

Hoofdstuk 10: Regressieanalyse als Lingua Franca


Doelstellingen:

Na dit hoofdstuk

- *weet je de meerwaarde van het centreren en standaardiseren van variabelen voor het uitvoeren van een regressieanalyse;*
- *kan je variabelen in R centreren en standaardiseren;*
- *weet je de meerwaarde van het uitvoeren van dummyregressie;*
- *kan je dummyvariabelen aanmaken in R;*
- *kan je een dummyregressie (al dan niet met inbegrip van interactie-effecten) uitvoeren, rapporteren en interpreteren in R.*

Nodige files:

- Wis2.RData;
een file met daarin voor 80 leerlingen of zij al dan niet geslaagd zijn voor een wiskundetoets, hun IQ, geslacht en aantal uren dat zij studeerden
- Studenten2.RData
een databestand met gegevens over de leerstrategieën en -concepties van studenten en enkele persoonlijkheidsschalen.
- Pisa3.RData
een file met daarin voor alle Vlaamse leerlingen uit de PISA-bevraging 2009 hun geslacht, immigratiestatus en sociaal-economische status, hun score op een wiskunde-, taal- en wetenschappentest, de attitude van leerlingen ten aanzien van de school en computers en enkele scores op leerschalen (samenvatten, begrijpen, memoriseren).
- Werk.RData
een file met daarin gegevens voor welbevinden op het werk, de mate waarin respondenten het bedrijfsklimaat als open omschrijven en hun statuut (arbeiders, bedienden en kaderleden).
- OLP2 Functies.R
een file met daarin aangepaste functies die bij dit OLP horen

-  In de voorgaande hoofdstukken hebben we verschillende analysemodellen geïntroduceerd die in feite allemaal varianten zijn van eenzelfde techniek. Zowel de t-test, een ANOVA en regressieanalyse zijn eigenlijk vormen van wat een lineair model heet. In dit hoofdstuk introduceren we verschillende manieren om het regressiemodel verder uit te bouwen om enerzijds de mogelijkheden tot interpretatie te verhogen (dit doen we in 10.1 en 10.2) en anderzijds om binnen de regressievergelijking ook gebruik te maken van kwalitatieve onafhankelijke variabelen en interactietermen (dit doen we in 10.3). Uit deze laatste vorm van uitbouwen zal blijken dat we eigenlijk de regressievergelijking kunnen hanteren als Lingua Franca voor alle lineaire modellen. Daardoor kunnen we “flexibeler” verschillende statistische modellen bouwen en de daarbij horende hypothesen toetsen. Bovendien is dit tevens het opstapje naar meer geavanceerde analysetechnieken waarbij dezelfde regressievergelijking wordt gehanteerd als vertrekpunt om complexere onderzoeksvragen te beantwoorden.

10.1. Gecentreerde variabelen

10.1.1



Het databestand Wis2.RData bevat de variabelen Wiskunde (een score op een wiskundetoets), Iq en Urengestudeerd.

Voer een multivariate regressieanalyse uit om het effect van Urengestudeerd op Wiskunde na te gaan, waarbij gecontroleerd wordt voor Iq en beantwoord onderstaande vraag:

Wat is de precieze betekenis van het intercept? Is het intercept informatief?

10.1.2



Het bovenstaande voorbeeld toont aan dat het intercept bij regressieanalyse niet altijd even informatief is. Indien je rapporteert over onderzoeksresultaten kan het interessant zijn om ook het intercept interpreteerbaar te maken.

10.1.3



Nu, we weten dat het intercept die waarde is op onze afhankelijke variabele voor een respondent die nul scoort op de onafhankelijke variabele(n). Bijgevolg wordt het intercept pas informatief indien de waarde nul op onze onafhankelijke variabele(n) ook een betekenis krijgt. Of anders gesteld, indien we onze onafhankelijke variabele(n) zodanig hercoderen dat de waarde nul een realistische en interessante waarde wordt dan wordt ook het intercept informatief.

Welke waarde(n) van bijvoorbeeld de variabele Iq kan (kunnen) interessant zijn **als referentiewaarde**?

10.1.4



Wat je vaak tegenkomt in onderzoek is ervoor zorgen dat de waarde nul staat voor gemiddeld scoren op de bewuste variabele. Dit heet **centreren rond het gemiddelde**. Een respondent die gemiddeld scoort op de onafhankelijke variabele(n) krijgt de waarde nul op die onafhankelijke variabele(n).

10.1.5



Als we weten dat de gemiddelde score voor Iq 100 punten bedraagt, hoe gaan we dan best verder te werk? Hoe zorgen we ervoor dat de waarde 0 gelijk staat aan het gemiddeld Iq zonder dat de rest van de informatie (verschillen in Iq tussen leerlingen) verloren gaat?

10.1.6



In R kan je vrij eenvoudig een gecentreerde variabele aanmaken door gebruik te maken van de rekenmogelijkheden van R. Zo wordt in de onderstaande R-code een variabele X gecentreerd door, via de functie `mean()`, voor elke waarde het gemiddelde af te trekken:

```
> X_gecentreerd <- X - mean(X, na.rm=TRUE)
```

10.1.7



Herneem het bestand Wis2.RData en centreer de variabelen Iq en Urengestudeerd rond de gemiddelde waarde en schrijf het resultaat weg in twee nieuwe variabelen: Iq_c en Urengestudeerd_c. Doe vervolgens de multivariate regressieanalyse uit 10.1.1 opnieuw met de gecentreerde variabelen.

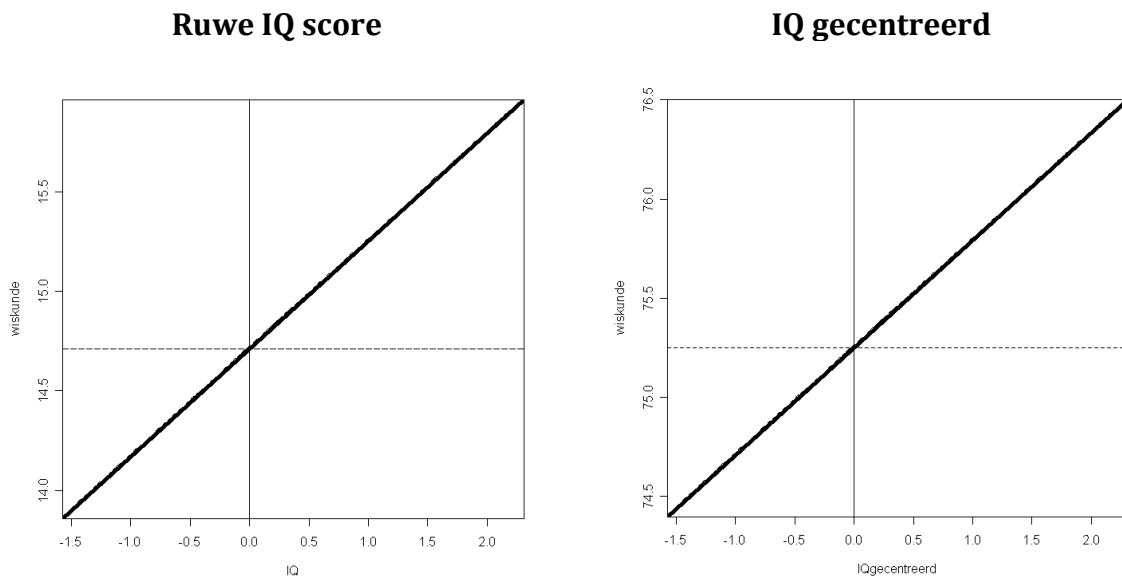
- Vergelijk de output met de output uit 10.1.1. Wat valt op?
- Interpreteer zowel het intercept als de regressiecoëfficiënten.

10.1.8



Het centreren van de onafhankelijke variabelen heeft geen invloed gehad op de regressiecoëfficiënten

Enkel het intercept is aangepast. Dit wordt duidelijk als we de rechte voor beide analyses naast elkaar zetten (zie onderstaande figuur). Het snijpunt van de regressieanalyse met de Y-as is veranderd. De hellingsgraad is in beide figuren identiek. Door te centreren rond het gemiddelde is in feite de regressielijn enkel naar boven geschoven.



Figuur 10.1: Regressierechte voor het effect van zowel de originele Iq variabele als Iq gecentreerd (Iq_c) op Wiskunde

10.1.9



Open het databestand Studenten2.RData. In dit databestand vind je de volgende variabelen terug:

- Diepteverw = de mate waarin een student de leerstof op een diepgaande wijze verwerkt;
- Opnamek = de mate waarin een student van mening is dat leren gelijk staat aan het enkel opslagen van kennis;
- Zelfontd = de mate waarin de student vindt dat leren voornamelijk een proces is dat de lerende al zelfontdekkend moet doorlopen.

Test de volgende hypothese aan de hand van een regressieanalyse waarbij je gebruik maakt van gecentreerde onafhankelijke variabelen en interpreteer voornamelijk het intercept:

De opvattingen van een student aangaande leren hebben een invloed op hoe de student leert. Indien een student van mening is dat leren voornamelijk gelijk staat aan het enkel opslagen van nieuwe kennis dan zal deze student minder geneigd zijn de leerstof op een diepe wijze te verwerken. Een student die vindt dat leren voornamelijk door middel van zelfontdekking dient plaats te vinden, zal dan weer eerder geneigd zijn hoger te scoren op een diepe verwerkingsstijl.

10.2 Gestandaardiseerde variabelen

10.2.1



Naast het intercept zijn de regressiecoëfficiënten ook niet altijd even rechttoe rechtaan te interpreteren. Neem bijvoorbeeld de output uit 10.1.9 (zie hieronder).

```
> summary(Model3)

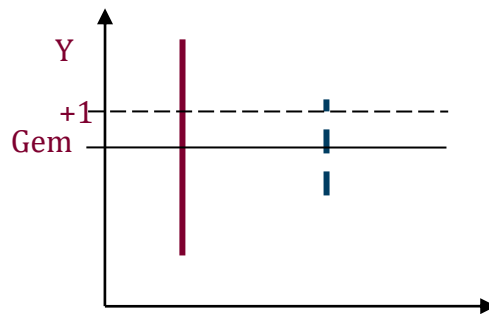
Call:
lm(formula = Diepteverw ~ Opnamek_c + Zelfontd_c, data = Studenten2)

Residuals:
    Min       1Q   Median       3Q      Max
-1.620309 -0.235525  0.005622  0.226418  1.895423

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.16470     0.03179   99.563 < 2e-16 ***
Opnamek_c     0.14958     0.04731    3.162  0.00185 **
Zelfontd_c    0.66491     0.05039   13.194 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

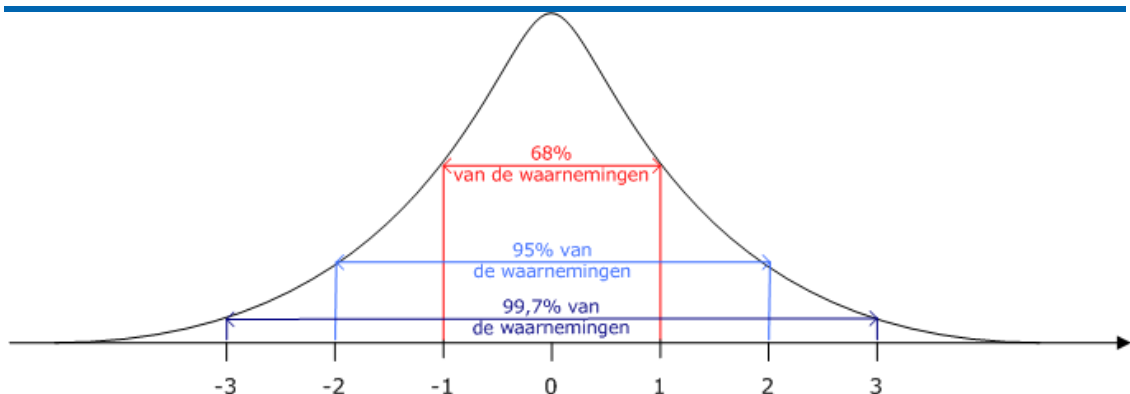
Er wordt daarbij bijvoorbeeld het effect nagegaan van de variabele 'Opname van kennis' (Opnamek_c). Eén punt hoger scoren op deze variabele leidt volgens deze analyse tot het scoren van 0,15 punten hoger op diepteverwerking.

Maar wat wil dit nu zeggen één punt hoger scoren op 'Opname van kennis'? Is dit zeer hoog scoren op deze variabele of is dit maar matig hoger scoren? Het antwoord op deze vragen is afhankelijk van de grootte van de verschillen tussen leerlingen. Anders gesteld, als de variantie in 'Opname van kennis' zeer klein is (bijna alle leerlingen verschillen nauwelijks van elkaar en scoren bijgevolg dicht rond het gemiddelde) dan zou één punt hoger scoren dan gemiddeld betekenen dat je veel hoger scoort dan de gemiddelde leerling. Is de variantie daarentegen zeer groot (leerlingscores liggen ruim verspreid rond het gemiddelde) dan is één punt hoger scoren minder drastisch. Het onderstaande figuurtje tracht dit te illustreren. De volle verticale lijn geeft de spreiding rond het gemiddelde weer voor een variabele waarvoor de variantie groter is. De gestipte verticale lijn is een variabele met een kleinere variantie. Uit de figuur kan je duidelijk aflezen dat bij de gestipte lijn 1 punt hoger scoren dan gemiddeld automatisch betekent dat je duidelijk een van de hoogst scorende respondenten bent. Scoor je bij de volle lijn één punt hoger, dan is dit minder extreem: je behoort nog niet bij de primussen van de klas.



Figuur 10.2 Illustratie voor de relatieve impact van “één punt hoger scoren” voor een variabele

Om dit interpretatieprobleem te omzeilen en zo de effectgrootte beter te kunnen afleiden uit de regressiecoëfficiënten kunnen we beroep doen op z-scores. Immers, voor alle variabelen geldt dat als we die omrekenen tot z-scores dat de schaal uniform is. Z-scores zijn immers variabelen die per respondent aangeven hoeveel standaardafwijkingen hij of zij verwijderd is van de gemiddelde score voor een gegeven kenmerk. Bijvoorbeeld 1 scoren op de variabele Opname van kennis in gestandaardiseerde vorm (= herrekenend naar z-score) betekent dat je 1 standaardafwijking hoger scoort dan de gemiddelde respondent. Indien de betrokken variabele normaal verdeeld is kunnen we beroep doen op de 68-95-99,7-regel om nog meer betekenis te geven aan de scores. De onderstaande figuur visualiseert die regel:



Figuur 10.3 De 68-95-99,7-regel visueel voorgesteld

68% van de waarnemingen behaalt een z-score tussen -1 en +1, of 16% scoort hoger dan 1 en 16% scoort lager dan -1.

95% van de waarnemingen behaalt een z-score tussen -2 en +2, of 2,5% scoort hoger dan 2 en 2,5% scoort lager dan -2.

99,7% van de waarnemingen behaalt een z-score tussen -3 en +3, of 0,15% scoort hoger dan 3 en 0,15% scoort lager dan -3.

10.2.2



We hernemen het voorbeeld van de wiskundescores. In de onderstaande output vind je de parameterschattingen terug van een regressieanalyse met als verklarende variabelen z-scores voor Iq en Urengestudeerd:

```
> summary(Model4)

Call:
lm(formula = Wiskunde ~ Iq_z + Urengestudeerd_z, data = Wis2)

Residuals:
    Min       1Q   Median       3Q      Max
-25.4678  -2.2676   0.1467   3.1231  16.7707

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    75.2500     0.6254 120.317 < 2e-16 ***
Iq_z            8.0856     0.6444  12.547 < 2e-16 ***
Urengestudeerd_z 2.1059     0.6444   3.268 0.00162 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.594 on 77 degrees of freedom
Multiple R-squared:  0.7166,    Adjusted R-squared:  0.7093
F-statistic: 97.36 on 2 and 77 DF,  p-value: < 2.2e-16
```

- Wat is de precieze betekenis van het intercept (75,25)?
- Hoe kan je nu de regressiecoëfficiënten interpreteren?

10.2.3



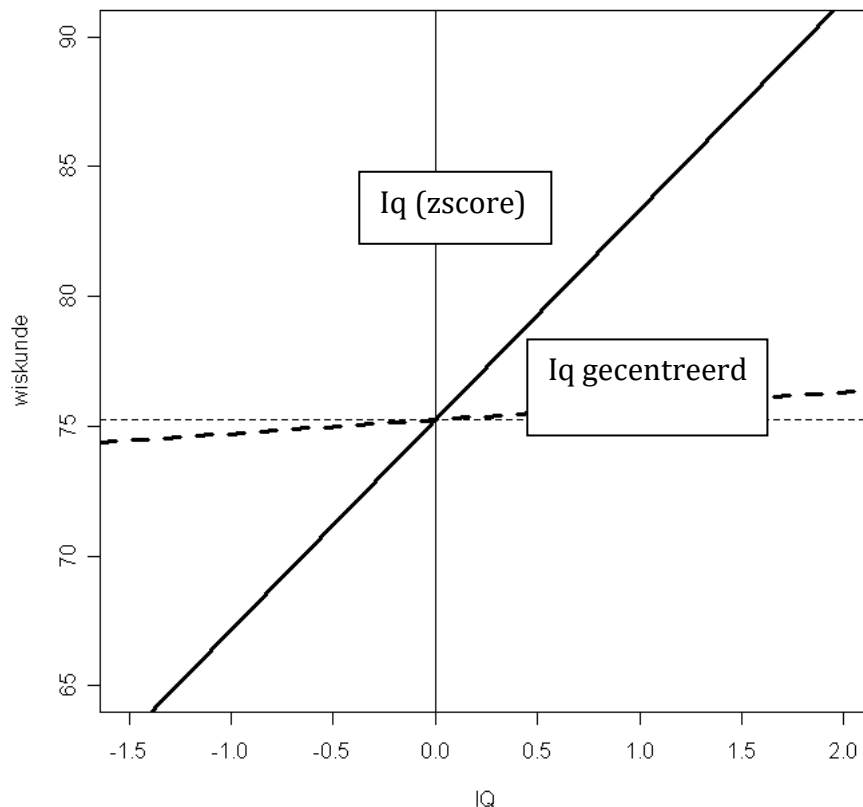
Een makkelijke manier om een z-score te maken van een variabele in R is door gebruik te maken van de functie `scale()`. In het onderstaand voorbeeld maken we een nieuwe variabele `X_z` aan die de z-scores voor een variabele `X` bevat.

```
> X_z<-scale(X)
```

10.2.4



In tegenstelling tot het centreren van de onafhankelijke variabelen heeft standaardiseren wel invloed op de regressiecoëfficiënten. Dat kan je afleiden uit de onderstaande figuur. Het snijpunt van de regressieanalyse met de Y-as is identiek aan de analyse met behulp van de gecentreerde Iq-variabele. De hellingsgraad is bij de variant met gestandaardiseerde scores (volle lijn) beduidend groter dan bij de variant met gecentreerde scores (stippelijn).



Figuur 10.4 Illustratie van de invloed van het omrekenen van een variabele naar een z-score

10.2.5



Open het databestand Studenten2.RData.

Test opnieuw de onderstaande hypothese aan de hand van een regressieanalyse waarbij je gebruik maakt van gestandaardiseerde onafhankelijke variabelen en interpreteer de parameters:

De opvattingen van een student aangaande leren hebben een invloed op hoe de student leert. Indien een student van mening is dat leren voornamelijk gelijk staat aan het enkel opslagen van nieuwe kennis dan zal deze student minder geneigd zijn de leerstof op een diepe wijze te verwerken. Een student die vindt dat leren voornamelijk door middel van zelfontdekking dient plaats te vinden, zal dan weer eerder geneigd zijn hoger te scoren op een diepe verwerkingsstijl.

10.2.6



Een laatste probleem dat zich kan stellen is dat de schaal van de afhankelijke variabele op zich ook weinigzeggend is. Analoog aan de vraagstelling in 10.2.1 kan je je afvragen wat 0,165 of 0,691 punten hoger scoren op diepteverwerking precies inhoudt. Ook voor de afhankelijke variabele kunnen we dezelfde stap zetten om de interpretatie te verhogen: standaardiseren.

10.2.7



Herneem de laatste analyse (10.2.5). Test opnieuw dezelfde hypothese, maar standaardiseer nu naast de onafhankelijke variabele eveneens de afhankelijke variabele. Hoe kan je nu het intercept en de regressiecoëfficiënten interpreteren?

10.2.8



De vraag die je kan opwerpen is, wat is de meerwaarde van standaardiseren? Het antwoord op deze vraag is sterk afhankelijk van de onderzoekssituatie waarin je belandt. Zit je bijvoorbeeld met variabelen die allen op zich een zeer zinnige schaal hebben (bijvoorbeeld hoeveel procent iemand haalt op het einde van een schooljaar, het aantal uren dat iemand opleiding gaat volgen in een bedrijf, ...) dan dien je niet te standaardiseren om de interpretatie te vergemakkelijken. Zit je daarentegen met variabelen met op zich niets zeggende schalen (zoals dat bijvoorbeeld vaak het geval is met schaalscores gaande van 1 tot 5) dan kan standaardiseren meer inzicht in de data leveren.

Een andere reden om te standaardiseren kan liggen in de vergelijkbaarheid van gegevens. Stel, jij hebt als onderzoeker je verdiept in onderzoek naar wat precies de determinanten zijn van een diepgaande leerstijl. Echter, in je eigen onderzoek maak je gebruik van een licht aangepaste vragenlijst met een andere wijze van berekenen van scores dan andere onderzoekers. Hoe ga je dan je resultaten kunnen aftoetsen ten aanzien van wat andere onderzoekers uitkomen? Je kan hooguit iets zeggen over de richting en significantie van de gevonden effecten, maar je kan geen uitspraken doen over het feit of het effect in jouw onderzoek sterker is dan wat in de literatuur wordt gerapporteerd. Werken zowel jij als je collega-onderzoekers met gestandaardiseerde variabelen dan worden de resultaten een stuk gemakkelijker te vergelijken. Het is daarom ook vaak een conventie geworden om te werk te gaan met gestandaardiseerde scores.

Ten slotte kunnen we ook stellen dat we deze werkwijze hebben aangebracht omdat dit in de verdere uitbreidingen van de regressieanalyse zal leiden tot handig interpreteerbare parameters.

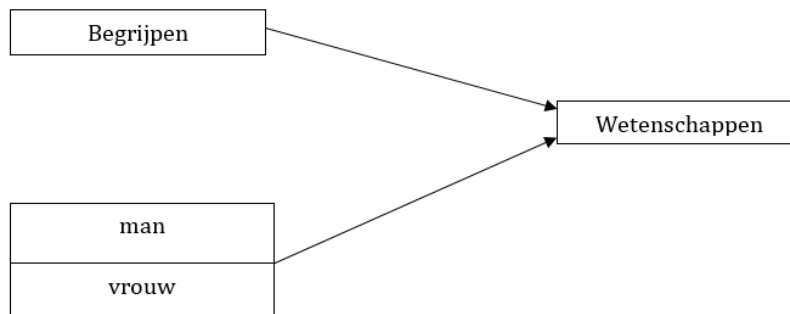
10.3 Dummyregressie

10.3.1



Toen we correlatie en bivariate regressieanalyse bespraken, hebben we de puntenwolk geïntroduceerd om het verband tussen variabelen te visualiseren.

Stel, onze onderzoeksvraag is: beïnvloedt de mate waarin studenten de leerstof proberen te begrijpen (schaalscore) hun score op een wetenschappentoets, ongeacht het geslacht van de student?

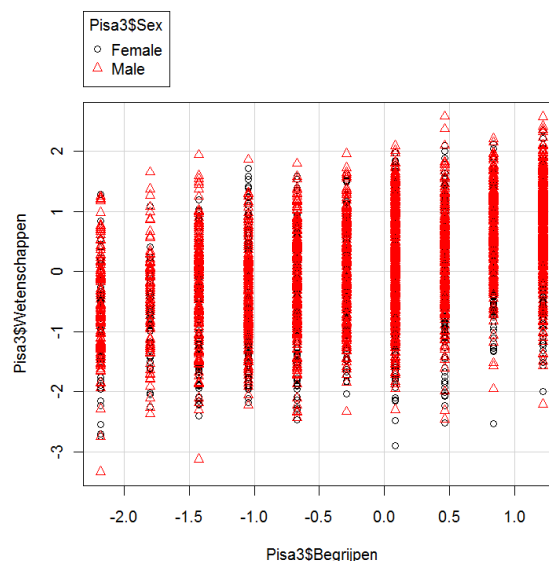


De onderstaande puntenwolk zet de scores van respondenten voor de schaa score 'begrijpen' uit tegen de scores op een wetenschappentoets die ze behalen. Om een eerste idee te krijgen van het antwoord op bovengestelde onderzoeksvraag, zouden we graag geslacht toevoegen aan deze figuur. Dit kan eenvoudig via volgend commando uit de library `car`:

```
> library(car)
> scatterplot(y~xkwantitatief/xkwalitatief, smooth=FALSE,
              regLine=FALSE)
```

Voor de onderzoeksvraag hierboven, geeft dit op de `Pisa3.RData`:

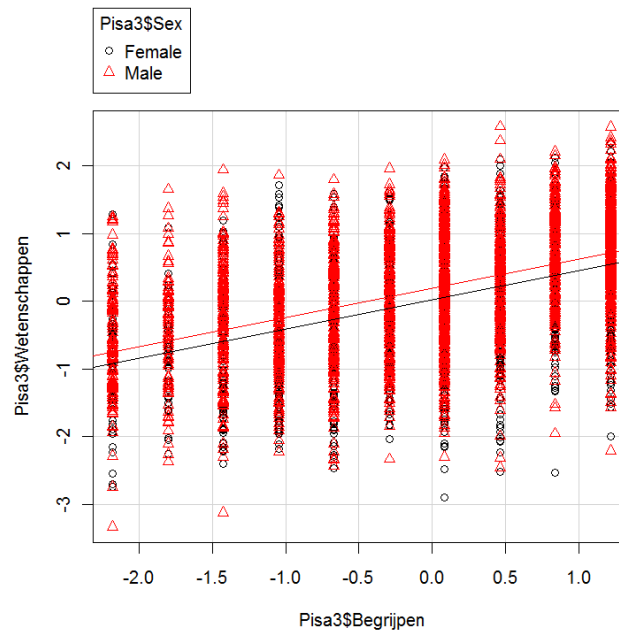
```
> scatterplot(Pisa3$Wetenschappenz~Pisa3$Begrijpenz/Pisa3$Sex,
              smooth=FALSE, regLine=FALSE)
```



Figuur 10.5 Scatterplot per geslacht

Algemeen zien we een positief verband: de leerstof meer proberen te begrijpen, lijkt samen te gaan met een betere score op Wetenschappen. Deze grafiek geeft nog meer informatie: elke rood driehoekje stelt de scores op begrijpen en wetenschappen voor een jongen voor en elk zwarte bolletje de scores voor een meisje.

In de eenvoudigste manier bestaat dummyregressie er eigenlijk uit om voor elke subgroep een afzonderlijke regressielijn te schatten die evenwijdig met elkaar lopen. In de onderstaande figuur zijn beide regressielijnen opgenomen.



Figuur 10.6 Scatterplot en regressielijn per geslacht

10.3.2



Beantwoord volgende vragen op basis van figuur 10.6.

- In welk opzicht verschillen jongens en meisjes van elkaar?
- Is het effect van begrijpen op wetenschapsscore onafhankelijk van geslacht?
- Is het verband van geslacht op wetenschapsscore onafhankelijk van de mate van begrijpen?

10.3.3



Natuurlijk zijn we opnieuw geïnteresseerd in hoe sterk de effecten van de onafhankelijke variabelen op de afhankelijke variabele zijn. Ook willen we graag weten of we de gevonden verbanden kunnen doortrekken naar de populatie.

Een regressieanalyse lijkt op het eerste zicht onmogelijk want regressieanalyse is enkel bestemd voor kwantitatieve variabelen. Maar net regressieanalyse is één van de meest gebruikte en meest flexibele analysetechnieken.

Een oplossing voor het probleem ligt in het werken met dummyvariabelen. Een **dummyvariabele** is een specifiek type van categorische variabele, meerbepaald een variabele die slechts twee categorieën bevat die respectievelijk de score 0 en 1 krijgen toegekend.

Nu, elke categorische variabele kan omgezet worden in een (reeks van) dummyvariabelen. Stel dat je een variabele Geslacht hebt. Deze variabele kan je simpelweg hercoderen tot de dummyvariabele **D** (zie hieronder):

	D
vrouw	0
man	1

We zeggen dan dat de dummyvariabele **aanstaat** voor een man. Of, met andere woorden, de dummyvariabele geeft weer of het al dan niet om een man gaat (1=waar).

Elke categorische variabele kan worden omgezet in een reeks van Dummyvariabelen die samen aangeven tot welke categorie een respondent behoort. Bijvoorbeeld een categorische variabele over het aantal uren gestudeerd, die bestaat uit 3 categorieën, nl. laag, gemiddeld, hoog:

	D1	D2
laag	0	0
gemiddeld	1	0
hoog	0	1

Op basis van deze dummycodering kunnen we eveneens afleiden tot welke categorie een respondent behoort. Zo behoort iemand die in het bovenstaande geval nul scoort op zowel D1 als D2 tot de categorie 'laag'. De categorie laag is in feite een referentiecategorie waartegen we de twee dummyvariabelen afzetten.

Door, met andere woorden, $c-1$ aantal dummyvariabelen aan te maken (waarbij c staat voor het aantal categorieën) kan je telkens afleiden tot welke categorie een waarneming behoort. De referentiecategorie is die categorie waarvoor alle dummyvariabelen op nul staan of **afstaan**.

10.3.4



Het voordeel van dummyvariabelen is dat je ze kan toevoegen aan een regressieanalyse en doen alsof het gaat om een kwantitatieve variabele.

Waarom mag je dat nu wel met een dummyvariabele en niet indien het gaat om bijvoorbeeld de categorische variabele Geslacht in zijn oorspronkelijke format, waarbij we de waarden 1 en 2 zouden hanteren in een regressieanalyse? Kan je daar een uitleg voor verzinnen?

10.3.5



In R kan je een dummyvariabele aanmaken via verschillende wegen. Eén van de mogelijkheden is werken via een voorwaarde. Bijvoorbeeld, je hebt een variabele X met twee categorieën: "Wit" en "Zwart". Je wilt een dummyvariabele aanmaken die aanstaat voor "Wit". Dit kan via het volgende commando, waardoor een nieuwe variabele met de naam Wit wordt aangemaakt die een waarde 1 heeft voor observatie-eenheden die "Wit" scoorden op X:

```
> Wit <- (X=="Wit")*1
```

Hierbij is het belangrijk om tussen de aanhalingstekens het label van de categorie correct te typen, dus identiek met aandacht voor hoofd-en kleine letters.

Bij meerdere categorieën dien je deze stappen een aantal keer te doorlopen om te komen tot de reeks dummyvariabelen.

10.3.6



Voor de onderzoeksvraag "Beïnvloedt de mate waarin studenten de leerstof proberen te begrijpen (schaalscore) hun score op een wetenschappentoets, ongeacht het geslacht van de student?" dienen we, voor het uitvoeren van een regressieanalyse, eerst een dummyvariabele aan te maken voor de variabele die het geslacht van de studenten weergeeft op basis van de variabele "Sex". We kunnen er bijvoorbeeld voor kiezen om een dummyvariabele te maken die aanstaat voor jongens:

```
> Pisa3$Jongen <- (Pisa3$Sex=="Male")*1
```

Via een `table()` functie kunnen we nagaan of we goed te werk zijn gegaan. In de rijen plaatsen we hieronder de scores voor de nieuwe variabele (0 en 1) en in de kolommen de waarden van de oorspronkelijke variabele. Hieruit lezen we af dat de nieuwe variabele aanstaat voor Jongen.

```
> table(Pisa3$Jongen, Pisa3$Sex)
```

	Female	Male
0	2271	0
1	0	2325

Vervolgens voeren we een regressieanalyse uit met de aangemaakte dummyvariabele. Hiervoor maken we gebruik van achtereenvolgens deze commando's:

```
Model <- lm(y ~ x1 + D, data = Data)
summary(Model)
```

Voor onze onderzoeksvraag "Beïnvloedt de mate waarin studenten de leerstof proberen te begrijpen (schaalscore) hun score op een wetenschappentoets, ongeacht het geslacht van de student?", geeft dit volgend resultaat:

```
> Model7 <- lm(Wetenschappenz ~ Begrijpenz + Jongen, data = Pisa3)
> summary(Model7)
```

```

Call:
lm(formula = Wetenschapenz ~ Begrijpenz + Jongen, data = Pisa3)

Residuals:
    Min       1Q   Median       3Q      Max
-2.95176 -0.53192  0.04277  0.56821  2.36323

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.02003    0.01731   1.157   0.247
Begrijpenz    0.43156    0.01228  35.151 < 2e-16 ***
Jongen        0.16968    0.02455   6.911 5.51e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8026 on 4321 degrees of freedom
(272 observations deleted due to missingness)
Multiple R-squared:  0.2238,    Adjusted R-squared:  0.2234
F-statistic: 622.8 on 2 and 4321 DF,  p-value: < 2.2e-16

```

In de output, meerbepaald de laatste kolom, lezen we af welke van de twee variabelen in ons model er toe doen om de score op de wetenschapstest te verklaren. In dit geval doen beide variabelen er toe, gezien de p-waarde voor beide variabelen kleiner is dan 0,05. Zowel geslacht als de mate van begrijpen zijn dus statistisch significante voorspellers van de score op wetenschappen.

Vervolgens kunnen we de parameters voor de regressielijnen terugvinden in de eerste kolom. Het intercept geeft hier de waarde van het intercept weer voor de referentiecategorie. Dit is de waarde die we verwachten voor wetenschappen voor een respondent die nul scoort op Begrijpenz en die behoort tot de referentiecategorie, nl. die categorie waarvoor de dummyvariabele op nul staat of **afstaat** (= meisjes). Het intercept voor meisjes die gemiddeld scoren op begrijpen is dus in de steekproef 0,020. Anders verwoord, in de steekproef scoren meisjes die gemiddeld scoren op begrijpen 0,020 SD hoger dan gemiddeld op de wetenschapstest. Het intercept voor jongens met eenzelfde gemiddelde score op begrijpen in de steekproef is 0,170 hoger (zie parameterschatting van de variabele Jongen), meerbepaald 0,190 (= 0,020 + 0,170). Anders verwoord, in de steekproef scoren jongens die gemiddeld scoren op begrijpen 0,190 SD hoger dan gemiddeld op de wetenschapstest.

Voor elke toename van 1 standaardafwijking (SD) in Begrijpenz, stijgt de score op de wetenschapstoets met 0,431 SD, ongeacht geslacht.

Of, als we dit alles samenbrengen in twee vergelijkingen:

$$\text{Wetenschapsscore}_{\text{meisjes}} = 0,020 + 0,431 \cdot \text{Begrijpenz}$$

$$\text{Wetenschapsscore}_{\text{jongens}} = 0,020 + 0,170 + 0,431 \cdot \text{Begrijpenz}$$

We zien dat het intercept voor meisjes niet significant is. We nemen dus aan dat deze parameterschatting even goed nul had kunnen zijn. Gezien de z-scores, betekent dit dat we verwachten dat in de populatie meisjes, die gemiddeld scoren op begrijpen, een gemiddelde score zullen behalen op de wetenschapstest. Voor begrijpen stellen we vast dat het een significant effect heeft ($p < 0,05$). Met andere

woorden de kans dat we in onze steekproef een verband zouden vaststellen indien er geen was in de populatie is wel zeer klein. Hetzelfde geldt voor geslacht.

10.3.7



Op basis van de parameterschattingen, kunnen we de scores van verschillende leerlingen gaan voorspellen. Geef de verwachte score voor wetenschappen voor onderstaande leerlingen, in de steekproef en in de populatie.

steekproef	Meisje	Jongen
-1 op Begrijpenz		
1,5 op Begrijpenz		

populatie	Meisje	Jongen
-1 op Begrijpenz		
1,5 op Begrijpenz		

10.3.8



Naast de informatie over de effecten van de onafhankelijke variabelen, bevat de output ook informatie over het model in zijn geheel.

```
> Model7 <- lm(Wetenschappenz ~ Begrijpenz + Jongen, data = Pisa3)
> summary(Model7)
```

Call:

```
lm(formula = Wetenschappenz ~ Begrijpenz + Jongen, data = Pisa3)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-2.95176 -0.53192  0.04277  0.56821  2.36323
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02003    0.01731   1.157    0.247
Begrijpenz   0.43156    0.01228  35.151 < 2e-16 ***
Jongen       0.16968    0.02455   6.911 5.51e-12 ***
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.8026 on 4321 degrees of freedom
```

```
(272 observations deleted due to missingness)
```

```
Multiple R-squared:  0.2238,    Adjusted R-squared:  0.2234
```

```
F-statistic: 622.8 on 2 and 4321 DF,  p-value: < 2.2e-16
```

De p-waarde onderaan de output geeft aan wat de p-waarde is voor ons model (p-value: < 2.2e-16). Deze p-waarde dienen we te interpreteren als de kans dat de nulhypothese geldig is. In dit geval luidt de nulhypothese dat ons model (bestaande uit twee onafhankelijke variabelen) in de populatie geen verschillen kan verklaren aangaande wetenschapsscores van leerlingen. Is deze p-waarde met andere woorden kleiner dan 0,05 dan kunnen we besluiten dat de kans wel heel klein is dat ons model in de populatie niet in staat is om vastgestelde verschillen te verklaren.

Bij Adjusted R-Squared lezen we af hoe goed ons model in z'n geheel is. Het geeft aan hoeveel procent van de variantie in onze afhankelijke variabele verklaard kan worden door ons model. Daarbij houden we steeds in het achterhoofd dat het model in dit geval moet beschouwd worden als de combinatie van onze onafhankelijke variabelen. In dit voorbeeld kunnen we stellen dat de combinatie van begrijpen en geslacht ons in staat stellen om 22,34% van de variantie in wetenschapsscores te verklaren. Deze Adjusted R-Squared geeft een preciezer beeld voor de verklaarde proportie variantie dan de Multiple R-Squared aangezien er een correctie is gemaakt voor modelcomplexiteit. Als we dus onnodige (zijnde niet-significante) parameters in ons model opnemen, zal de Adjusted R-squared dit in rekening brengen.

10.3.9



We hernemen het voorbeeld waarin we nagingen wat de invloed was van Iq en Urengestudeerd op de wiskundescore van leerlingen. Na het uitvoeren van de analyse zegt een collega dat je ook dient te controleren voor het effect van Geslacht.

Open het databestand Wis2.RData en standaardiseer eerst de variabelen Iq en Urengestudeerd. Maak voor de variabele Geslacht een dummyvariabele aan die aangeeft of het al dan niet om een jongen gaat en noem deze variabele "Jongen". Bereken de frequentietabel voor de originele variabele en de dummyvariabele en ga na of je juist te werk bent gegaan.

Voer vervolgens de nodige analyse uit om na te gaan of deze variabelen, ongeacht het geslacht een effect hebben op de variabele wiskundescore.

10.3.10



We werken verder met databestand Wis2.RData. Maak voor de variabele Iqcategorisch (een categorische variant voor de Iq variabele met drie categorieën) een reeks van dummyvariabelen aan, met als referentiecategorie "gemiddeld". Bereken de frequentietabel voor de originele variabele en de dummyvariabele en ga na of je juist te werk bent gegaan.

Herhaal de regressieanalyse uit 10.3.10 maar vervang de Iq_z variabele door de twee dummyvariabele die je aanmaakte voor de variabele Iqcategorisch. Ook de dummyvariabele Jongen laat je in de analyse. Beantwoord daarna de volgende vragen:

- a) Wat betekent de waarde van het intercept?
- b) Scoren leerlingen uit de categorie met een hoog Iq hoger dan leerlingen met een gemiddeld Iq?
- c) Kan je eveneens nagaan of leerlingen uit de categorie met een hoog Iq significant hoger scoren op wiskunde dan leerlingen uit de categorie met een laag Iq? Zo nee, welke stappen zou je kunnen suggereren om dit toch na te gaan?

10.3.11



We gaan aan de slag met het databestand `Werk.RData`. Maak voor de variabele "Statuut" een reeks van dummyvariabelen aan, met als referentiecategorie "arbeider".

Ga vervolgens na of er een effect is van open bedrijfsklimaat ("Klimaatz") op het welbevinden op het werk ("Welbevindenz") na controle voor het statuut van medewerkers (arbeiders, bedienden en kaderleden)?

Bereken tot slot voor onderstaande respondenten hun verwachte welbevinden.

steekproef	Arbeider	Bediende	Kader
-1,5 op klimaat			
0,5 op klimaat			

populatie	Arbeider	Bediende	Kader
-1,5 op klimaat			
0,5 op klimaat			

10.3.12



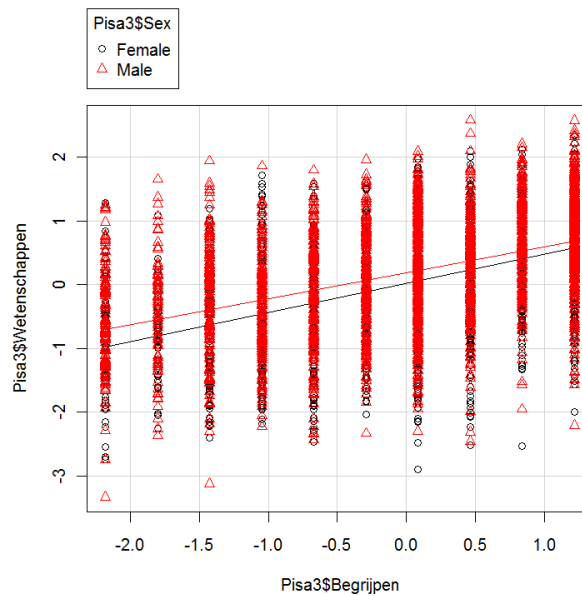
Tot hiertoe zijn we er vanuit gegaan dat twee parallelle regressielijnen de beste benadering is van de realiteit. We veronderstellen in ons dummyregressie model (10.3.6) dat het effect van begrijpen gelijk loopt voor beide geslachten: de hellingsgraad is identiek voor jongens en meisjes.

We kunnen ons afvragen of dit wel klopt. Bepaalde onderzoeksvragen verwachten ook expliciet dat je dit gaat nagaan. Bijvoorbeeld, "Is het effect van de mate waarin studenten de leerstof proberen te begrijpen op de toetsscore voor wetenschappen, verschillend voor jongens dan meisjes?"

Wederom kunnen we in eerste instantie visueel gaan kijken wat de data ons vertellen. Hiervoor hernemen we het commando `scatterplot`, maar laten we het laatste stukje (`regLine=FALSE`) weg.

```
> scatterplot(y~xkwantitatief/xkwalitatief, smooth=FALSE)

> scatterplot(Pisa3$Wetenschappenz~Pisa3$Begrijpenz/Pisa3$Sex,
              smooth=FALSE)
```



Figuur 10.7 Scatterplot met interactie tussen geslacht en begrijpen

10.3.13

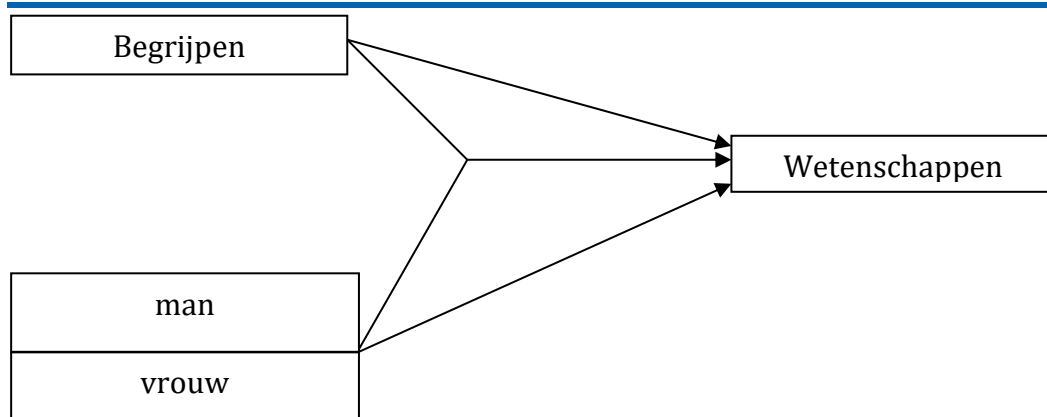


Wat zou jij uit Figuur 10.7 concluderen?

10.3.14



We kunnen aan de hand van een dummyregressie ook nagaan of het beschrijven van het verband tussen begrijpen, geslacht en wetenschapsscore aan de hand van deze twee niet parallelle lijnen een goede voorstelling is van de realiteit. We kunnen dit model als volgt formeel visualiseren:



Figuur 10.8 Analyseschema dummyregressie mét interactie-effect

Dit model zegt dat we verwachten dat er een verschil is tussen meisjes en jongens aangaande hun wetenschapsprestaties. Dit wordt voorgesteld door de directe pijl van geslacht naar wetenschappen. Daarnaast verwachten we dat er een interactie-effect is tussen geslacht en begrijpen: de invloed van de mate van begrijpen op de wetenschapsscores is niet identiek voor beide geslachten. Ten derde nemen we ook de directe pijl tussen begrijpen en wetenschappen op. Dit is een afspraak die

we maakten in hoofdstuk 1: als we een interactie-effect meenemen, nemen we altijd de directe effecten ook mee.

Hoe gaan we nu gelijkaardige interactie-effecten toevoegen aan een regressieanalyse met dummyvariabelen?

Vanuit de invalshoek van regressieanalyse zijn er verschillende methoden om interactieeffecten te gaan schatten¹. We lichten hier de meest gebruikte toe: via **producttermen**.

Als we de regressievergelijking van de laatste analyse, zonder interactie, opschrijven dan ziet die er als volgt uit:

$$Y = \beta_1 + \beta_2 * \text{Begrijpenz} + \beta_3 * \text{Jongen}$$

Hierbij zijn de beta's de verschillende onbekenden die we willen schatten (= de parameters). Net als voor de rechtstreekse verbanden willen we ook voor ons interactieeffect een parameter schatten. We vullen met andere woorden de regressievergelijking aan met een extra "term", de interactieterm:

$$Y = \beta_1 + \beta_2 * \text{Begrijpenz} + \beta_3 * \text{Jongen} + \beta_4 * \text{interactieterm}$$

Deze interactieterm is in feite een nieuwe variabele die we willen toevoegen aan de analyse om grip te krijgen op hoe het effect tussen een onafhankelijke en afhankelijke variabele beïnvloed wordt door een andere onafhankelijke variabele. Deze nieuwe variabele maken we aan door het **product** te nemen tussen de twee onafhankelijke variabelen waartussen we een interactie verwachten.

10.3.15



Laat we terugkeren naar ons voorbeeld. Op basis van Figuur 10.7 veronderstellen we dat de invloed van de mate van begrijpen op de wetenschapsscores niet identiek is voor beide geslachten. Indien we dit willen toetsen in een analyse, dan zouden we een nieuwe variabele moeten aanmaken die het product vormt tussen de dummyvariabele Jongen en de variabele Begrijpenz. In de onderstaande tabel hebben we de scores van vier hypothetische leerlingen op de betrokken variabelen geplaatst.

	Jongen	Begrijpenz	Interactieterm
l1n1	1	1	1
l1n2	0	1	0
l1n3	1	0	0
l1n4	0	0	0

¹ Jaccard, J., Turrissi, R., en Wan, C.K. (1990). *Interaction effects in multiple regression*. Newsbury Park: Sage Publications.

Stel dat we nu de volgende regressievergelijking schatten:

$$Y = \beta_1 + \beta_2 \text{ Jongen} + \beta_3 \text{ Begrijpenz} + \beta_4 \text{ Interactieterm}$$

- a) Voor welke van de vier leerlingen zou het intercept gelden als verwachte score?
- b) Voor welke leerling(en) dienen we enkel β_2 toe te voegen aan het intercept om de verwachte score te schatten?
- c) Voor welke leerling(en) dienen we enkel β_3 toe te voegen aan het intercept om de verwachte score te schatten?
- d) Welke beta's moeten we hanteren om een verwachte score te berekenen voor ln1?

10.3.16



Het toevoegen van een interactieterm in R, is een vrij eenvoudig ingreep. Je dient in de feiten niet echt een nieuwe variabele aan te maken. Je kan in de regressievergelijking bij de functie `lm()` simpelweg een extra term opnemen die de interactieterm voorstelt. Enkele voorbeelden maken dit duidelijk.

Stel we willen het effect van een variabele en een dummyvariabele (x1 en D1) én het interactieeffect van beiden op Y nagaan. Daartoe volstaat de volgende R-code. `x1*D1` is de interactieterm die we extra toevoegen aan de analyse.

```
Model<-lm(y~x1+D1+x1*D1)
summary(Model)
```

Stel we willen het effect van twee dummyvariabelen (D1 en D2) en het interactieeffect van beide op Y nagaan. Daartoe volstaat de volgende R-code. `D1*D2` is de interactieterm die we extra toevoegen aan de analyse.

```
Model<-lm(Y~D1+D2+D1*D2)
Summary(Model)
```

Je hebt dus altijd evenveel termen na de `~` dan pijlen in het analyseschema, in dit geval drie.

We keren terug naar ons voorbeeld waarin we veronderstellen dat de invloed van de mate van begrijpen op de wetenschapsscores verschilt naargelang geslacht. De analyse resulteert in volgende output:

```
> Model11 <- lm(Wetenschappenz ~ Begrijpenz + Jongen
+ Begrijpenz*Jongen, data = Pisa3)

> summary(Model11)

Call:
lm(formula = Wetenschappenz ~ Begrijpenz + Jongen + Begrijpenz *
    Jongen, data = Pisa3)

Residuals:
    Min       1Q   Median       3Q      Max
-2.95120 -0.53140  0.04267  0.56707  2.33401
```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.01727   0.01736   0.995  0.3199
Begrijpenz     0.45736   0.01805  25.335 < 2e-16 ***
Jongen         0.17007   0.02454   6.929 4.87e-12 ***
Begrijpenz:Jongen -0.04798  0.02462  -1.949  0.0513 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8023 on 4320 degrees of freedom
(272 observations deleted due to missingness)
Multiple R-squared:  0.2244,    Adjusted R-squared:  0.2239
F-statistic: 416.7 on 3 and 4320 DF,  p-value: < 2.2e-16

```

Als we kijken naar de parameters van het model, zien we dat zowel begrijpen als geslacht een significante invloed hebben op wetenschapsscore. Voor de interactie tussen beiden zit dit op de grens van statistische significantie ($p=0,0513$). Voor een meisje dat gemiddeld scoort op begrijpen verwachten we in de steekproef een score op wetenschappen van 0,017. Dit is niet statistisch significant, dus in de populatie verwachten voor zo'n leerling een gemiddelde score op de wetenschapstest. De hellingsgraad van de regressielijn voor meisjes bedraagt 0,457.

$$\text{Wetenschapsscore}_{\text{meisjes}} = 0,017 + 0,457 \cdot \text{Begrijpenz}$$

Voor een jongen wordt het ietwat complexer. Net zoals in het vorige voorbeeld, is er een ander intercept voor jongens ($0,187 = 0,017 + 0,170$). Daarnaast is nu ook de hellingsgraad voor jongens anders. Per toename van 1 standaardafwijking op Begrijpenz, stijgt de voorspelde wetenschapsscore met 0,409 standaardafwijkingen ($= 0,457 - 0,048$). Het effect van begrijpen is voor jongens dus minder uitgesproken dan voor meisjes.

$$\text{Wetenschapsscore}_{\text{jongens}} = 0,017 + 0,170 \cdot \text{Jongen} + 0,457 \cdot \text{Begrijpenz} - 0,048 \cdot \text{Begrijpenz} \cdot \text{Jongen}$$

10.3.17



Op basis van de parameterschattingen kunnen we de scores van verschillende leerlingen gaan voorspellen. Geef de verwachte score op wetenschappen voor onderstaande leerlingen, in de steekproef en in de populatie. Beschouw hierbij de interactieterm als significant.

steekproef	Meisje	Jongen
Gemiddelde score op begrijpen		
-1 op begrijpen		
1,5 op begrijpen		

populatie	Meisje	Jongen
Gemiddelde score op begrijpen		
-1 op begrijpen		
1,5 op begrijpen		

10.3.18



Als we kijken naar het model in zijn geheel (10.3.16), merken we dat het een goed model is dat door te trekken is naar de populatie. De p-waarde (p-value: $< 2.2e-16$) is namelijk kleiner dan 0,05. In zijn geheel, de drie parameters samen, verklaren 22,39% van de variantie in de score op de wetenschapstest. Dit is echter nauwelijks meer dan het model zonder interactieterm (zie 10.3.8), wat in vraag stelt of die interactieterm wel zo nuttig is in het verklaren van de scores op wetenschappen.

We kunnen dit vermoeden in R ook expliciet gaan testen. Is het model mét interactieterm beter is dan het model zonder? Hiervoor moeten we opnieuw eerst een dataframe aanmaken zonder missings erin. Daarna kunnen we met behulp van het `anova` commando de `lm`-modellen vergelijken.

```
> Data2<-na.omit(Data[ c("x1","x2","x3","y")])
> anova(Model1, Model2)
```

In ons voorbeeld geeft dit:

```
> Pisa4<-na.omit(Pisa3[ c("Wetenschappenz","Begrijpenz","Jongen")])
> Model12 <- lm(Wetenschappenz ~ Begrijpenz + Jongen, data = Pisa4)
> Model13 <- lm(Wetenschappenz ~ Begrijpenz + Jongen
+ Begrijpenz*Jongen, data = Pisa4)

> anova(Model12, Model13)
Analysis of Variance Table

Model 1: Wetenschappenz ~ Begrijpenz + Jongen
Model 2: Wetenschappenz ~ Begrijpenz + Jongen + Begrijpenz * Jongen
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1   4321 2783.3
2   4320 2780.9   1    2.4455 3.7991 0.05135 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Uit de bovenstaande output kunnen we ten eerste aflezen dat de RSS (Residual Sum of Squares) nauwelijks kleiner is bij Model 13 (2780,9 t.o.v. 2783,3). Model 13 leidt dus niet tot veel betere voorspellingen dan Model 12. Bovendien lezen we af dat het verschil in 'Model fit' net niet significant ($p=0,0513$). We weten dat dit hier niet te wijten is aan een kleine dataset. Daarom besluiten we dat Model 12 als beste model is voor de populatie.

10.3.19



We bouwen door op de regressieanalyse uit 10.3.11. Ga na of het effect van open bedrijfsklimaat op het welbevinden op het werk verschilt naar het statuut van medewerkers (arbeiders, bedienden en kaderleden). Negeer hierbij even de bevindingen uit 10.3.11.

- a) Is dit model een beter dan het model zonder interactie-effect?
- b) Stel visueel voor en bereken de verwachte score in de populatie voor een kaderlid dat 1,5 standaardafwijkingen scoort op klimaat.

10.3.20



We bouwen door op de regressieanalyse uit 10.3.10 en voegen er een interactieterm aan toe tussen Iq_laag en Jongen.

- a) Is in de populatie het effect van geslacht anders voor leerlingen met een laag Iq dan voor de andere leerlingen?
- b) Bereken de verwachte score in de populatie voor jongens die behoren tot de categorie Iq hoog en voor jongens die behoren tot de categorie Iq laag, die een gemiddeld aantal uren gestudeerd hebben?

10.3.21



Een collega-onderzoeker oppert de mogelijkheid dat er ook een interactie is tussen het aantal uren dat een leerling studeert en het Iq. Volgens hem zou het aantal uren dat een leerling studeert een sterker effect moeten hebben bij kinderen met een laag Iq. Ga deze hypothese na door de laatste regressieanalyse uit 10.3.20 verder uit te bereiden.

10.3.22



Deze laatste oefening toont meteen de meerwaarde van het werken met dummyvariabelen. Je kan namelijk zeer soepel omgaan met het toetsen van hypothesen, ook als het gaat om interactie-effecten.

Een bijkomend voordeel van het werken met dummyvariabelen is dat je meer mogelijkheden hebt om effecten van het behoren tot één of meerdere categorieën na te gaan. Een voorbeeld maakt duidelijk wat we hiermee bedoelen. Stel dat we goede redenen hebben om aan te nemen dat enkel leerlingen met een laag Iq anders gaan scoren voor een wiskundetoets dan de overige leerlingen. Dit vraagt eigenlijk om een vergelijking van leerlingen in de categorie Iq laag met de leerlingen uit de twee overige categorieën (Iq gemiddeld en Iq hoog). Door enkel en alleen de dummyvariabele Iq_laag op te nemen in het regressiemodel (samen met evt. andere controle- of interactievariabelen) toetsen we rechtstreeks deze hypothese.

Daarnaast kan je via dummyvariabelen ook meerdere categorieën samen nemen. We zouden bijvoorbeeld een dummyvariabele kunnen aanmaken die aanstaat voor leerlingen die ofwel gemiddeld ofwel hoog scoren voor Iq. Dus, in één dummyvariabele combineren we twee categorieën. Dit bespaart ons soms een hoop hercodeerwerk.

Respons bij hoofdstuk 10

RESPONS 10.1.1

De uitwerking van deze analyse gaat als volgt.

```
> Modell<-lm(Wiskunde~Urengestudeerd+Iq,data=Wis2)
> summary(Moell)
```

Call:

```
lm(formula = Wiskunde ~ Urengestudeerd + Iq, data = Wis2)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.4678	-2.2676	0.1467	3.1231	16.7707

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	14.71005	4.38329	3.356	0.00123	**
Urengestudeerd	0.74867	0.22911	3.268	0.00162	**
Iq	0.54182	0.04318	12.547	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

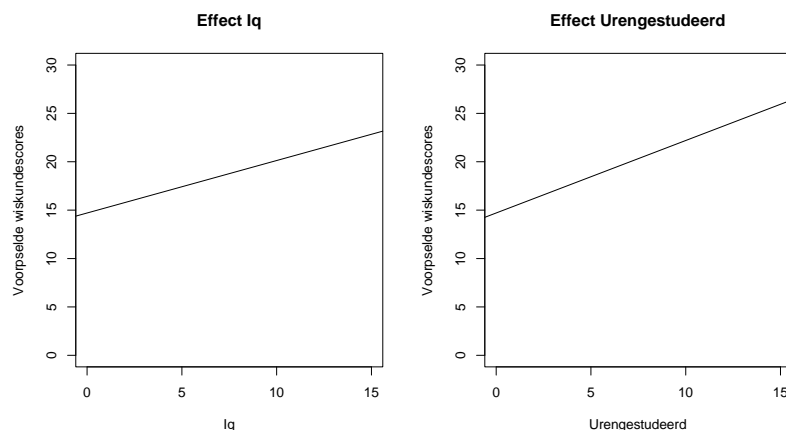
Residual standard error: 5.594 on 77 degrees of freedom

Multiple R-squared: 0.7166, Adjusted R-squared: 0.7093

F-statistic: 97.36 on 2 and 77 DF, p-value: < 2.2e-16

Geometrisch kunnen we de waarde 14,71 interpreteren als het snijpunt van de regressielijn met de Y-as (zie figuur 10.1.1 hieronder).

Dit is m.a.w. de wiskundescore die een leerling met een nulscore op Iq en Urengestudeerd zou behalen. Deze informatie is eigenlijk van weinig of geen waarde, voornamelijk voor variabele Iq. Een leerling zonder IQ bestaat niet. Zelfs een chimpansee zou een zekere score moeten behalen op een IQ-test.



Figuur 10.1.1 Regressierechte voor het effect van zowel Iq als Urengestudeerd op Wiskunde

RESPONS 10.1.3

Een interessante referentiewaarde zou bijvoorbeeld het gemiddeld Iq kunnen zijn. Mochten we erin slagen om een variabele Iq te creëren die de waarde nul heeft voor personen met een gemiddeld Iq, dan is ook het intercept informatiever. Het intercept is dan de verwachte wiskundescore voor een leerling met een gemiddeld Iq.

Andere mogelijk interessante referentiewaarden kunnen zijn:

- de empirische minimumscore: stel dat je een Iq variabele maakt die de waarde nul heeft voor de persoon met het laagste Iq, dan krijgt het intercept de volgende betekenis: de verwachte wiskundescore voor “de domste leerling”;
- de mediaan

RESPONS 10.1.5

We kunnen best een nieuwe variabele aanmaken die de waarde nul heeft voor leerlingen met een gemiddeld Iq. Hiertoe kunnen we simpelweg voor elke leerling de Iq-score verminderen met het gemiddelde Iq (100).

In de onderstaande tabel staan 3 hypothetische leerlingen uit het bestand. Voor elk van hen hebben we de oorspronkelijke Iq-score weergegeven en de nieuwe score op de variabele die we Iq_gecentreerd hebben genoemd.

Uit deze tabel kan je duidelijk aflezen dat we de verschillen tussen leerlingen niet verliezen door deze stap. Leerling 1 blijft 10 punten hoger scoren dan leerling 2 en leerling 2 blijft op zijn beurt 35 punten hoger scoren dan leerling 3.

Iq		Iq_gecentreerd
110	(-100) →	10
100	(-100) →	0
65	(-100) →	-35

Een positieve score op deze variabele wil met andere woorden zeggen dat je als leerling hoger scoort dan gemiddeld, een negatieve score wijst daarentegen op lager scoren dan gemiddeld.

RESPONS 10.1.7

De onderstaande R-code geeft de nodige stappen weer om beide variabelen te centreren rond het gemiddelde en vervolgens de analyse opnieuw te doen.

```
> Wis2$Iq_c<-Wis2$Iq-mean(Wis2$Iq,na.rm=TRUE)
> Wis2$Urengestudeerd_c<-Wis2$Urengestudeerd-
  mean(Wis2$Urengestudeerd,na.rm=TRUE)
> Model2<-lm(Wiskunde~Urengestudeerd_c+Iq_c,data=Wis2)
> summary(Model2)
```

```

Call:
lm(formula = Wiskunde ~ Urengestudeerd_c + Iq_c, data = Wis2)

Residuals:
    Min       1Q   Median       3Q      Max
-25.4678  -2.2676   0.1467   3.1231  16.7707

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    75.25000     0.62543  120.317  < 2e-16 ***
Urengestudeerd_c  0.74867     0.22911   3.268  0.00162 **
Iq_c            0.54182     0.04318  12.547  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.594 on 77 degrees of freedom
Multiple R-squared:  0.7166,    Adjusted R-squared:  0.7093
F-statistic: 97.36 on 2 and 77 DF,  p-value: < 2.2e-16

```

- a) De regressiecoëfficiënten van zowel de variabele Iq als de variabele Urengestudeerd in bovenstaande output zijn identiek aan de parameters die we in de output van 10.1.1. kunnen terugvinden. Het centreren heeft met andere woorden geen invloed gehad op de schattingen. Enkel het intercept is een stuk hoger (75,25).
- b) Het intercept kunnen we nu interpreteren als de verwachte wiskundescore voor een leerling met een gemiddeld Iq die een gemiddeld aantal uren gestudeerd heeft voor de toets. Deze leerling behaalt 75,25 punten volgens onze analyse. Voor elk uur dat een leerling meer studeert wint hij 0,749 punten voor wiskunde. En per punt dat hij of zij hoger scoort op de Iq-toets scoort hij/zij 0,542 punten hoger voor wiskunde.

RESPONS 10.1.9

In deze hypothese is de variabele Diepteverw de afhankelijke variabele. De verklarende variabelen zijn Zelfontd en Opnamek. Deze laatste twee dienen we te centeren. Dit doen we via de volgende commando's.

```

> Studenten2$Opnamek_c<-Studenten2$Opnamek-
  mean(Studenten2$Opnamek, na.rm=TRUE)

> Studenten2$Zelfontd_c<-Studenten2$Zelfontd-
  mean(Studenten2$Zelfontd, na.rm=TRUE)

```

Daarna doen we een regressieanalyse met de twee nieuwe variabelen als onafhankelijke variabelen. Dit resulteert in de volgende parameterschattingen:

```

> Model3<-lm(Diepteverw~Opnamek_c+Zelfontd_c,data=Studenten2)
> summary(Model3)

Call:
lm(formula = Diepteverw ~ Opnamek_c + Zelfontd_c, data = Studenten2)

Residuals:
    Min       1Q   Median       3Q      Max
-1.620309 -0.235525  0.005622  0.226418  1.895423

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.16470    0.03179  99.563 < 2e-16 ***
Opnamek_c     0.14958    0.04731   3.162  0.00185 **
Zelfontd_c    0.66491    0.05039  13.194 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4228 on 174 degrees of freedom
(3 observations deleted due to missingness)
Multiple R-squared:  0.7941,    Adjusted R-squared:  0.7918
F-statistic: 335.6 on 2 and 174 DF,  p-value: < 2.2e-16

```

Het intercept kunnen we interpreteren als de verwachte score van een gemiddelde student aangaande hun opvattingen over leren. Of anders gesteld, een student die gemiddeld scoort op zowel de variabele Opname van kennis en Zelfontdekkend leren scoort volgens deze analyse 3,165 punten op de variabele diepteverwerking. Beide verklarende variabelen blijken bovendien een significant positief effect te hebben op de afhankelijke variabele. Het eerste deel van de hypothese wordt met andere woorden niet ondersteund door deze analyse.

RESPONS 10.2.2

```

> summary(Model4)

Call:
lm(formula = Wiskunde ~ Iq_z + Urengestudeerd_z, data = Wis2)

Residuals:
    Min       1Q   Median       3Q      Max
-25.4678  -2.2676   0.1467   3.1231  16.7707

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    75.2500    0.6254 120.317 < 2e-16 ***
Iq_z            8.0856    0.6444  12.547 < 2e-16 ***
Urengestudeerd_z  2.1059    0.6444   3.268  0.00162 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.594 on 77 degrees of freedom
Multiple R-squared:  0.7166,    Adjusted R-squared:  0.7093
F-statistic: 97.36 on 2 and 77 DF,  p-value: < 2.2e-16

```

- a) Het intercept kunnen we interpreteren als de verwachte score van een leerling met een gemiddelde Iq-score en die een gemiddeld aantal uren heeft gestudeerd. Het is namelijk, net zoals bij gecentreerde scores, dat een waarde nul bij een z-score staat voor gemiddeld scoren op...
- b) De regressiecoëfficiënt voor bijvoorbeeld Iq kunnen we als volgt interpreteren: een leerling met een Iq-score die 1 standaardafwijking hoger is dan gemiddeld zal 8,09 punten hoger scoren op de wiskundetoets. Of nog anders, als je tot de 16% slimste leerlingen behoort, scoor je volgens deze analyse 8,09 punten hoger op wiskunde dan de gemiddelde leerling.

RESPONS 10.2.5

Om deze vraag te voltooien dienen we in een eerste stap de Zscores te berekenen voor zowel Opname van kennis als voor Zelfontdekkend leren.

```
> Studenten2$Opnamek_z<-scale(Studenten2$Opnamek)
> Studenten2$Zelfontd_z<-scale(Studenten2$Zelfontd)
```

Vervolgens voegen we de nieuwe variabelen in als verklarende variabelen bij een regressieanalyse. Dit resulteert in de volgende schattingen:

```
> Model5<-lm(Diepteverw~Opnamek_z+Zelfontd_z,data=Studenten2)
> summary(Model5)
```

Call:

```
lm(formula = Diepteverw ~ Opnamek_z + Zelfontd_z, data = Studenten2)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.620309	-0.235525	0.005622	0.226418	1.895423

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.16470	0.03179	99.563	< 2e-16 ***
Opnamek_z	0.16502	0.05220	3.162	0.00185 **
Zelfontd_z	0.69124	0.05239	13.194	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4228 on 174 degrees of freedom
(3 observations deleted due to missingness)

Multiple R-squared: 0.7941, Adjusted R-squared: 0.7918

F-statistic: 335.6 on 2 and 174 DF, p-value: < 2.2e-16

Een student die gemiddeld scoort op zowel de variabele Opname van kennis en Zelfontdekkend leren scoort volgens deze analyse 3,165 punten op de variabele diepteverwerking. Beide verklarende variabelen blijken bovendien een significant positief effect te hebben op de afhankelijke variabele. Eén standaardafwijking hoger dan gemiddeld scoren op de variabele Opname van kennis leidt tot een stijging van 0,165 punten op de variabele diepteverwerking. Met andere woorden leerlingen die tot de 16% hoogst scorende leerlingen voor de variabele opname van kennis behoren scoren minstens 0,165 punten hoger op diepteverwerking. Voor Zelfontdekkend leren is het effect sterker. Behoor je daar bij de 16% hoogst scorende studenten dan scoor je minstens 0,691 punten hoger op diepteverwerking.

RESPONS 10.2.7

Dit resulteert in de volgende schattingen:

```
> Studenten2$Diepteverw_z<-scale(Studenten2$Diepteverw)
> Model6<-lm(Diepteverw_z~Opnamek_z+Zelfontd_z,data=Studenten2)
> summary(Model6)
```

```

Call:
lm(formula = Diepteverw_z ~ Opnamek_z + Zelfontd_z, data = Studenten2)

Residuals:
    Min       1Q   Median       3Q      Max
-1.747776 -0.254053  0.006064  0.244230  2.044533

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01901    0.03429  -0.555  0.57992
Opnamek_z    0.17801    0.05630   3.162  0.00185 **
Zelfontd_z    0.74562    0.05651  13.194 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4561 on 174 degrees of freedom
(3 observations deleted due to missingness)
Multiple R-squared: 0.7941, Adjusted R-squared: 0.7918
F-statistic: 335.6 on 2 and 174 DF, p-value: < 2.2e-16

```

Een student die gemiddeld scoort op zowel de variabele Opname van kennis en Zelfontdekkend leren scoort volgens deze analyse 0,019 standaardafwijkingen lager dan gemiddeld op de variabele diepteverwerking. De schatting voor het intercept is niet statistisch significant (zie p-waarde). Dit wil zeggen dat in de populatie deze parameter net zo goed nul zou kunnen zijn. Dat brengt ons tot de conclusie dat gemiddeld scores op beide opvattingen ook leidt tot gemiddeld scores op diepteverwerking.

De regressiecoëfficiënten in analyses waarbij alle variabelen gestandaardiseerd zijn, worden ook wel eens gerapporteerd onder de noemer gestandaardiseerde regressiecoëfficiënten. In heel wat publicaties worden zowel de ongestandaardiseerde als de gestandaardiseerde regressiecoëfficiënten samen gerapporteerd.

Beide verklarende variabelen blijken bovendien een significant positief effect te hebben op de afhankelijke variabele. Eén standaardafwijking hoger dan gemiddeld scores op de variabele Opname van kennis leidt tot een stijging van 0,178 standaardafwijkingen op de variabele diepteverwerking. Met andere woorden, leerlingen die tot de 16% hoogst scorende leerlingen voor de variabele opname van kennis behoren scoren minstens 0,178 SD hoger op diepteverwerking. Voor Zelfontdekkend leren is het effect sterker. Behoor je daar bij de 16% hoogst scorende studenten dan scoor je minstens 0,746 SD hoger op diepteverwerking.

RESPONS 10.3.2

a) In welk opzicht verschillen jongens en meisjes van elkaar?

Als we kijken naar de regressielijn voor jongens enerzijds en meisjes anderzijds, merken we dat er enkel een verschil is in intercept. Jongens scoren gemiddeld gezien iets hoger op wetenschappen dan meisjes.

- b) Is het effect van begrijpen op wetenschapsscore onafhankelijk van geslacht?

Wat er niet verschilt tussen beide regressielijnen is de hellingsgraad of slope. De twee lijnen lopen parallel. Dit wil zeggen dat het effect van begrijpen op wetenschapsscore onafhankelijk is van geslacht. Één standaardafwijking hoger scoren op begrijpen, geeft dezelfde toename in wetenschapsscore voor jongens dan voor meisjes.

- c) Is het effect van geslacht op wetenschapsscore onafhankelijk van de mate van begrijpen?

Ja, het effect van geslacht op wetenschapsscore is onafhankelijk van de mate van begrijpen. De afstand tussen de twee regressielijnen is voor elke score van begrijpen even groot. Er is dus ook een effect voor geslacht, na controle voor de mate van begrijpen.

RESPONS 10.3.4

In tegenstelling tot de variabele geslacht in zijn oorspronkelijk formaat heeft de waarde nul en één een eenduidige betekenis gekregen bij een dummyvariabele. Stel dat je de variabele geslacht zou toevoegen, wat is dan de betekenis van het intercept? De verwachte score indien iemand 0 scoort op de variabele geslacht, maar voor wie geldt dit? Dit is onmogelijk. Bovendien, hoe zou je de regressiecoëfficiënt moeten interpreteren. Bij een dummyvariabele gaat dit niet op. Als je die toevoegt dan krijgt het intercept de betekenis van verwachte score voor de referentiecategorie en de regressiecoëfficiënt voor het effect van het behoren tot de categorie waarvoor de dummyvariabele aan staat. In dit geval bijvoorbeeld voor het effect van een jongen zijn.

RESPONS 10.3.7

steekproef	Meisje	Jongen
-1 op Begrijpenz	$0,020 + (-1 \cdot 0,431) = -0,411$	$0,020 + 0,170 + (-1 \cdot 0,431) = -0,241$
1,5 op Begrijpenz	$0,020 + (1,5 \cdot 0,431) = 0,667$	$0,020 + 0,170 + (1,5 \cdot 0,431) = 0,837$

populatie	Meisje	Jongen
-1 op Begrijpenz	$0 + (-1 \cdot 0,431) = -0,431$	$0 + 0,170 + (-1 \cdot 0,431) = -0,261$
1,5 op Begrijpenz	$0 + (1,5 \cdot 0,431) = 0,647$	$0 + 0,170 + (1,5 \cdot 0,431) = 0,817$

1) Variabelen Iq en Urengestudeerd standaardiseren:

```
> Wis2$Iq_z<-scale(Wis2$Iq)
> Wis2$Urengestudeerd_z<-scale(Wis2$Urengestudeerd)
```

2) Dummyvariabele aanmaken voor Geslacht die aanstaat voor "Jongen":

```
> Wis2$Jongen<-(Wis2$Geslacht=="Jongen")*1
> table(Wis2$Jongen,Wis2$Geslacht)
```

```
      Jongen Meisje
0         0      40
1        40       0
```

3) Regressie met dummyvariabele:

```
> Model8<-lm(Wiskunde~Iq_z+Urengestudeerd_z+Jongen,data=Wis2)
> summary(Model8)
```

Call:

```
lm(formula = Wiskunde ~ Iq_z + Urengestudeerd_z + Jongen, data = Wis2)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-22.9159  -2.9149   0.3939   2.8551  14.9111
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    73.4746    0.8748   83.986 < 2e-16 ***
Iq_z            7.5852    0.6434   11.789 < 2e-16 ***
Urengestudeerd_z 1.8603    0.6241    2.981 0.00386 **
Jongen          3.5507    1.2740    2.787 0.00672 **
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 5.363 on 76 degrees of freedom

Multiple R-squared: 0.7429, Adjusted R-squared: 0.7328

F-statistic: 73.2 on 3 and 76 DF, p-value: < 2.2e-16

In de output, meerbepaald de laatste kolom, lezen we af welke van de drie variabelen in ons model er toe doen om de score op de wiskundetoets te verklaren. In dit geval doen alle variabelen er toe, gezien de p-waarde voor elke variabele kleiner is dan 0,05. Zowel Iq, urengestudeerd als geslacht zijn dus statistisch significante voorspellers van de score op wiskunde zowel in de steekproef als in de populatie.

Vervolgens kunnen we de parameters voor de regressielijnen terug vinden in de eerste kolom. Het intercept geeft hier de waarde van het intercept weer voor de referentiecategorie. Dit is de waarde die we verwachten voor wiskunde voor een respondent die nul scoort op Iq_z en Urengestudeerd_z en die behoort tot de referentiecategorie, nl. die categorie waarvoor de dummyvariabele op nul staat of afstaat (= meisje). Het intercept voor meisjes met een gemiddeld Iq die een gemiddeld aantal uren hebben gestudeerd, is 73,47. Anders verwoord, in de steekproef en in de populatie ($p < 0.05$) scoren meisjes met een gemiddeld Iq die een gemiddeld aantal uren hebben

gestudeerd 73,47 punten op de wiskundetoets. Het intercept voor jongens met eenzelfde gemiddelde score op Iq_z en $Urengestudeerd_z$ is 3,55 hoger, meer bepaald 77,02 (= 73,47 + 3,55). Anders verwoord, in de steekproef en de populatie ($p < 0.05$) scoren jongens met een gemiddeld Iq die een gemiddeld aantal uren hebben gestudeerd 77,02 punten op de wiskundetoets.

Voor elke toename van 1 standaardafwijking in Iq_z , stijgt de score op de wiskundetoets met 7,585 punten, ongeacht geslacht.

Voor elke toename van 1 standaardafwijking in $Urengestudeerd_z$, stijgt de score op de wiskundetoets met 1,860 punten, ongeacht geslacht.

Of, als we dit alles samenbrengen in twee vergelijkingen:

$$\begin{aligned} \text{Wiskundescore}_{\text{meisjes}} &= 73,47 + 7,585 \cdot Iq_z + 1,860 \cdot Urengestudeerd_z \\ \text{Wiskundescore}_{\text{jongens}} &= 77,02 + 7,585 \cdot Iq_z + 1,860 \cdot Urengestudeerd_z \end{aligned}$$

De p-waarde onderaan de output geeft aan wat de p-waarde is voor ons model (p-value: $< 2.2e-16$). Deze p-waarde dienen we te interpreteren als de kans dat de nulhypothese geldig is. In dit geval luidt de nulhypothese dat ons model (bestaande uit drie onafhankelijke variabelen) in de populatie geen verschillen kan verklaren aangaande wiskundescores van leerlingen. De p-waarde is beduidend lager dan 0,05 waardoor we kunnen besluiten dat de kans wel heel klein is dat ons model in de populatie niet in staat is om vastgestelde verschillen te verklaren.

Bij Adjusted R-Squared lezen we af hoe goed ons model in z'n geheel is. Op basis van onze analyse kunnen we stellen dat de combinatie van Iq , $urengestudeerd$ en $geslacht$ ons in staat stellen om 73,28% van de variantie in wiskundescores te verklaren.

RESPONS 10.3.10

Indien we een reeks dummyvariabelen willen aanmaken om de variabele Iq categorisch te weerspiegelen, met daarbij gebruikmakend van de categorie “gemiddeld” als referentiecategorie, dan moeten we twee nieuwe dummyvariabelen aanmaken. Er zijn namelijk drie categorieën en we hebben één dummyvariabele minder nodig dan dat er categorieën zijn. We maken eerst een dummy die aanstaat voor een leerling die “laag” scoort op Iq categorisch. Vervolgens maken we een tweede dummy die aanstaat voor een leerling die “hoog” scoort op Iq categorisch.

```
> Wis2$Iq_laag <- (Wis2$Iq_categorisch == "laag") * 1
> Wis2$Iq_hoog <- (Wis2$Iq_categorisch == "hoog") * 1
```

Om te controleren of we goed te werk zijn gegaan maken we opnieuw frequentietabellen aan:

```
> table(Wis2$Iq_laag, Wis2$Iq_categorisch)
```

	laag	gemiddeld	hoog
0	0	28	27
1	25	0	0


```
> table(Wis2$Iq_hoog,Wis2$Iqcategorisch)
```

	laag	gemiddeld	hoog
0	25	28	0
1	0	0	27

De analyse zou er als volgt moeten uitzien:

```
> Model9<-lm(Wiskunde~Iq_laag+Iq_hoog+Urengestudeerd_z+Jongen,data=Wis2)
> summary(Model9)
```

Call:

```
lm(formula = Wiskunde ~ Iq_laag + Iq_hoog + Urengestudeerd_z +
    Jongen, data = Wis2)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.6190	-3.6785	-0.3321	3.7005	17.6879

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	72.0671	1.3532	53.256	< 2e-16 ***
Iq_laag	-6.4692	1.7539	-3.689	0.000425 ***
Iq_hoog	9.9393	1.8075	5.499	5.06e-07 ***
Urengestudeerd_z	2.2594	0.7374	3.064	0.003032 **
Jongen	3.7000	1.5334	2.413	0.018266 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.37 on 75 degrees of freedom

Multiple R-squared: 0.6421, Adjusted R-squared: 0.623

F-statistic: 33.64 on 4 and 75 DF, p-value: 4.625e-16

- Voor een meisje met een gemiddeld Iq (= de referentiecategorie), dat een gemiddeld aantal uren gestudeerd heeft verwachten we een score van 72,07 (het intercept).
- Scoren leerlingen uit de categorie met een hoog Iq hoger dan leerlingen met een gemiddeld Iq? Het antwoord luidt ja, ze scoren 9,94 punten hoger. Dit verschil is statistisch significant ($p < 0,001$).
- Kan je eveneens nagaan of leerlingen uit de categorie met een hoog Iq significant hoger scoren op wiskunde dan leerlingen uit de categorie met een laag Iq? Hierbij luidt het antwoord nee. Je kan dat niet rechtstreeks afleiden uit deze tabel (ook al kan je het wel vermoeden uiteraard). Als je dat per se wil toetsen kan je een extra dummyvariabele aanmaken die aanstaat voor Iq "gemiddeld" en vervolgens Iq_hoog of Iq_laag uit de regressieanalyse vervangen door Iq_gemiddeld. Zo krijg je een nieuwe referentiecategorie in de analyse.

Indien we een reeks dummyvariabelen willen aanmaken om de variabele Statuut te weerspiegelen, met daarbij gebruikmakend van de categorie “arbeider” als referentiecategorie, dan moeten we twee nieuwe dummyvariabelen aanmaken.

```
> Werk$Statuut_bediende<-(Werk$Statuut=="bediende")*1
> Werk$Statuut_kader<-(Werk$Statuut=="kader")*1
```

Om te controleren of we goed te werk zijn gegaan maken we opnieuw frequentietabellen aan:

```
> table(Werk$Statuut_bediende,Werk$Statuut)
```

	arbeider	bediende	kader
0	346	0	308
1	0	346	0

```
> table(Werk$Statuut_kader,Werk$Statuut)
```

	arbeider	bediende	kader
0	346	346	0
1	0	0	308

De analyse zou er als volgt moeten uitzien:

```
> Model10<-lm(Welbevindenz~Klimaatz+Statuut_bediende+Statuut_kader,
               data=Werk)
```

```
> summary(Model10)
```

Call:

```
lm(formula = Welbevindenz ~ Klimaatz + Statuut_bediende + Statuut_kader,
    data = Werk)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.51977	-0.49050	-0.00481	0.50087	2.21750

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.006039	0.039861	0.152	0.88
Klimaatz	0.636028	0.023481	27.087	< 2e-16 ***
Statuut_bediende	0.240358	0.056412	4.261	2.23e-05 ***
Statuut_kader	-0.289621	0.058071	-4.987	7.22e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7403 on 996 degrees of freedom

Multiple R-squared: 0.4537, Adjusted R-squared: 0.452

F-statistic: 275.7 on 3 and 996 DF, p-value: < 2.2e-16

Het model verklaart 45,2% van de variantie in welbevinden op het werk. Dit is bovendien significant we verwachten dus dat in de populatie deze onafhankelijken er ook toe doen in het voorspellen van welbevinden op het werk ($p < 0,05$).

Als we vervolgens kijken naar de richting van de effecten, zien we dat we voor een arbeider die de openheid van het bedrijfsklimaat gemiddeld scoort, een score van 0,006 op welbevinden verwachten. Dit is niet significant ($p > 0,05$), waardoor we voor zulk persoon een gemiddelde score op welbevinden verwachten in de populatie. Per standaardafwijking dat een respondent hoger scoort op klimaat, stijgt zijn welbevinden met 0,636 standaardafwijkingen. Dit effect verwachten we ook in de populatie terug te vinden ($p < 0,05$). Een bediende scoort gemiddeld 0,240 standaardafwijkingen hoger dan een arbeider op welbevinden op het werk. Een kaderlid daarentegen scoort gemiddeld 0,290 punten lager. Beide effecten verwachten we ook terug te vinden in de populatie ($p < 0,05$).

steekproef	Arbeider	Bediende	Kader
-1,5 op klimaat	$0,006 + (-1,5 * 0,636)$ = - 0,948	$0,006 + 0,240 + (-1,5 * 0,636)$ = - 0,708	$0,006 - 0,290 + (-1,5 * 0,636)$ = - 1,238
0,5 op klimaat	$0,006 + (0,5 * 0,636)$ = 0,324	$0,006 + 0,240 + 0,5 * 0,636$ = 0,564	$0,006 - 0,290 + (0,5 * 0,636)$ = 0,034

populatie	Arbeider	Bediende	Kader
-1,5 op klimaat	$0 + (-1,5 * 0,636)$ = - 0,954	$0 + 0,240 + (-1,5 * 0,636)$ = - 0,714	$0 - 0,290 + (-1,5 * 0,636)$ = - 1,244
0,5 op klimaat	$0 + (0,5 * 0,636)$ = 0,318	$0 + 0,240 + (0,5 * 0,636)$ = 0,558	$0 - 0,290 + (0,5 * 0,636)$ = 0,028

RESPONS 10.3.13

Wat het meest opvallende is in deze figuur is dat de best passende regressielijnen voor elke groep afzonderlijk niet parallel lopen. Hoe meer leerlingen beroep doen op het begrijpen van de leerstof, hoe kleiner het verschil tussen jongens en meisjes. Of omgekeerd, bij leerlingen die minder beroep doen op het begrijpen van de leerstof, speelt geslacht een grotere rol. De hellingsgraad voor de lijn van de meisjes is steiler dan die voor de jongens. De invloed van het begrijpen is dus sterker voor meisjes dan voor jongens.

RESPONS 10.3.15

	Jongens	Begrijpenz	Interactieterm
lln1	1	1	1
lln2	0	1	0
lln3	1	0	0
lln4	0	0	0

Stel dat we nu de volgende regressievergelijking schatten:

$$Y = \beta_1 + \beta_2 \text{ Jongens} + \beta_3 \text{ Begrijpenz} + \beta_4 \text{ Interactieterm}$$

- a) Voor welke van de vier leerlingen zou het intercept gelden als verwachte score?

Enkel voor lln4, want enkel deze leerling scoort nul op alle onafhankelijke variabelen in de regressieanalyse.

- b) Voor welke leerling(en) dienen we enkel β_2 toe te voegen aan het intercept om de verwachte score te schatten?

Enkel voor lln3. Deze leerling is de enige die één scoort op Jongen en nul op de overige variabelen.

- c) Voor welke leerling(en) dienen we enkel β_3 toe te voegen aan het intercept om de verwachte score te schatten?

Enkel voor lln2 aangezien dit de enige leerling is die één scoort op Begrijpenz en niet 1 scoort op Jongen.

- d) Welke beta's moeten we hanteren om een verwachte score te berekenen voor lln1?

Hiertoe dienen we rekening te houden met alle beta's. We vertrekken van het intercept, voegen er de β_2 aan toe omdat het gaat om een leerling met score 1 op Jongen. We voegen er ook β_3 aan toe omdat het gaat om een leerling met score 1 op Begrijpenz. En aangezien de interactievariabele ook aanstaat voor deze leerling dienen we bovendien β_4 mee in de berekening op te nemen.

RESPONS 10.3.17

steekproef	Meisje	Jongen
Gemiddelde score op begrijpen	$0,017 + 0 \cdot 0,457$ = 0,017	$0,017 + 0,170 + 0 \cdot 0,457 + 0 \cdot (-0,048) \cdot 1$ = 0,187
-1 op begrijpen	$0,017 + (-1) \cdot 0,457$ = - 0,440	$0,017 + 0,170 + (-1) \cdot 0,457 + (-1) \cdot (-0,048) \cdot 1$ = - 0,222
1,5 op begrijpen	$0,017 + 1,5 \cdot 0,457$ = 0,703	$0,017 + 0,170 + 1,5 \cdot 0,457 + 1,5 \cdot (-0,048) \cdot 1$ = 0,801

populatie	Meisje	Jongen
Gemiddelde score op begrijpen	$0 + 0 \cdot 0,457$ = 0	$0 + 0,170 + 0 \cdot 0,457 + 0 \cdot (-0,048) \cdot 1$ = 0,170
-1 op begrijpen	$0 + (-1) \cdot 0,457$ = - 0,457	$0 + 0,170 + (-1) \cdot 0,457 + (-1) \cdot (-0,048) \cdot 1$ = - 0,239
1,5 op begrijpen	$0 + 1,5 \cdot 0,457$ = 0,686	$0 + 0,170 + 1,5 \cdot 0,457 + 1,5 \cdot (-0,048) \cdot 1$ = 0,784

RESPONS 10.3.19

```
> Werk2<-na.omit(Werk[ c("Welbevindenz","Klimaatz","Statuut_bediende",
  "Statuut_kader","Statuut")])

> scatterplot(Werk2$Welbevindenz~Werk2$Klimaatz/Werk2$Statuut,
  smooth=FALSE)

> Model14<-lm(Welbevindenz~Klimaatz+Statuut_bediende+Statuut_kader,
  data=Werk2)

> Model15<-lm(Welbevindenz~Klimaatz + Statuut_bediende + Statuut_kader
  + Klimaatz*Statuut_bediende + Klimaatz*Statuut_kader,
  data=Werk2)
> summary(Model15)

Call:
lm(formula = Welbevindenz ~ Klimaatz + Statuut_bediende + Statuut_kader +
  Klimaatz * Statuut_bediende + Klimaatz * Statuut_kader, data = Werk2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.04557 -0.42962 -0.00806  0.42714  1.93297

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.02246    0.03373   0.666   0.506
Klimaatz        0.80537    0.03398  23.699 < 2e-16 ***
Statuut_bediende 0.19870    0.04765   4.170 3.31e-05 ***
Statuut_kader   -0.28730    0.04905  -5.858 6.37e-09 ***
Klimaatz:Statuut_bediende 0.20649    0.04829   4.276 2.09e-05 ***
Klimaatz:Statuut_kader  -0.72874    0.04853 -15.015 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6245 on 994 degrees of freedom
Multiple R-squared:  0.612,    Adjusted R-squared:  0.61
F-statistic: 313.5 on 5 and 994 DF,  p-value: < 2.2e-16

> anova(Model14, Model15)
Analysis of Variance Table

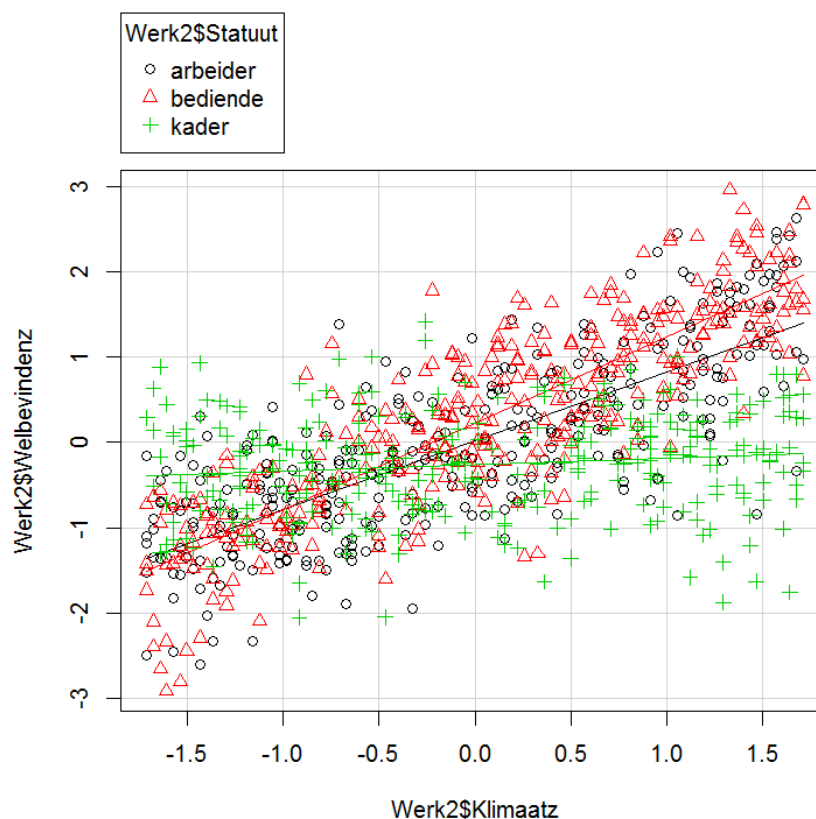
Model 1: Welbevindenz ~ Klimaatz + Statuut_bediende + Statuut_kader
Model 2: Welbevindenz ~ Klimaatz + Statuut_bediende + Statuut_kader +
  Klimaatz * Statuut_bediende + Klimaatz * Statuut_kader
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     996 545.78
2     994 387.65  2     158.13 202.74 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Het model met interactieeffect verklaart 61% van de variantie in welbevinden op het werk, wat significant is ($p < 2.2e-16$). Dit model is beter dan het model zonder interactie-term, dat 45,2% van de variantie in welbevinden op het werk verklaarde. Het toevoegen van deze interactieterm deed de RSS sterk dalen. Dit effect wordt als significant bestempeld ($p < 2.2e-16$). Het model mét interactieterm zal in de gehele populatie ook meer variantie verklaren dan het model zonder.

Wanneer we kijken naar de richting van de effecten, merken we dat alle coëfficiënten significant zijn. Alle effecten zijn dus door te trekken naar de populatie. Voor een arbeider die gemiddeld scoort op welbevinden verwachten we een score van 0,022 in de steekproef en een gemiddelde score in de populatie. Een bediende scoort bij een gemiddeld bedrijfsklimaat 0,199 hoger, terwijl een kaderlid gemiddeld 0,287 lager scoort.

Voor een arbeider verwachten we per standaardafwijking stijging in open bedrijfsklimaat een toename van 0,805 standaardafwijkingen in welbevinden op het werk. Voor een bediende is dit verband tussen open bedrijfsklimaat en welbevinden sterker, met name 1,011 ($=0,805+0,206$). Voor een kaderlid daarentegen is er nauwelijks een effect van open bedrijfsklimaat op welbevinden ($0,805-0,729=0,076$).

Volgende figuur illustreert dit. De regressielijn voor kaderleden is bijna een constante. De regressielijn voor bediende is bovendien steiler dan die voor arbeiders.



Figuur 10.3.19 Scatterplot voor effect van klimaat en statuut op welbevinden, met interactie-effect tussen klimaat en statuut

De verwachte score in de populatie op welbevinden op het werk voor een kaderlid dat 1,5 scoort op klimaat is -0,173 standaardafwijkingen.

$$\text{Verwachte score} = 0 - 0,287 + 1,5 \cdot 0,805 + 1,5 \cdot (-0,729) = -0,287 + 1,208 - 1,094 = -0,173$$

De uitwerking van deze opdracht ziet er als volgt uit::

```
> Model16<-lm(Wiskunde~Iq_laag+Iq_hoog+Urengestudeerd_z+Jongen
               +Jongen*Iq_laag,data=Wis2)
> summary(Model16)
```

Call:

```
lm(formula = Wiskunde ~ Iq_laag + Iq_hoog + Urengestudeerd_z +
    Jongen + Jongen * Iq_laag, data = Wis2)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-24.0018	-3.2983	0.3247	3.4613	15.1747

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	72.8215	1.3864	52.525	< 2e-16 ***
Iq_laag	-8.7537	2.0939	-4.181	7.89e-05 ***
Iq_hoog	10.6774	1.8171	5.876	1.12e-07 ***
Urengestudeerd_z	2.1058	0.7289	2.889	0.00507 **
Jongen	1.7389	1.8200	0.955	0.34247
Iq_laag:Jongen	6.1412	3.1978	1.920	0.05865 .

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.259 on 74 degrees of freedom

Multiple R-squared: 0.6591, Adjusted R-squared: 0.6361

F-statistic: 28.61 on 5 and 74 DF, p-value: 4.953e-16

- a) Is het effect van geslacht anders voor leerlingen met een laag Iq dan voor de andere leerlingen?

Jongen heeft niet langer een significant effect. Met andere woorden, volgens deze analyse scoren zowel meisjes als jongens nagenoeg hetzelfde voor wiskunde. De interactievariabele is daarentegen maar net op het nippertje niet significant. Gegeven de beperkte steekproef mogen we wat soepeler omgaan met de vuistregel van $p < 0,05$ en kunnen we, zij het met enige onzekerheid, concluderen dat er enkel bij leerlingen met een laag Iq een verschil te merken is tussen jongens en meisjes: het effect van jongen zijn is 6,141 punten sterker voor leerlingen met een laag Iq dan voor leerlingen met een gemiddeld Iq (=de referentiegroep).

- b) Bereken de verwachte score in de populatie voor jongens die behoren tot de categorie Iq hoog en voor jongens die behoren tot de categorie Iq laag, die een gemiddeld aantal uren hebben gestudeerd?

Voor Iq hoog kunnen we deze score berekenen:

$$72,822 + 10,677 = 83,499$$

Voor Iq laag kunnen we deze score berekenen:

$$72,822 - 8,754 + 6,141 = 70,209$$

We kunnen de logica die we eerder opbouwden rond de interactie tussen categorieën doortrekken. Dit veronderstelt het opnemen van een extra term die het product vormt tussen de dummyvariabele `Iq_laag` enerzijds en de variabele `Urengestudeerd_z` anderzijds. De parameterschattingen van het model met deze extra term erbij zien er zo uit:

```
> Model17<-lm(Wiskunde~Iq_laag+Iq_hoog+Urengestudeerd_z+Jongen
               + Jongen*Iq_laag+Urengestudeerd_z*Iq_laag,data=Wis2)
> summary(Model17)
```

Call:

```
lm(formula = Wiskunde ~ Iq_laag + Iq_hoog + Urengestudeerd_z +
    Jongen + Jongen * Iq_laag + Urengestudeerd_z * Iq_laag, data = Wis2)
```

Residuals:

Min	1Q	Median	3Q	Max
-23.6786	-3.6197	-0.0789	3.3008	14.5461

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	72.7292	1.3890	52.362	< 2e-16	***
Iq_laag	-8.2889	2.1422	-3.869	0.000235	***
Iq_hoog	10.8586	1.8252	5.949	8.59e-08	***
Urengestudeerd_z	1.5283	0.9224	1.657	0.101831	
Jongen	1.8212	1.8213	1.000	0.320637	
Iq_laag:Jongen	5.4606	3.2656	1.672	0.098774	.
Iq_laag:Urengestudeerd_z	1.5361	1.5045	1.021	0.310626	

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.257 on 73 degrees of freedom
 Multiple R-squared: 0.6639, Adjusted R-squared: 0.6363
 F-statistic: 24.03 on 6 and 73 DF, p-value: 1.701e-15

Uit deze tabel kunnen we aflezen dat de interactieterm niet significant afwijkt van nul. Bijgevolg kunnen we op basis van deze analyse de veronderstelling van onze collega niet onderschrijven.

Gehanteerde functies bij hoofdstuk 10

Functie	Doelstelling	Bron
<code>lm(y~x)</code>	Voert een regressieanalyse uit met y als afhankelijke variabele en x als onafhankelijke variabele.	R basispakket
<code>scale()</code>	Herschaalt een bestaande variabele in een variabele met z-scores als waarden.	R basispakket
<code>scatterplot(y~xkwantitatief/ xkwalitatief, smooth=FALSE)</code>	Stelt de resultaten van een dummyregressie met interactie-effect visueel voor. Indien je <code>regLine=FALSE</code> toevoegt, geeft het je een gewone maar met verschillende kleur en vorm van punten per groep van de categorische variabele.	Car
<code>summary()</code>	Door te verwijzen naar een object met daarin het resultaat van een <code>lm()</code> commando tussen de haakjes krijg je een samenvatting van de regressieanalyse: de parameterschattingen, de significantietoetsen voor de parameterschattingen, de R-kwadraat en de significantie van het regressiemodel zelf.	R basispakket