e = \frac{\sum_{i=1}^{k} n_i \sigma^2_{e_i}}{\sum_{i=1}^{k} n_i}.$$

De variantie in de geobserveerde correlaties noteren we met σ^2_r. Op basis van het bovenstaande kunnen we stellen dat $\sigma^2_r = \sigma^2_\rho + \sigma^2_e$ waarbij σ^2_ρ de variantie van de *populatiecorrelaties* voorstelt, m.a.w. de variantie in de geobserveerde correlaties kan opgedeeld worden in 2 componenten. Het onderscheid tussen de fixed en random effects modellen is dan als volgt:

- Bij een fixed effects model veronderstelt men dat $\rho_i = \rho$ (en dus dat $\sigma^2_\rho = 0$), m.a.w. de variantie in geobserveerde correlaties is volledig te wijten aan steekproeffouten.
- Bij een random effects model stelt men dat $\rho_i = \overline{\rho} + u_i$, met $u_i \sim N(0, \sigma^2_\rho)$ zodat $\rho_i \sim N(\overline{\rho}, \sigma^2_\rho)$, m.a.w. de variantie in populatiecorrelaties en steekproeffouten zijn 2 bronnen van variantie die aan de basis liggen van de variantie in geobserveerde correlaties.

Het technische verschil tussen beide modellen (fixed en random) is het berekenen van de standaardfout van de schatter van de gecombineerde effectgrootte.

- Fixed effects modellen beschouwen enkel de variatie binnen de studie (Engels: within-study variability).
- Random effects modellen brengen de fouten mee in rekening van het trekken van een steekproef uit populaties die op hun beurt getrokken zijn uit een superpopulatie. De standaardfout bevat dus 2 componenten: de variatie binnen de de studie en de variatie ontstaan uit verschillen tussen studies (Engels: between-study variability).

Gevolg: de geschatte standaardfouten bij random effects modellen zijn (veel) groter dan deze bij fixed effects modellen wanneer effectgroottes daadwerkelijk heterogeen zijn over de studies en daarom zijn statistische toetsen voor gecombineerde effecten conservatiever. Dit betekent dat men minder snel de nulhypothese van geen effect kan verwerpen.

Methodes voor meta-analyse

In deze cursus bespreken we

- de methode ontwikkeld door Hedges en Olkin (1985): zij ontwikkelden een fixed en random effects model voor het combineren van effectgroottes.
- de methode van Hunter en Schmidt (1990) waarbij een random effects model beschouwd wordt.

Stap 5 Meer geavanceerde analyses

- Analyse van moderatoren
- Schatten van publicatiebias

6 Literatuuronderzoek

- Gebruik van electronische databases om artikels rond een bepaald onderwerp te vinden:

  ISI web of knowledge, PubMed, PsycINFO en PsycLIT

- Niet enkel bestaande artikels rond een bepaald onderwerp omvatten studies die potentieel in de meta-analyse gebruikt kunnen worden maar ook auteurs die actief zijn in een bepaald domein moeten geïdentificeerd worden.

- Doe ook een forward search: welke auteurs citeren artikels in dit domein?

- Let op voor *publication bias* (file-drawer probleem).

  Statistisch significante resultaten hebben meer kans om gepubliceerd te worden. Daarom kan een meta-analyse populatie-effecten mogelijks overschatten.

  Niet alleen gepubliceerde studies moeten opgenomen worden maar de zoektocht moet ook uitgebreid worden naar papers van conference proceedings en experten in het vakgebied moeten gecontacteerd worden zodat ook niet-gepubliceerde bevindingen ter beschikking gesteld kunnen worden.

  Hoe? via e-mail of via het posten van een boodschap bij een specifieke 'newsgroup'.

Voorbeeld 1

Gepubliceerde of in-press artikels, dissertaties, hoofdstukken uit boeken en ongepubliceerde manuscripten:

- PsycLIT en PsycINFO: databanken voor psychologische onderzoekspapers
- Dissertation abstracts: databanken voor doctoraatsverhandelingen
- E-mails naar verscheidene mailing lists voor cognitieve, sociale en persoonlijkheidspsychologie & een lijst met participanten van een workshop over evaluatieve conditionering in 2007

Start: Februari 2007; Einde: December 2008

Resultaat: 282 citaties; na exclusie: 253 studies afkomstig van 145 citaties (zie verder)

7 Criteria voor inclusie

- Slecht uitgevoerd onderzoek mee in meta-analyse opnemen, kan ook vertekende resultaten opleveren.
- Zorg dat je geen appelen met peren vergelijkt.
- Inclusiecriteria hangen af van de onderzoeksvraag en methodologische kwesties.

  Voorbeeld

  Bij onderzoek naar cognitieve gedragstherapie: duidelijke definitie van wat een dergelijke therapie inhoudt en bvb. exclusie van studies die geen gebruik maken van controlegroepen, etc.
- Gebruik transparante criteria en wees daarin duidelijk bij het rapporteren!
- Het is ook mogelijk om studies te classificeren volgens methodologie, bvb. al dan niet gebruiken van controlegroepen, en op die manier nagaan of dergelijke zaken moderatoren zijn voor de effectgroottes.

  Voorbeeld

  Heeft het type van controlegroep een invloed op de sterkte van het effect van een cognitieve behandelingstherapie?

Voorbeeld 1

Na initiële exclusie van 84 citaties, 5 criteria:

1. De studie is experimenteel of quasi-experimenteel voor EC waarbij 1 of meerdere CS's gekoppeld worden met 1 of meerdere US's met een gegeven valentie.
2. In de EC studies wordt onderzocht of het koppelen van de CS met de US de valentie wijzigt van de CS.
3. De metingen die de studies minstens moeten bevatten, worden gedefinieerd (zie ook figuur 2 uit de paper).
4. Studies moeten de data rapporteren zodat ten minste 1 relevante effectgrootte gecodeerd kan worden.

   Auteurs werden eerst gecontacteerd om nodige maar ontbrekende data te voorzien; toch kon voor een deel van de studies geen effectgrootte berekend worden.
5. Data worden niet opgenomen indien ze al opgenomen zijn in een andere studie om duplicatie te vermijden.

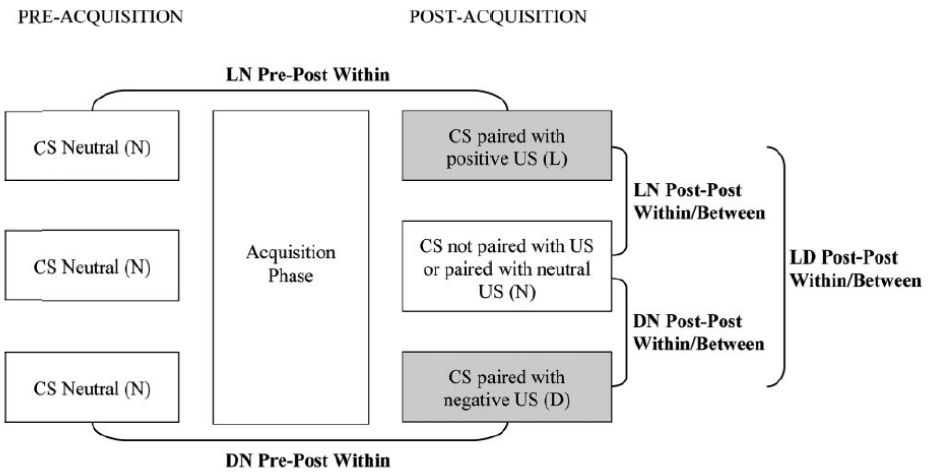
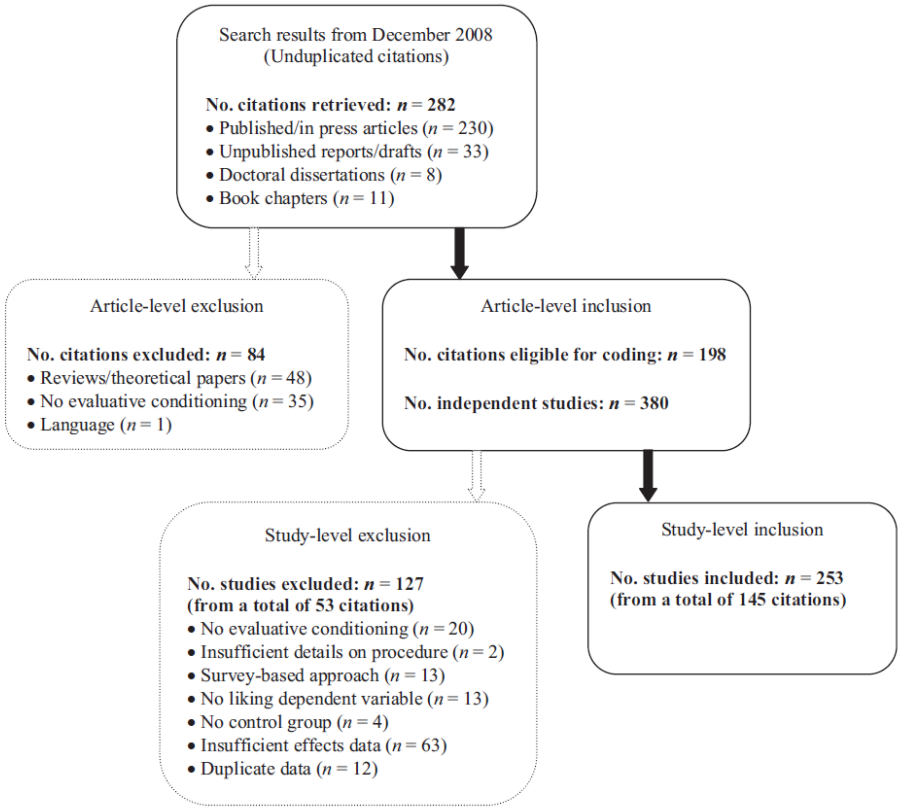


Figure 2. Illustration of possible effect size contrasts (boldface) in within- or between-subjects evaluative conditioning designs. CS = conditioned stimulus; US = unconditioned stimulus; L = "liking"; D = "disliking"; N = "neutral" (see text for details).

4 studies worden uitgesloten aangezien ze enkel L of D meten (geen controlegroep).



8 Berekenen van effectgroottes

- Naast r en d zijn ook niet-gestandaardiseerde effectgroottes mogelijk (dit laten we hier buiten beschouwing).
- Vaak wordt r aangeraden aangezien:
  * $-1 \leq r \leq 1$
  * onafhankelijk van wat onderzocht wordt, betekent $r = 0$ geen (linear) effect en een waarde van -1 of 1 een perfecte correlatie (lineair!).
- Bij vergelijken van groepen waarbij groepgroottes heel verschillend zijn, raadt men aan om d te gebruiken aangezien r mogelijks een vertekende schatter oplevert.
- Na de keuze van de effecgrootte: bereken de effectgrootte voor elk effect dat opgenomen moet worden in de meta-analyse.

- Het is mogelijk dat 1 studie verschillende effectgroottes rapporteert.

  Voorbeeld

  posttraumatische stress kan op verschillende manieren gemeten zijn in 1 studie

  Opties:

  - Laat verschillende effectgroottes voor verschillende metingen van dezelfde uitkomst uit 1 studie toe in de meta-analyse.
  - Bereken een gemiddelde effectgrootte over alle metingen van dezelfde uitkomst.

- Artikels rapporteren niet noodzakelijk effectgroottes of gebruiken andere effectgroottes dan deze in de meta-analyse.

  In dat geval moeten effectgroottes berekend worden of moeten de gebruikte effectgroottes getransformeerd worden.

- Het berekenen van effectgroottes kan aan de hand van toetsingsgrootheden (bvb. van t naar r), op basis van kanswaarden (van een kans overgaan naar een toetsingsgrootheid z) of d kan berekend worden op basis van gemiddeldes en standaarddeviaties.
- Indien er onvoldoende data voorhanden is om effectgroottes te bepalen, dient men de auteurs te contacteren.
- Verschillende manieren om effectgroottes te transformeren of te berekenen zijn beschikbaar in de literatuur; enkele daarvan hebben we behandeld in het hoofdstuk 'Statistische en praktische significantie'.

Voorbeeld 1

- Figuur 2: verschillende contrasten mogelijk (ook onderscheid tussen between-subject en within-subject design)

  Om na te gaan of EC effecten variëren in functie van het gebruikte contrast, wordt het contrast type gespecifieerd bij het coderen van de effectgrootte, alsook de aard van de controlegroep bij between-subjects designs.

- EC effecten vergelijken in essentie 2 gemiddeldes $\rightarrow$ gebruik van Cohen's d
- Berekening op basis van gemiddeldes, standaarddeviaties en steekproefgroottes of via transformatie van andere effectgroottes of toetsingsgrootheden

  Transformaties op basis van t-statistieken, F-statistieken met 1 vrijheidsgraad en enkel 2 groepen, gemiddelde winstscores, verschillen in proporties, Pearson correlatiecoëfficiënt r

- Afhankelijke variabelen / uitkomsten: zelfrapportering, keuze, impliciete metingen

9 Methodes voor meta-analyse

In dit stuk bespreken we het fixed en random effects model van Hedges en Olkin en het random effects model van Hunter en Schmidt waarbij een correlatiecoëfficiënt r gebruikt wordt om de grootte van een effect te kwantificeren. De beschreven methodes zijn ook toepasbaar wanneer effectgroottes d beschouwd worden.

9.1 Fixed effects model (Hedges en Olkin)

Correlaties worden eerst omgezet in een standaard normale metriek gebruik makend van de omzetting van ρ naar z van Fisher (Fisher's z-transformatie):

$$Z_i = \frac{1}{2}\ \log\left(\frac{1+r_i}{1-r_i}\right)$$

Deze score volgt bij benadering een normale verdeling met gemiddelde $\bar{Z}_\rho$ en variantie $1/(n_i-3)$. Het terugtransformeren naar r_i gebeurt dan als volgt

$$r_i = \frac{e^{2Z_i}-1}{e^{2Z_i}+1}.$$

In een volgende stap wordt een gewogen gemiddelde berekend van de getransformeerde effectgroottes:

$$\overline{Z_r} = \frac{\sum_{i=1}^{k} w_i Z_i}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k}(n_i-3)Z_i}{\sum_{i=1}^{k}(n_i-3)}. \tag{1}$$

Merk op dat $w_i = 1/\mathrm{Var}(Z_i) = n_i - 3$.

Een schatter voor ρ is dan:

$$\hat{\rho} = \frac{e^{2\overline{Z_r}}-1}{e^{2\overline{Z_r}}+1}.$$

De standaardfout van $\overline{Z_r}$ is

$$SE(\overline{Z_r}) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}.$$

We kunnen een betrouwbaarheidsinterval rond de gemiddelde effectgrootte construeren door te steunen op de normale verdeling, een 95% betrouwbaarheidsinterval is dan:

$$\left[\overline{Z_r} - 1.96 \times SE(\overline{Z_r}), \overline{Z_r} + 1.96 \times SE(\overline{Z_r})\right].$$

Dit interval kan als volgt getransformeerd worden in een 95% betrouwbaarheidsinterval voor ρ:

$$\left[\frac{e^{2(\overline{Z_r}-1.96\times SE(\overline{Z_r}))}-1}{e^{2(\overline{Z_r}-1.96\times SE(\overline{Z_r}))}+1}, \frac{e^{2(\overline{Z_r}+1.96\times SE(\overline{Z_r}))}-1}{e^{2(\overline{Z_r}+1.96\times SE(\overline{Z_r}))}+1}.\right].$$

Het gewogen gemiddelde in (1) kan gebruikt worden om de homogeniteit van de correlaties te toetsen. Hierbij wordt de volgende toetsingsgrootheid gebruikt:

$$Q = \sum_{i=1}^{k}(n_i - 3)(Z_i - \bar{Z}_r)^2. \tag{2}$$

Er kan aangetoond worden dat onder de nulhypothese van homogene correlaties geldt dat $Q \sim \chi^2(k-1)$. Indien de nulhypothese verworpen wordt, kan men overstappen op het random effects model (sectie 9.2).

Een andere optie is om verder te werken met fixed effects modellen maar op te gaan splitsen volgens 1 of meerdere *meta-analytische predictoren* (i.e. mogelijke moderatoren) totdat de nulhypothese binnen iedere categorie niet langer verworpen wordt.

Voorbeeld

Veronderstel dat in 8 scholen de correlatie tussen intelligentie en examenresultaten onderzocht werd en dat men deze resultaten wil combineren over scholen. Bij deze 8 scholen kunnen we 2 types onderwijs onderscheiden. De resultaten van de studie staan hieronder weergegeven:

School	Type onderwijs	n	r
1	I	100	0.34
2	I	100	0.16
3	I	50	0.12
4	I	50	0.38
5	II	100	0.19
6	II	100	0.01
7	II	50	-0.03
8	II	50	0.23

Via de fixed effects benadering van Hedges en Olkin vinden we:

School	n	w	r	z
1	100	97	0.34	0.35
2	100	97	0.16	0.16
3	50	47	0.12	0.12
4	50	47	0.38	0.40
5	100	97	0.19	0.19
6	100	97	0.01	0.01
7	50	47	-0.03	-0.03
8	50	47	0.23	0.23

$$
\begin{aligned}
w_i &= n_i - 3 \\
z_i &= \tfrac{1}{2}\log\left(\tfrac{1+r_i}{1-r_i}\right) \\
\overline{z} &= \frac{\sum_{i=1}^{k} w_i z_i}{\sum_{i=1}^{k} w_i} = 0.18 \\
\hat{\rho} &= \frac{e^{2\overline{z}}-1}{e^{2\overline{z}}+1} = 0.178 \\
\mathrm{se}(\overline{Z}) &= \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}} = 0.042 \\
Q &= \sum_{i=1}^{k} w_i (z_i - \overline{z})^2 \\
&= 10.444 \\
&< 14.07 = \chi^2_{(0.95)}(7)
\end{aligned}
$$

Een 95% betrouwbaarheidsinterval voor ρ is als volgt:

$$\left[\frac{e^{2(\overline{z}-1.96\times\mathrm{se}(\overline{z}))}-1}{e^{2(\overline{z}-1.96\times\mathrm{se}(\overline{z}))}+1}, \frac{e^{2(\overline{z}+1.96\times\mathrm{se}(\overline{z}))}-1}{e^{2(\overline{z}+1.96\times\mathrm{se}(\overline{z}))}+1}\right] = [0.098, 0.256]$$

We stellen vast dat de nulhypothese van homogene correlaties niet verworpen kan worden op het 5% significantieniveau aangezien de geobserveerde toetsingsgrootheid kleiner is dan de kristische waarde uit de χ^2-verdeling met 7 ($= k - 1$) vrijheidsgraden ($10.444 < 14.07$).

9.2 Random effects model (Hedges en Olkin)

Hierbij wordt opnieuw gebruik gemaakt van een gewogen gemiddelde zoals in (1) maar de gewichten bevatten nu een variantiecomponent die zowel de variantie binnen als tussen de studies in rekening brengt.

We noteren de variantie tussen de groepen als τ^2 (schatter: $\hat{\tau}^2$). Vergelijking (1) wordt dus vervangen door

$$\overline{Z_r^*} = \frac{\sum_{i=1}^{k} w_i^* Z_i}{\sum_{i=1}^{k} w_i^*} \tag{3}$$

waarbij

$$w_i^* = \left(\frac{1}{w_i} + \tau^2\right)^{-1}.$$

τ^2 kan op verschillende manieren geschat worden, bvb. (Hedges & Vevea, 1998):

$$\tau^2 = \frac{Q - k + 1}{c}$$

met Q uit vergelijking (2) en

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}.$$

Wanneer we met correlaties werken, is $w_i = n_i - 3$ en dus

$$c = \sum_{i=1}^{k} (n_i - 3) - \frac{\sum_{i=1}^{k} (n_i - 3)^2}{\sum_{i=1}^{k} (n_i - 3)}.$$

Het is mogelijk dat de schatting voor de variabiliteit tussen de groepen, $\hat{\tau}^2$, negatief is. In dat geval wordt deze gelijk aan 0 gesteld.

Een schatter voor $\bar{\rho}$ is dan:

$$\hat{\bar{\rho}} = \frac{e^{2\overline{Z_r^*}} - 1}{e^{2\overline{Z_r^*}} + 1}.$$

De standaardfout van $\overline{Z_r^*}$ is

$$SE(\overline{Z_r^*}) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}}.$$

Analoog als voorheen ziet een 95% betrouwbaarheidsinterval voor de gemiddelde effectgrootte er als volgt uit (onder de veronderstelling van normaliteit):

$$\left[\overline{Z_r^*} - 1.96 \times SE(\overline{Z_r^*}), \overline{Z_r^*} + 1.96 \times SE(\overline{Z_r^*})\right].$$

Dit interval kan als volgt getransformeerd worden in een 95% betrouwbaarheidsinterval voor $\bar{\rho}$:

$$\left[\frac{e^{2(\overline{Z_r^*} - 1.96 \times SE(\overline{Z_r^*}))} - 1}{e^{2(\overline{Z_r^*} - 1.96 \times SE(\overline{Z_r^*}))} + 1}, \frac{e^{2(\overline{Z_r^*} + 1.96 \times SE(\overline{Z_r^*}))} - 1}{e^{2(\overline{Z_r^*} + 1.96 \times SE(\overline{Z_r^*}))} + 1}.\right].$$

Voorbeeld

We hernemen het voorbeeld waarbij de samenhang tussen intelligentie en examenresultaten onderzocht werd. Via de random effects benadering van Hunter en Schmidt vinden we:

School	n	w	$w*$	r	z
1	100	97	57.9	0.34	0.35
2	100	97	57.9	0.16	0.16
3	50	47	35.4	0.12	0.12
4	50	47	35.4	0.38	0.40
5	100	97	57.9	0.19	0.19
6	100	97	57.9	0.01	0.01
7	50	47	35.4	-0.03	-0.03
8	50	47	35.4	0.23	0.23

$$w_i = n_i - 3$$
$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} = 495.3$$
$$Q = 10.444$$
$$\tau^2 = \frac{Q-k+1}{c} = 0.007$$
$$w_i^* = \left(\frac{1}{w_i} + \tau^2\right)^{-1}$$
$$z_i = \frac{1}{2}\log\left(\frac{1+r_i}{1-r_i}\right)$$
$$\overline{z}^* = \frac{\sum_{i=1}^{k} w_i^* z_i}{\sum_{i=1}^{k} w_i^*} = 0.18$$
$$\hat{\overline{\rho}} = \frac{e^{2\overline{z}^*}-1}{e^{2\overline{z}^*}+1} = 0.178$$
$$\text{se}(\overline{Z}^*) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}} = 0.052$$

Een 95% betrouwbaarheidsinterval voor $\bar{\rho}$ is als volgt:

$$\left[\frac{e^{2(\overline{z}^* - 1.96\times\text{se}(\overline{z}^*))}-1}{e^{2(\overline{z}^* - 1.96\times\text{se}(\overline{z}^*))}+1}, \frac{e^{2(\overline{z}^* + 1.96\times\text{se}(\overline{z}^*))}-1}{e^{2(\overline{z}^* + 1.96\times\text{se}(\overline{z}^*))}+1}\right] = [0.079, 0.274]$$

9.3 Random effects model (Hunter en Schmidt)

Hunter en Schmidt stellen enkel een methode voor random effecten voor. Hun argumentering is dat fixed effects modellen niet geschikt zijn voor empirische data en voor het type conclusies dat onderzoekers wensen te trekken.

Hun methode laat toe om te corrigeren voor verschillende bronnen van fouten of artefacten zoals steekproeffouten, de betrouwbaarheid van metingen en een beperkte variatie in scores. Hier bespreken we de 'bare bones' versie: hierbij wordt enkel gecorrigeerd voor steekproeffouten.

In tegenstelling tot de vorige methodes, worden de correlaties niet eerst getransformeerd om een gewogen gemiddelde te berekenen:

$$\hat{\bar{\rho}} = \bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i}.$$

We zien ook dat de studie-specifieke gewichten op een andere manier gekozen worden dan bij de voorgaande methodes.

Verder vertrekt men van de volgende decompositie van σ_ρ^2:

$$\sigma_r^2 = \sigma_\rho^2 + \sigma_e^2 \Rightarrow \sigma_\rho^2 = \sigma_r^2 - \sigma_e^2.$$

σ_r^2 en σ_e^2 kunnen als volgt geschat worden:

$$\begin{aligned}\hat{\sigma}_r^2 &= \frac{\sum_{i=1}^{k} n_i(r_i - \bar{r})^2}{\sum_{i=1}^{k} n_i} \\ \hat{\sigma}_e^2 &= \frac{(1-\bar{r}^2)^2}{\bar{n}-1} \text{ met } \bar{n} = \frac{\sum_{i=1}^{k} n_i}{k}\end{aligned}$$

en bijgevolg kan de variantie van de populatiecorrelaties geschat worden als $\hat{\sigma}_\rho^2 = \hat{\sigma}_r^2 - \hat{\sigma}_e^2$.

Een 95% betrouwbaarheidsinterval voor $\bar{\rho}$ is:

$$\left[\bar{r} - 1.96 \times \frac{\hat{\sigma}_r}{\sqrt{k}}, \bar{r} + 1.96 \times \frac{\hat{\sigma}_r}{\sqrt{k}}\right]$$

aangezien $SE(\bar{r}) = \hat{\sigma}_r/\sqrt{k}$. De breedte van het betrouwbaarheidsinterval weerspiegelt de grootte van de fout in de schatting van $\bar{\rho}$ te wijten aan de variabiliteit van $\bar{r}$.

Een 95% credibility-interval voor $\bar{\rho}$ is:

$$[\bar{r} - 1.96 \times \hat{\sigma}_\rho, \bar{r} + 1.96 \times \hat{\sigma}_\rho]$$

De breedte van het credibility-interval weerspiegelt de variabiliteit van ρ_i rond $\bar{\rho}$ in de populatie van studies.

Een credibility-interval heeft betrekking op de verdeling van de parameter terwijl een betrouwbaarheidsinterval betrekking heeft op de schatter van één effect $\bar{\rho}$. De steekproeffout is omvat in de standaardfout van $\bar{r}$ die ook afhangt van de steekproeffgrootte.
Credibility-intervals hangen niet af van de steekproeffout aangezien variantie veroorzaakt door steekproeffouten verwijderd werd uit de schatter voor σ_ρ^2.

Interpretatie 95% betrouwbaarheidsinterval: bij een oneindig aantal meta-analyses (intervallen) zal 95% van de bekomen intervallen $\bar{\rho}$ bevatten.

Interpretatie 95% credibility-interval: 95% van de waarden in de distributie van ρ_i liggen binnen dit interval.

De homogeniteit van de effectgroottes kan getoetst worden a.d.h.v. de volgende toetsingsgrootheid:

$$Q^* = \sum_{i=1}^{k} \frac{(n_i - 1)(r_i - \bar{r})^2}{(1-\bar{r}^2)^2}.$$

Onder de nulhypothese van homogene correlaties is Q^* χ^2-verdeeld met $k-1$ vrijheidsgraden.

Voorbeeld

We combineren de resultaten voor de samenhang tussen intelligentie en studieresultaten nu aan de hand van de random effects benadering an Hunter en Schmidt.

1. Bepaling $\bar{r}$

$$\begin{aligned}\bar{r} &= \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i} \\ &= \frac{100(0.34) + \ldots + 50(0.23)}{100 + \ldots + 50} = 0.175\end{aligned}$$

2. Bepaling $\hat{\sigma}_r^2$

$$\begin{aligned}\hat{\sigma}_r^2 &= \frac{\sum_{i=1}^{k}[n_i(r_i - \bar{r})^2]}{\sum_{i=1}^{k} n_i} \\ &= \frac{100(0.34 - 0.175)^2 + \ldots + 50(0.23 - 0.175)^2}{100 + \ldots + 50} = 0.0167\end{aligned}$$

3. Bepaling $\hat{\sigma}_e^2$

$$\begin{aligned}\hat{\sigma}_e^2 &= \frac{(1 - \bar{r}^2)^2}{\bar{n} - 1} \\ &= \frac{(1 - 0.175^2)^2}{74} = 0.0127\end{aligned}$$

4. $\hat{\sigma}_\rho^2 = \hat{\sigma}_r^2 - \hat{\sigma}_e^2$. Bijgevolg

$$\hat{\sigma}_\rho^2 = 0.0167 - 0.0127 = 0.0040.$$

5. 95% credibility-interval voor $\bar{\rho}$:

$$\begin{aligned}&[\bar{r} - 1.96 \times \hat{\sigma_\rho}, \bar{r} + 1.96 \times \hat{\sigma_\rho}] \\ &\left[0.175 - 1.96 \times \sqrt{0.004}, 0.175 + 1.96 \times \sqrt{0.004}\right] \\ &[0.052, 0.298]\end{aligned}$$

6. 95% betrouwbaarheidsinterval voor $\bar{\rho}$:

$$\begin{aligned}&\left[\bar{r} - 1.96 \times \frac{\hat{\sigma_r}}{\sqrt{k}}, \bar{r} + 1.96 \times \frac{\hat{\sigma_r}}{\sqrt{k}}\right] \\ &[0.175 - 1.96 \times 0.046, 0.175 + 1.96 \times 0.046] \\ &[0.086, 0.264]\end{aligned}$$

7. Toetsingsgrootheid homogeniteit:

$$\sum_{i=1}^{k} \frac{(n_i - 1)(r_i - \bar{r})^2}{(1 - \bar{r}^2)^2} = 10.482 < 14.07 \quad (\chi^2_{(0.95)}(7)).$$

We zien dat $\hat{\sigma}_e^2/\hat{\sigma}_r^2 = 0.76$. Dit betekent dat slechts 76% van de variatie in correlatie verklaard wordt door steekproeffouten. M.a.w. een beduidend stuk van de variantie in de geobserveerde correlaties wordt niet verklaard door steekproeffouten. Het type onderwijs is mogelijks een moderator voor de relatie tussen intelligentie en examenresultaten. Dit gaat men na door een aparte meta-analyse voor de 2 types van onderwijs uit te voeren.

1. **Onderwijs Type I**

$$\begin{aligned}
\bar{r} &= \frac{100(0.34) + \ldots + 50(0.38)}{300} = 0.25 \\
\hat{\sigma}_r^2 &= \frac{100(0.34 - 0.25)^2 + \ldots + 50(0.38 - 0.25)^2}{300} = 0.0110 \\
\hat{\sigma}_e^2 &= \frac{(1 - \bar{r}^2)^2}{(\bar{n}' - 1)} = \frac{(1 - 0.25^2)^2}{300/4 - 1} = 0.0119 \\
\hat{\sigma}_\rho^2 &= \hat{\sigma}_r^2 - \hat{\sigma}_e^2 = 0.011 - 0.0119 = -0.0009
\end{aligned}$$

   Merk op dat een variantie niet negatief kan zijn. Wat we hier beschouwen is echter een schatting, die schatting kan een overschatting of een onderschatting zijn van de echte variantie. Vandaar dat we soms schattingen voor varianties bekomen die negatief zijn. Het is dus zeker niet zo dat de variantie die te wijten is aan steekproeffouten groter kan zijn dan de totale variantie van de correlatie.

2. **Onderwijs Type II**

$$\begin{aligned}
\bar{r} &= \frac{100(0.19) + \ldots + 50(0.23)}{300} = 0.10 \\
\hat{\sigma}_r^2 &= \frac{100(0.19 - 0.10)^2 + \ldots + 50(0.23 - 0.10)^2}{300} = 0.0110 \\
\hat{\sigma}_e^2 &= \frac{(1 - \bar{r}^2)^2}{(\bar{n}' - 1)} = \frac{(1 - 0.10^2)^2}{300/4 - 1} = 0.0132 \\
\hat{\sigma}_\rho^2 &= \hat{\sigma}_r^2 - \hat{\sigma}_e^2 = 0.011 - 0.0132 = -0.002
\end{aligned}$$

We vinden dat $\bar{r}$ inderdaad verschillend is voor beide types wat er op wijst dat ρ anders is voor beide types. Dit betekent dat het type onderwijs inderdaad een moderator is. We zien dat binnen de types van onderwijs de variatie in correlatie (bijna) volledig verklaard wordt door steekproeffouten.

9.4 Voorbeeld 1

- Effectgroottes worden gecorrigeerd voor vertekening in het geval van kleine steekproefgroottes; de bijhorende standaardfouten worden geschat.
- Combineren van verschillende effectgroottes binnen 1 studie:
  - Voor iedere analyse wordt de relevante effectgrootte van een studie gekozen; bij meerdere effectgroottes wordt een gemiddelde bepaald (aggregatie).
  - Alle meta-analytische berekeningen worden op basis van de geaggregeerde effectgroottes uitgevoerd.

Meta-analytische berekeningen:

- Gebruik van mixed-effects model (speciaal type van random-effects model, Lipsey en Wilson, 2001):
  - Een random effects model veronderstelt niet dat variatie in effectgroottes enkel afkomstig is van steekproeffouten.
  - Assumptie van fixed effects in EC literatuur is niet realistisch.
  - Random effects modellen: conservatiever
  - Het random effects model convergeert naar het fixed effects model indien steekproeffouten wel enige bron van variatie zijn.
- De gewichten van de studies zijn de inverse van de varianties.
- Toets voor heterogeniteit van effectgroottes en schatting voor de graad van heterogeniteit.
- Er wordt nagegaan of heterogeniteit (deels) verklaard kan worden door de moderatie van de studiekarakterstieken (zie ook verder).

  Categorische moderatoren: variantie-analyse

  Continue moderatoren: gewogen regressie-analyse

Analyses vooraf:

- Nagaan of er potentieel sprake kan zijn van publication bias (zie verder)
- Selectie van set van studies voor de hoofdanalyse

  Vooraf wordt variantie-analyse aangewend op de set van 253 studies om na te gaan of design-gerelateerde aspecten een invloed hebben op de grootte van de effectgroottes. Meer bepaald wordt de invloed nagegaan van:

- type contrast: LD, LN, DN (zie figuur 2)
- designtype (within versus between)
- meting (pre-post versus post-post)
- type controlegroep
- onderzoeksonderwerp
- uitkomst

Bevindingen:

Enkel het type van uitkomst beïnvloedt de grootte van de effectgroottes.

→ De hoofdanalyse en moderator analyses worden uitgevoerd op de set van zelf-rapporteringsmaten.

Dit omvat zowel within- als between-subjects data en alle types van contrasten voor effectgroottes.

Hoofdanalyse:

- 215 effectgroottes afkomstig van 652 gecodeerde effectgroottes binnen studies

  1 studie wordt als outlier beschouwd omwille van extreem grote effectgrootte en wordt weggelaten.

  ⇒ 214 effectgroottes afkomstig van 9 149 deelnemers

- $\overline{d} = 0.524$, $SE(\overline{d}) = 0.03$ (*Cohen: medium effect*)
- 95% betrouwbaarheidsinterval voor $\overline{\delta}$: $[0.466, 0.582]$; effect significant verschillend van 0
- Q statistiek voor homogeniteit is 706.97
  De nulhypothese van homogeniteit wordt verworpen ($p < 0.001$)
- I^2: % variatie dat te wijten is aan heterogeniteit

$$I^2 = \frac{Q - df}{Q} = \frac{706.97 - 213}{706.97} = 0.70$$

  df: vrijheidsgraden geassocieerd met toets voor homogeniteit, i.e. $k - 1$

  ⇒ 70% van de variantie in effectgroottes over studies is te wijten aan systematische variatie i.p.v. steekproeffouten.

10 Moderator analyses

We hebben tot nu toe een onderscheid gemaakt tussen fixed en random effects modellen. Bij moderator analyses maakt men vaak gebruik van *mixed effects* modellen, i.e. modellen die zowel fixed als random effecten bevatten. In dergelijke modellen gaat men er, net als bij random effects modellen, van uit dat de effectgroottes variëren maar een deel van deze variatie wordt expliciet gemodelleerd in functie van potentiële moderator variabelen (die als 'fixed' beschouwd worden). Mixed effects modellen laten toe dat effectgroottes nog heterogeen zijn ('random' component) na het modelleren van de variatie tussen studies in functie van moderatoren, dit in tegenstelling tot hun fixed effects tegenhangers.

Aan de hand van mixed effects modellen kan men nagaan of er evidentie is voor moderatie.

- Voor categorische variabelen gebeurt dit via een variantie-analyse procedure.
- Voor continue moderatoren is dit gebaseerd op een gewogen regressieprocedure.

We laten technische details achterwege (zie ook Field en Gillett, 2010).

Merk op dat niet iedere studie informatie geeft over het niveau van een moderator variabele.

Voorbeeld

In een meta-analyse over experimenteel onderzoek naar bepaalde cognitieve vaardigheden, is het al dan niet geven van een beloning bij de experimenten een potentiële moderator. Dit wil zeggen dat deze de grootte van het manipulatie-effect mogelijks beïnvloedt. Het is echter mogelijk dat niet elke studie informatie verschaft over het al dan niet geven van een beloning.

In dergelijke gevallen hebben we te maken met ontbrekende data omdat niet alle studies in de moderator analyse opgenomen kunnen worden.

- Het is niet zo dat de meta-analyse hierdoor met zekerheid niet-valide wordt en bijgevolg heeft het wel zin om dergelijke analyses uit te voeren.
- Echter, wanneer sommige niveaus van de moderator over- of onder-gerepresenteerd zijn, dient men voorzichtig te zijn met de interpretatie en de mogelijkheid van bias (vertekening) moet duidelijk gemaakt worden aan de lezer (Field & Gillett, 2010).

Voorbeeld 1

Moderator analyses (hoofdbevindingen):

EC effecten zijn sterker

- voor hogere 'contingency awareness' (expliciete leercontext, bewust van CS-US koppeling) dan voor lagere 'contingency awareness' (evaluatie op spontane manier),
- voor supraliminale dan voor subliminale US presentatie,
- voor 'postacquisition' (na fase CS-US koppelingen) dan voor 'postextension' effecten (na fase met niet-gekoppelde CS),
- voor zelfrapporteringsmaten dan voor impliciete maten (dit werd in een aparte analyse nagegaan).

11 Publication bias

Probleem: studies kunnen op systematische wijze uitgesloten worden voor publicatie, typisch omwille van het ontbreken van *statistisch* significante resultaten. In dit geval treedt een vertekening of bias op in de resultaten. Men spreekt ook van *File-drawer probleem* of *availability bias*.

Het feit dat bepaalde studies niet gepubliceerd worden, kan liggen aan policy van de journals of de auteurs zelf, andere redenen zijn ook mogelijk.

Daarom is het bij een meta-analyse van belang om ook niet-gepubliceerde studies te identificeren.

Publication bias?

We bespreken hier kort enkele manieren om na te gaan of er sprake is van publication bias.

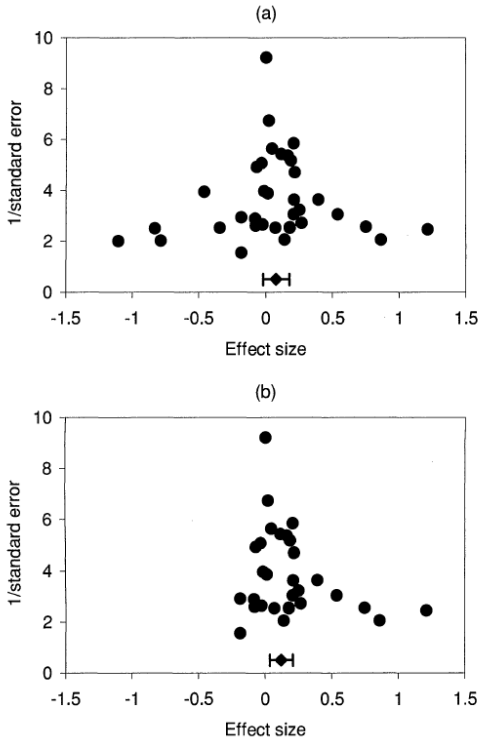
- Funnel-plot: plot van een maat voor de precisie van gemeten effecten (bvb. steekproefgrootte, 1/standaardfout) tegenover de geobserveerde effectgrootte

  Aanwijzingen voor publication bias worden gevonden door de afwezigheid van studies in het linkerdeel van de plot.

  Er bestaan ook methodes om de relatie tussen effectgrootte en zijn variantie te kwantificeren (bvb. correlatie).

  Publication bias wordt dan aangetoond via de aanwezigheid van een sterk verband.

  Duval en Tweedie (2000): plot in geval van geen publication bias (a) en wel publication bias (b).



- 'fail-safe' (N_{fs}) aantal van Rosenthal (1979): hoeveel niet-gepubliceerde of ontbrekende studies zonder effect of samenhang zijn nodig om de resultaten van een meta-analyse te veranderen van 'statistisch significant' naar 'statistisch niet-significant'?

  Om dit aantal te berekenen, moeten de effectgroottes eerst geconverteerd worden naar een z-score. De uitdrukking is dan als volgt:

  $$N_{fs} = \frac{\left(\sum_{i=1}^{k} z_i\right)^2}{2.706} - k$$

  Kritiek: focus op NHST

  Alternatief (Orwin, 1983): hoeveel ontbrekende studies zijn nodig om de effectgrootte onder een vooropgestelde grootte te brengen?

Correctie voor publication bias

Er zijn verschillende technieken mogelijk om in zekere mate te corrigeren voor publication bias:

- Aanvullen van een asymmetrische funnelplot

  Assumptie: alle ontbrekende studies zijn deze met kleine effectgroottes. Deze techniek leidt daardoor mogelijks tot een over-correctie.

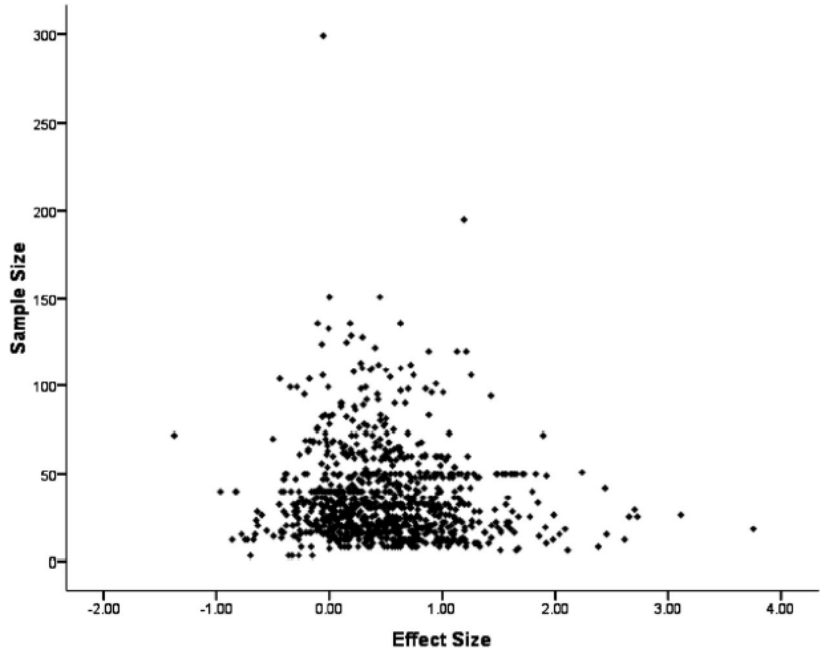
- Het bepalen van gewichten die het proces weergeven waarmee sommige studies meer kans hebben om gepubliceerd te worden dan anderen op basis van studiekarakteristieken, zoals statistische significantie.

  Methodes die gebruik maken van deze gewichten, laten toe om publication bias te detecteren en een aangepaste schatting voor de gemiddelde effectgrootte te bepalen die in zekere mate beter aangeeft wat de echte effectgrootte is (gesteld dat de modelassumpties gelden).

  Deze methodes zijn vrij technisch (we gaan hier niet dieper op in) en enkel effectief wanneer $k > 100$.

- Vevea en Woods (2005): aanpassing van voorgaande methodes die kan gebruikt worden wanneer het aantal studies k niet toelaat om gewichten accuraat te schatten.

voorbeeld 1



Geen duidelijke aanwijzingen voor publication bias

Statistische analyse van relatie tussen effectgrootte en steekproefgrootte: geen significante relatie

12 Rapportering

Het rapporteren van de resultaten van een meta-analyse verloopt vrij analoog als het rapporteren van onderzoeksresultaten, dit komt verder in de cursus aan bod.

Het rapport moet zo compleet en duidelijk mogelijk zijn. Iedere genomen beslissing die de analyse beïnvloed heeft, moet vermeld worden.

- Duidelijk de zoek- en inclusiecriteria beschrijven.
- Definieer duidelijk de gehanteerde maat voor effectgrootte.
- Motiveer keuze voor gebruikte techniek (random versus fixed effects aanpak).
- Moderatoren: effectgroottes voor subgroepen presenteren.
- Rapporteer statistieken die in verband staan met de variantie op de effectgroottes (schatting + toetsen).
- Rapporteer een schatting voor de effectgrootte op populatieniveau én bijhorend betrouwbaarheidsinterval of credibility-interval.
- Geef informatie over de aanwezigheid van publication bias en eventuele correcties hiervoor.

In de APA aanbevelingen omtrent 'MARS' (Meta-Analysis Reporting Standards; https://www.apa.org/pubs/journals/releases/amp-amp0000191.pdf) staat weergegeven hoe het rapporteren van een meta-analyse moet gebeuren en welke elementen noodzakelijk zijn.

13 Voorbeeld 2: Effectiveness of Long-term Psychodynamic Psychotherapy

In dit stuk worden kort de grote stappen van een meta-analyse herhaald aan de hand van de studie in de paper van Leichsenring en Rabung (2008): Effectiveness of Long-term Psychodynamic Psychotherapy (LTPP).

Men wenst de effecten van LTPP voor complexe mentale stoornissen (persoonlijkheidsstoornissen, angststoornissen, etc.) te beoordelen.

1. Het formuleren van een **onderzoeksvraag**: welke samenhang of welk effect wil ik onderzoeken?

   LTPP:

   - *Duidelijke definitie van wat als LTPP beschouwd wordt (duur therapie: minstens 1 jaar of 50 sessies).*
   - *Onderzoeksvragen:*
     - *Is LTPP superieur in vergelijking met andere kortere psychotherapeutische behandelingen?*
     - *Hoe effectief is LTPP?*
     - *Welke factoren hebben een significante bijdrage voor de uitkomst van LTPP?*

2. Welke studies zullen opgenomen worden, m.a.w. wat zijn de **inclusiecriteria**? Maken we enkel gebruik van gepubliceerde studies?

   LTPP: een lijst van criteria wordt opgesomd. Zowel observationele studies als randomised controlled trials worden opgenomen.

3. Het vinden van de studies en hun resultaten.

   LTPP: studies gepubliceerd tussen 1960 en mei 2008 via een automatische zoekopdracht in MEDLINE, PsycInfo en Current Contents. De zoektermen worden opgegeven. Daarnaast werden ook nog handmatig artikels en tekstboeken doorzocht en heeft men gecommuniceerd met auteurs en experten in dit vakgebied.

4. **Coderen** van de gegevens in de verzamelde studies. In de eerste stap worden schema's opgesteld voor het coderen. Er moet in principe ook altijd een tweede persoon zijn die dezelfde studies codeert. Op die manier kan de betrouwbaarheid / consistentie van het codeerschema onderzocht worden.

   De volgende zaken dienen gecodeerd en gerapporteerd te worden:

   - identificatie van de studie en referentie
   - de moderatoren die onderzocht worden
   - kwaliteit van de studies (kan gebruikt worden als moderator of als insluitingscriterium)
   - opzet van de studies (bvb. observationeel, experimenteel)
   - informatie over de berekening van effectgroottes

- effectgrootte en alle berekeningen om deze te bekomen.

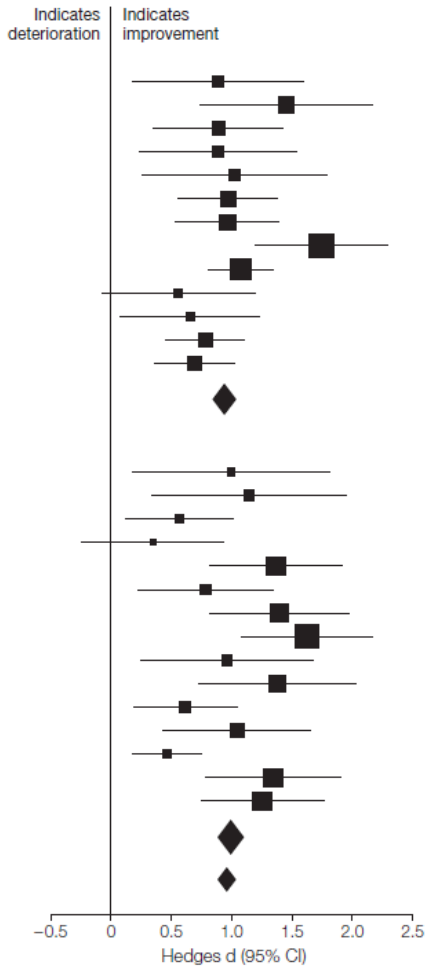
LTPP: er wordt een opsomming gegeven van de informatie die uit de verschillende artikels gehaald wordt. De 2 auteurs hebben deze informatie onafhankelijk van elkaar eruit gehaald. Betrouwbaarheid werd bepaald voor de bekomen effectgroottes door de 2 auteurs ($r \geq 0.80$).

5. De eigenlijke meta-analyse

   Zie de mogelijke werkwijzes beschreven in de voorgaande secties.

   LTPP: Voor 'within-group' effecten wordt d voor iedere uitkomst berekend als het gemiddelde voor de behandeling min het gemiddelde na de behandeling, gedeeld door de standaarddeviatie voor de behandeling. In het geval van meer dan 1 behandelingsgroep gebruikt men een gecorrigeerde d (Hedges), een positieve d betekent dat LTPP voor een verbetering zorgde .

   Resultaten 'overall outcome' (bovenaan: RCT's, onderaan: observationele studies):



Algemene conclusie: op basis van de meta-analyse vindt men bewijzen die de hypothese dat LTPP een effectieve behandeling is bij complexe mentale stoornissen, ondersteunen.

6. File-drawer probleem

   LTPP: Spearman rangcorrelatie wordt berekend tussen effectgrootte en steekproefgroottes. Een significante correlatie is een indicatie voor mogelijke publication bias. Het fail-safe aantal van Rosenthal wordt bepaald voor de effectgroottes van verschillende uitkomsten. Er worden geen belangrijke aanwijzingen voor publication bias gevonden.

14 Nabeschouwingen

- Een nuttige referentie bij het praktisch uitvoeren van een meta-analyse is de volgende:

  Field, A. P., & Gillet, R. (2010). How to do a meta-analysis (expert tutorial). *British Journal of Mathematical and Statistical Psychology, 63*, 665-694.

- De resultaten van een meta-analyse hangen niet enkel af van de studies die uiteindelijk opgenomen zijn maar ook van codering van de karakteristieken, welke statistische modellen gebruikt worden en de manier waarop studies gescreend worden. Om die reden zijn *sensitiviteitsanalyses* bijzonder nuttig. Hierbij worden de data geanalyseerd onder verschillende assumpties of met verschillende methodes. Indien de resultaten niet wijzigen onder de verschillende assumpties/methodes dan kan men stellen dat de resultaten robuust zijn m.b.t. deze assumpties/methodes.

  In case studie 1 werden sensitiviteitsanalyses uitgevoerd waarbij modellen beschouwd werden (1) zonder weglaten van de outlier, (2) met enkel within-subjects data, (3) enkel LN en DN contrasten en (5) met alle uitkomstmaten.

  De resultaten zijn in grote mate hetzelfde als voor de originele analyse.

  In case studie 2 werden ook sensitiviteitsanalyses uitgevoerd waarbij het effect van verschillende variabelen op de uitkomst werd nagegaan.

- In de praktijk is er software beschikbaar om meta-analyes uit te voeren.
  - Field en Gillet (2010) geven een overzicht en referenties van enkele mogelijke packages.
  - R bevat twee packages om meta-analyses uit te voeren, `meta` en `metafor`. In R kan ook een analyse voor publication bias uitgevoerd worden.
  - Er bestaat een plug-in voor Excel (Kontopantelis en Reeves, 2009).
  - Field en Gillet (2010) illustreren in hun tutorial het gebruik van SPSS bij het uitvoeren van een meta-analyse. SPSS heeft hiervoor geen ingebouwde tools; er wordt gebruik gemaakt van syntax die ze hiervoor geschreven hebben.
  - In case studie 1 maakte men gebruik van SPSS macros (Lipsey en Wilson, 2001). In case studie 2 gebruikte men SPSS en MetaWin (Rosenberg, 1999).

15 Referenties

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Becker, B.J. (1996). The generalizability of empirical research results. In Benbow, C.P. & Lubinski, D. (Eds.) *Intellectual talent: Psychological and social issues* (pp. 363-383). Baltimore, M.D.: Johns Hopkins University Press.

Duval, S., Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98.

Field, A.P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10*, 444-467.

Field, A. P., & Gillet, R. (2010). How to do a meta-analysis (expert tutorial). *British Journal of Mathematical and Statistical Psychology, 63*, 665-694.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L.V. (1992). Meta Analysis. *Journal of Educational Statistics, 17*, 279-296.

Hedges, L.V., & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Hedges, L.V., & Pigott, T.D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*, 203217.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative Conditioning in Humans: A Meta-Analysis. *Psychological Bulletin, 136,* 390-421.

Howitt, D., & Cramer, D. (2007). *Methoden en technieken in de psychologie.* Amsterdam: Pearson Education Benelux.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J.E., & Schmidt, F.L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. *International Journal of Selection and Assessment, 8*, 275-292.

Journal Article Reporting Standards (JARS), Meta-Analysis Reporting Standards (MARS), and Flow of Participants Through Each Stage of an Experiment or Quasi-Experiment. (2017, January 11). Retrieved from http://www.apastyle.org/manual/related/JARS-MARS.pdf

Kontopantelis, E., & Reeves, D. (2009). MetaEasy: A meta-analysis add-in for Microsoft Excel. *Journal of Statistical Software, 30*, 1-25.

Leichsenring, F., & Rabung, S. (2008). Effectiveness of Long-term Psychodynamic Psychotherapy. *Journal of the American Medical Association, 300*, 1551-1565.

Lipsey, M. W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Orwin, R.G. (1983). A fail-save N for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157-159.

Rosenberg MS, Adams DC, & Gurevitch J. (1999) *MetaWin. Statistical software for meta-analysis. Version 2.0*. Sunderland, MA: Sinauer Associates.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Vevea, J.L., & Woods, C.M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*, 428-443.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*, 26-30.