University of Antwerp

GKN - Contactmoment 1

Deel 1: Intro + herhaling

Sven De Maeyer & Bea Mertens

14/10/2021

Waarom *Gevorderde* Kwantitatieve Analyses?

Weten we dan nog niet genoeg?

Wat zijn statistische modellen?

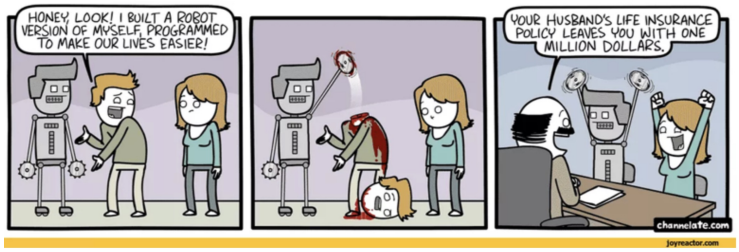
Statistische modellen = Golems… = robots…

Wat zijn statistische modellen?

Ze voeren uit wat je hen vraagt (zonder zelf na te denken) …

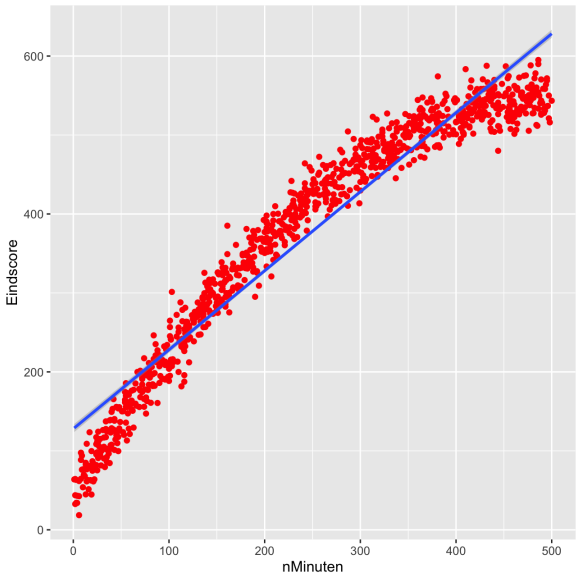
Wat zijn statistische modellen?

Voor bepaalde onderzoeksproblemen zijn bepaalde statistische modellen handig. Maar! Ze kunnen ook gevaarlijk zijn!



Wat zijn statistische modellen?

Niet alle statistische modellen zijn geschikt voor alle data ...



Wat als?

vb. studie van Silberzahn et al. 2018 (voor artikel zie BB)

> **Onderzoeksvraag:** Is het zo dat scheidsrechters in voetbal meer geneigd zijn gele en rode kaarten te geven aan spelers met een andere huidskleur?

29 onderzoeksteams gingen aan de slag met **dezelfde data**!

Resultaat:

- 69% v/d teams vond een significant effect, 31% niet...
- Analyse-aanpak van alle teams verschilde van elkaar

→ Conclusie afhankelijk van statistisch model dat werd gehanteerd!

suRplus van GKN

Meervoudige regressie = één (breed) statistisch model

Maar wat als je andere types van onderzoeksproblemen tegen komt?

Hier verruimen we je pallet aan beschikbare modellen:

- Structurele vergelijkingsmodellen (SEM)
- Multilevel modellen
- Logistische regressie modellen

Hoe gaan we de lessen aanpakken en welk leermateriaal is er?

Het openleerpakket

Op Blackboard vind je **OLP's** voor de verschillende onderdelen

Neem deze zoveel mogelijk **VOOR** de contactmomenten door

Maak een **script** aan dat je ook kan hanteren tijdens de lessen

De lessen zelf

Deel 1:

- Overlopen van de theorie | belangrijkste punten uit het OLP
- Voorbeelden waarbij jullie actief mee gaan denken

Deel 2:

- Ruimte voor oefenen (oefeningen met respons ter beschikking)
- Ruimte om aan de groepsopdracht te werken

→ Breng een laptop mee indien mogelijk en/of werk samen!

ZSO's

Per analysetechniek vind je op Blackboard een ZSO. Deze kan je facultatief maken.

Doel ZSO's:

→ Toepassing van de leerstof

→ Voorbereiding voor de groepsopdracht

Hoe gaan we evalueren?

Groepsopdracht (paper) + Mondeling examen

Groepsopdracht - Doel

- Aantonen dat jullie de analysetechnieken beheersen en kunnen uitvoeren in R
- De essentie van een analyse rapporteren en bespreken (cfr. nodig voor masterproef)
- Helder rapporteren en komen tot conclusies op basis van analyses

Groepsopdracht - Middel

(1) Een **paper** schrijven

Jullie krijgen een dataset uit het TIMSS 2019 onderzoek en bijhorende onderzoeksvragen

Zelfstandig de nodige analyses uitvoeren + rapporteren!

Groepsopdracht - Middel

(2) Ook een **'net' scRipt** opleveren

- Bevat enkel de definitieve commando's
- Geannoteerd
- Volgt analyses zoals gepresenteerd in de paper
- Moet door een andere onderzoeker kunnen worden uitgevoerd en tot dezelfde resultaten leiden (= reproduceerbaar)

Groepsopdracht - Praktisch

De opdracht is een groepsopdracht!

- Groepen van 3 studenten
- Vrij om zelf 2 medestudenten te kiezen
- Inschrijven in een groep via Blackboard (DEADLINE = zondag 17/10)
- Studenten zonder groep worden door ons aan elkaar of aan een groep met 2 studenten toegewezen

Groepsopdracht - Good practices en tips

- Mogelijkheid tot feedback via Comproved (info volgt)
- Werk samen!
- Get to know your data!
- 💡Vergeet statistiek A & B niet 💡

Groepsopdracht - Paper

Structuur van een wetenschappelijk artikel (bv. onderzoeksartikels in Pedagogische Studiën als inspiratie)

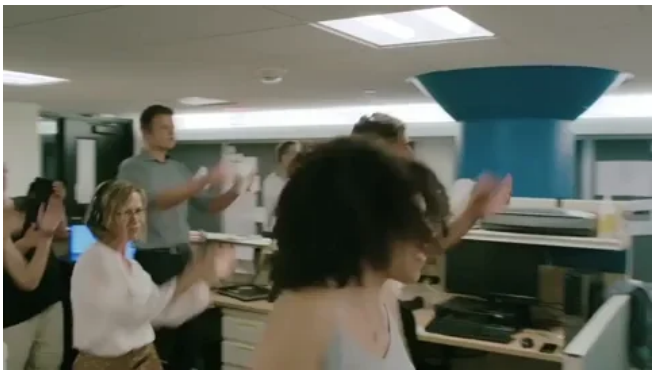
- 6 A4's
- Times New Roman, 11pt

Onderdelen:

- Titelblad
- Methodologiesectie
- Resultaten per onderzoeksvraag
- Bijdrage van de auteurs

Groepsopdracht - Wat brengt de opdracht op?

Veel plezieR



Maar vooral: **12 punten** te verdienen

Mondeling examen - Doel

Aantonen dat je inzicht hebt in de materie!

Dus we testen begrip en niet 'reproductieve kennis'

Mondeling examen - Vorm

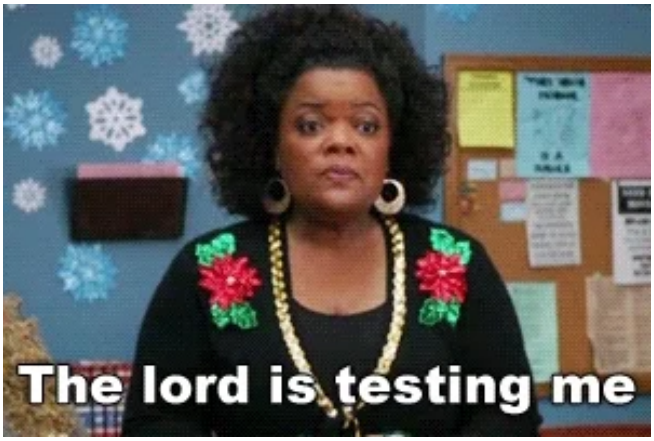
Tijdens het mondeling examen stellen we vragen gerelateerd aan de paper die je indiende

Op het moment zelf geen voorbereiding

Kom voorbereid naar het examen: je eigen paper is de leidraad om het mondeling examen af te leggen

Mondeling examen - Wat brengt het mondeling examen op?

Liefst niet te veel st**R**ess



En vooral: **8 punten** te verdienen

Herhaling statistiek B

Analysemodellen

Statistiek B:

van onderzoeksvraag naar analysemodel

Vertrekpunt:

schematische voorstelling

🏃 Analysemodellen - Oefening 1 🏃

Visualiseer onderstaande onderzoeksvragen (samen met je buur)

> *Context: onderzoek naar effectiviteit van online trainingstraject 'Machine Learning'*

- OV1: Wat is het effect van 'kijktijd' (aantal minuten kennisclip bekeken) op de eindscore die trainees behalen (z-score)?
- OV2: Is er een effect van 'kijktijd' op de eindscore die trainees behalen (z-score) ongeacht de voorkennis (gemeten adhv een parallelle toets, z-score)?
- OV3: Is het effect van 'kijktijd' op de eindscore die trainees behalen (z-score) afhankelijk van de voorkennis (z-score)

🏃 Analysemodellen - Oefening 2 🏃

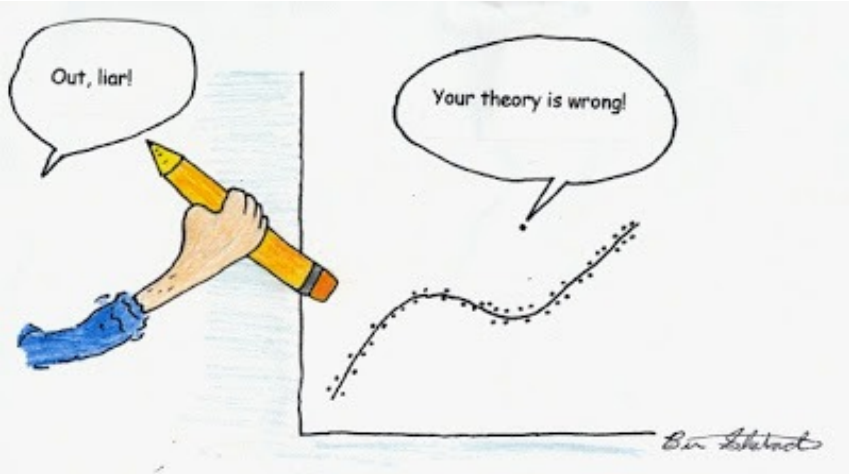
De training wordt gemodereerd door 20 verschillende mentoren en elke mentor begeleidt 15 trainees

- OV4: Leidt voorkennis (z-score) tot een andere leeractiviteit ('kijktijd') en bijgevolg tot andere eindresultaten?
- OV5: Is er een effect van de mentor die een trainee krijgt toegewezen op de eindscore die trainees behalen (z-score)?
- OV6: Is het effect van 'kijktijd' op de eindscore die trainees behalen (z-score) afhankelijk van de mentor die een trainee krijgt toegewezen?
- OV7: In hoeverre is de eindscore die trainees behalen (z-score) voorspellend voor het al dan niet behalen van het certificaat?

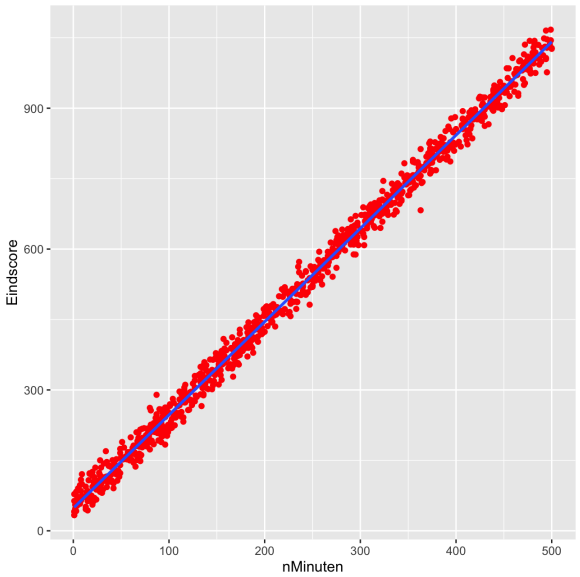
Regressieanalyse?

Wanneer pas je regressieanalyse toe?

Wat zijn de assumpties achter regressieanalyse?



Regressieanalyse: visueel



Regressieanalyse: formule

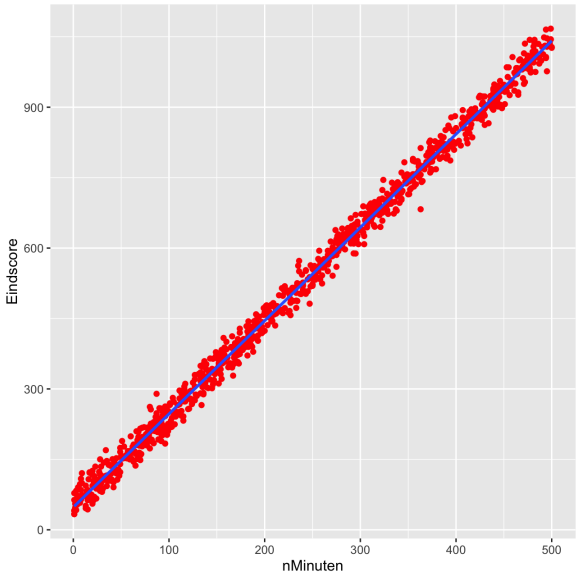
$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_{ij}$$

met:

- β_0 = het intercept (verwachte score voor y indien x_i gelijk is aan 0)
- β_1 = slope (verwachte stijging/daling in score y als x_i met 1 eenheid stijgt)

Regressieanalyse: formule

$$Eindscore_i = 45 + 2 * nMinuten + \epsilon_{ij}$$



- Verwachte eindscore als je 0 minuten kennisclips bekeek = ?
- Verwachte eindscore als je 2 minuten kennisclips bekeek = ?
- Verwachte eindscore als je 20 minuten kennisclips bekeek = ?

Een eenvoudig voorbeeld...

Een voorbeeldje met PIRLS 2016 data van Vlaanderen ...

> Heeft de mate waarin leerlingen vinden dat ze actief betrokken worden in de taallessen (Betrokkenheid, variabele '`ASBGERL`') een invloed op de score van begrijpend lezen (Leesvaardigheid, variabele '`ASRREA01`')?

Bivariate regressieanalyse (1)

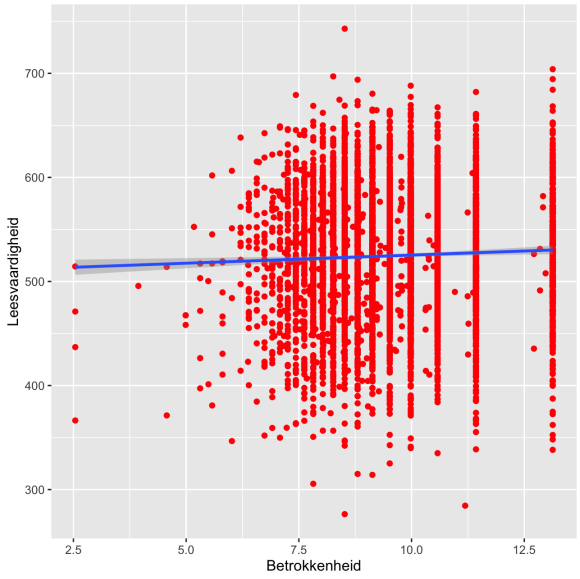
Eerst de dataset en de variabelen herbenoemen

```
load(here("C1", "Data", "Vlaanderen_1_2_3.RData"))
Vlaanderen                 <- Vlaanderen_1_2_3
Vlaanderen$Betrokkenheid   <- as.numeric(Vlaanderen$ASBGERL)
Vlaanderen$Leesvaardigheid <- as.numeric(Vlaanderen$ASRREA01)
```

Hier gebruik ik het pakket `here` wat maakt dat ik geen lange paden naar files op m'n pc dien in te voeren in het stukje `load()`. Zelf moet je vooral goed verwijzen naar de plaats waar je file staat. Of je moet leren werken met `here` natuurlijk http://jenrichmond.rbind.io/post/how-to-use-the-here-package/

Bivariate regressieanalyse (2)

Dan een grafiek maken



Bivariate regressieanalyse (3)

Schat het model

```
Model1 <- lm(Leesvaardigheid ~ Betrokkenheid,
             data = Vlaanderen)
summary(Model1)
```

```
##
## Call:
## lm(formula = Leesvaardigheid ~ Betrokkenheid, data = Vlaanderen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -246.654  -39.698    2.369   42.286  219.881
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   509.7203     5.0201 101.536  < 2e-16 ***
## Betrokkenheid   1.5633     0.5206   3.003  0.00269 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.3 on 5133 degrees of freedom
##   (63 observations deleted due to missingness)
## Multiple R-squared:  0.001754,    Adjusted R-squared:  0.001559
## F-statistic: 9.018 on 1 and 5133 DF,  p-value: 0.002686
```

Intercept?

Betrokkenheidheid?

Model?

Wat als er meerdere onafhankelijke variabelen zijn? (1)

Meerdere onafhankelijke variabelen…

- Variabele '`ASBHPLR`' = mate waarin ouders aangeven graag te lezen (Leesplezier)
- Variabele '`ASBGDDH`' = het aantal digitale devices in huis (Devices)

Meervoudige regressieanalyse (1)

Herbenoem de variabelen en schat het model

```
Vlaanderen$Leesplezier <- as.numeric(Vlaanderen$ASBHPLR)
Vlaanderen$Devices     <- as.numeric(Vlaanderen$ASBGDDH)

Model2 <- lm(
  Leesvaardigheid ~ Betrokkenheid + Leesplezier + Devices,
  data = Vlaanderen)
summary(Model2)
```

```
Call:
lm(formula = Leesvaardigheid ~ Betrokkenheid + Leesplezier +
    Devices, data = Vlaanderen)

Residuals:
     Min       1Q   Median       3Q      Max
-251.552  -37.444    2.554   40.204  218.280

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   437.9744     8.7168  50.245   <2e-16 ***
Betrokkenheid   1.1552     0.5393   2.142   0.0323 *
Leesplezier     6.7014     0.4501  14.888   <2e-16 ***
Devices         1.4722     0.5407   2.723   0.0065 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.39 on 4576 degrees of freedom
  (618 observations deleted due to missingness)
Multiple R-squared:  0.04961,    Adjusted R-squared:  0.04899
```

Welke variabele heeft het sterkste effect?

Meervoudige regressieanalyse (2)

Standaardiseer alle kwantitatieve variabelen

```
Vlaanderen$LeesvaardigheidZ <- scale(Vlaanderen$Leesvaardigheid)
Vlaanderen$BetrokkenheidZ   <- scale(Vlaanderen$Betrokkenheid)
Vlaanderen$LeesplezierZ     <- scale(Vlaanderen$Leesplezier)
Vlaanderen$DevicesZ         <- scale(Vlaanderen$Devices)
```

Meervoudige regressieanalyse (3)

Herschat het model

```
Call:
lm(formula = LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ +
    DevicesZ, data = Vlaanderen)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1640 -0.6198  0.0423  0.6655  3.6133

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.05197    0.01428   3.639 0.000277 ***
BetrokkenheidZ  0.03091    0.01443   2.142 0.032254 *
LeesplezierZ    0.21332    0.01433  14.888  < 2e-16 ***
DevicesZ        0.03900    0.01432   2.723 0.006499 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9666 on 4576 degrees of freedom
  (618 observations deleted due to missingness)
Multiple R-squared:  0.04961,    Adjusted R-squared:  0.04899
F-statistic: 79.62 on 3 and 4576 DF,  p-value: < 2.2e-16
```


Oh nee... een categorische onafhankelijke variabele?

Wat als we ook categorische variabelen als onafhankelijke variabelen willen toevoegen?

Stel we willen controleren voor de variabele 'Geslacht'

(variabele `ASBG01` waarbij 1 = Girl en 2 = Boy)

Regressieanalyse met een categorische onafh. variabele (1)

Herbenoem de variabele

Controleer of het een factor is

```
Vlaanderen$Geslacht <- Vlaanderen$ASBG01
is.factor(Vlaanderen$Geslacht)
```

```
## [1] FALSE
```

Wat zou er gebeuren als we de variabele `Geslacht` zo zouden toevoegen aan het model?

Regressieanalyse met een categorische onafh. variabele (2)

Schat het model

```
Call:
lm(formula = LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ +
    DevicesZ + Geslacht, data = Vlaanderen)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0877 -0.6239  0.0405  0.6652  3.5331

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.28304    0.04479   6.320 2.87e-10 ***
BetrokkenheidZ  0.02431    0.01444   1.684  0.09230 .
LeesplezierZ    0.21264    0.01428  14.886  < 2e-16 ***
DevicesZ        0.03796    0.01428   2.658  0.00788 **
Geslacht       -0.15560    0.02859  -5.442 5.55e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9636 on 4575 degrees of freedom
  (618 observations deleted due to missingness)
Multiple R-squared:  0.05572,    Adjusted R-squared:  0.0549
F-statistic: 67.49 on 4 and 4575 DF,  p-value: < 2.2e-16
```

Hoe moeten we hier het intercept interpreteren?

Regressieanalyse met een categorische onafh. variabele (3)

Maak een factor van de variabele `Geslacht`

```
Vlaanderen$GeslachtF <- as.factor(Vlaanderen$Geslacht)
```


Regressieanalyse met een categorische onafh. variabele (4)

Herschat het model

```
Model3 <- lm(LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ +
               DevicesZ + GeslachtF, data = Vlaanderen)
summary(Model3)
```

```
##
## Call:
## lm(formula = LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ +
##     DevicesZ + GeslachtF, data = Vlaanderen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0877 -0.6239  0.0405  0.6652  3.5331
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.12744    0.01988   6.412 1.59e-10 ***
## BetrokkenheidZ  0.02431    0.01444   1.684  0.09230 .
## LeesplezierZ    0.21264    0.01428  14.886  < 2e-16 ***
## DevicesZ        0.03796    0.01428   2.658  0.00788 **
## GeslachtF2     -0.15560    0.02859  -5.442 5.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9636 on 4575 degrees of freedom
##   (618 observations deleted due to missingness)
## Multiple R-squared:  0.05572,    Adjusted R-squared:  0.0549
## F-statistic: 67.49 on 4 and 4575 DF,  p-value: < 2.2e-16
```

Hoe kunnen we hier het intercept interpreteren?

Regressieanalyse met een categorische onafh. variabele (5)

Is hetzelfde model als met een dummy-variabele

```
Vlaanderen$Jongen <- (Vlaanderen$Geslacht == 2)*1
Model3b <- lm(LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ +
               DevicesZ + Jongen, data = Vlaanderen)
summary(Model3b)
```

```
##
## Call:
## lm(formula = LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ +
##     DevicesZ + Jongen, data = Vlaanderen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0877 -0.6239  0.0405  0.6652  3.5331
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.12744    0.01988   6.412 1.59e-10 ***
## BetrokkenheidZ  0.02431    0.01444   1.684  0.09230 .
## LeesplezierZ    0.21264    0.01428  14.886  < 2e-16 ***
## DevicesZ        0.03796    0.01428   2.658  0.00788 **
## Jongen         -0.15560    0.02859  -5.442 5.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9636 on 4575 degrees of freedom
##   (618 observations deleted due to missingness)
## Multiple R-squared:  0.05572,    Adjusted R-squared:  0.0549
## F-statistic: 67.49 on 4 and 4575 DF,  p-value: < 2.2e-16
```


Een laatste voorbeeld...

Hoe teken je alweer het model dat bij volgende onderzoeksvraag hoort?

> Is het verband tussen de mate dat ouders aangeven graag te lezen en de scores die leerlingen halen op begrijpend lezen anders voor jongens dan voor meisjes?

Hoe te modelleren in R? Wat impliceert dit?

Interactie-effecten (1)

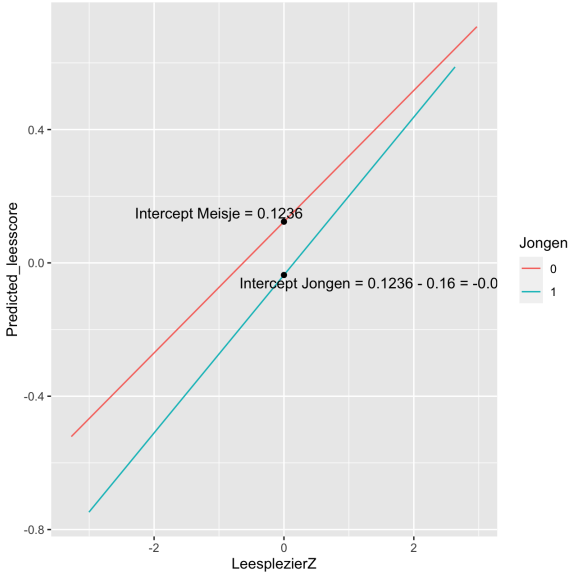
```
Model4 <- lm(LeesvaardigheidZ ~ LeesplezierZ +
               Jongen + LeesplezierZ*Jongen,
             data = Vlaanderen)
summary(Model4)
```

```
Call:
lm(formula = LeesvaardigheidZ ~ LeesplezierZ + Jongen + LeesplezierZ *
    Jongen, data = Vlaanderen)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0933 -0.6273  0.0416  0.6693  3.4991

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.12357    0.01978   6.248 4.53e-10 ***
LeesplezierZ        0.19689    0.01975   9.971  < 2e-16 ***
Jongen             -0.16006    0.02839  -5.639 1.81e-08 ***
LeesplezierZ:Jongen  0.03565    0.02841   1.255     0.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9658 on 4631 degrees of freedom
  (563 observations deleted due to missingness)
Multiple R-squared:  0.05363,    Adjusted R-squared:  0.05302
F-statistic: 87.48 on 3 and 4631 DF,  p-value: < 2.2e-16
```

Conclusie?

Interactie-effecten (2)

Het interactie-effect (ook al is het niet significant) even visueel



Interactie-effecten (3)

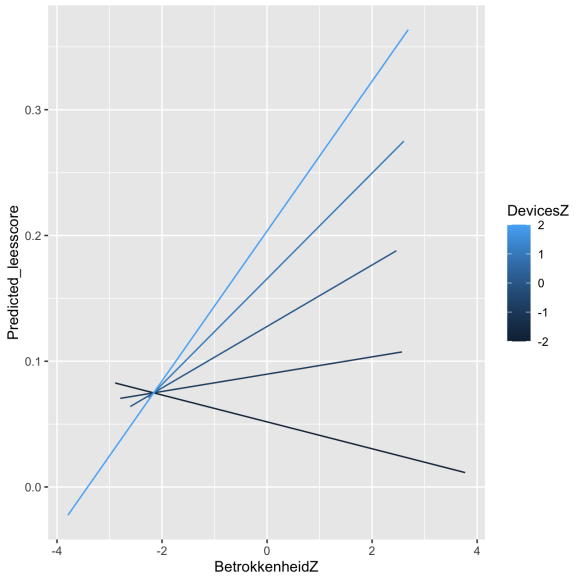
Interactie-effecten tussen twee KWANtitatieve verklarende variabelen?

```
Model5 <- lm(LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ + DevicesZ + GeslachtF + BetrokkenheidZ*DevicesZ , data = Vlaanderen)
summary(Model5)
```

```
Call:
lm(formula = LeesvaardigheidZ ~ BetrokkenheidZ + LeesplezierZ +
    DevicesZ + GeslachtF + BetrokkenheidZ * DevicesZ, data = Vlaanderen)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0869 -0.6228  0.0376  0.6654  3.5265

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.12765    0.01988   6.422 1.48e-10 ***
BetrokkenheidZ        0.02445    0.01444   1.694  0.09042 .
LeesplezierZ          0.21268    0.01428  14.889  < 2e-16 ***
DevicesZ              0.03794    0.01428   2.657  0.00791 **
GeslachtF2           -0.15560    0.02859  -5.442 5.54e-08 ***
BetrokkenheidZ:DevicesZ  0.01757    0.01484   1.184  0.23650
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9635 on 4574 degrees of freedom
  (618 observations deleted due to missingness)
Multiple R-squared:  0.05601,    Adjusted R-squared:  0.05498
F-statistic: 54.28 on 5 and 4574 DF,  p-value: < 2.2e-16
```

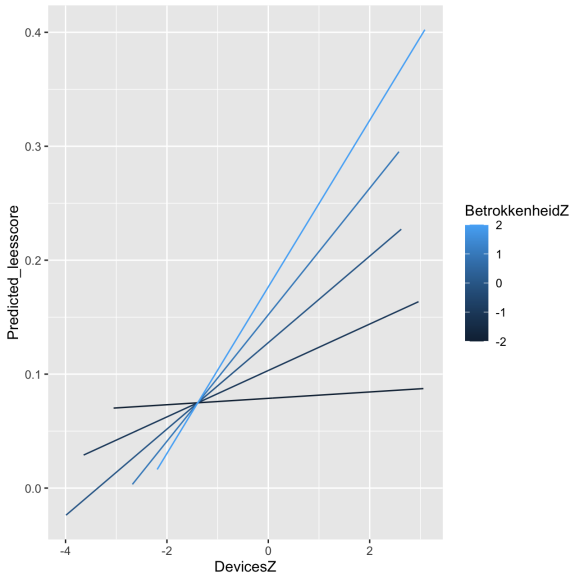

Interactie-effecten (4)

Het interactie-effect (ook al is het niet significant) even visueel



Interactie-effecten (4)

Hetzelfde interactie-effect even visueel (maar dan "andersom")



We zijn weer helemaal bij

Tijd voor een pauze