GKN 2020-2021: Contactmoment 6

Logistische regressie-analyse (2)

Tine van Daal & Sofie Vermeiren

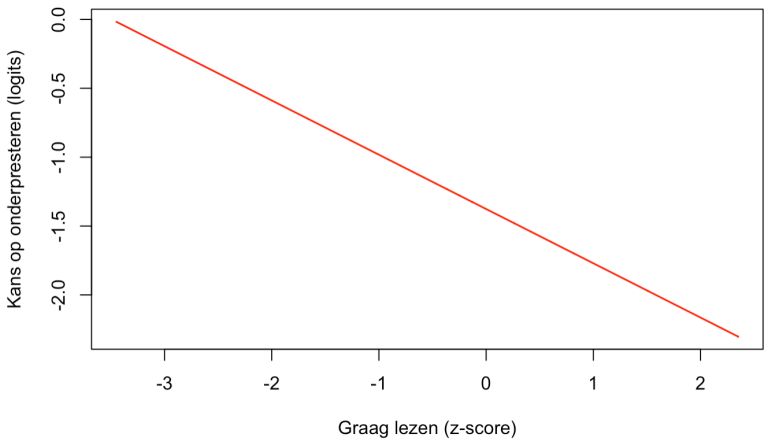
Opleidings- en onderwijswetenschappen

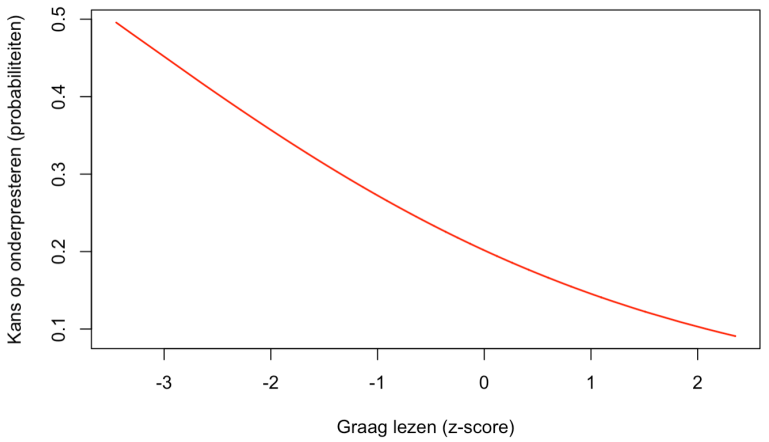
Recap

Laatste model vorig contactmoment

```
Call:
glm(formula = Onderpresteren ~ Gender + Ouders_GraagLezenZ, family = binomial(),
    data = Vlaanderen_1_2_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1701  -0.6869  -0.5991  -0.4531   2.2943

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.37641    0.05319  -25.88  < 2e-16 ***
GenderGirls       -0.25424    0.07611   -3.34 0.000836 ***
Ouders_GraagLezenZ -0.39405    0.03856  -10.22  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4496.7  on 4634  degrees of freedom
Residual deviance: 4377.4  on 4632  degrees of freedom
  (563 observations deleted due to missingness)
AIC: 4383.4

Number of Fisher Scoring iterations: 4
```

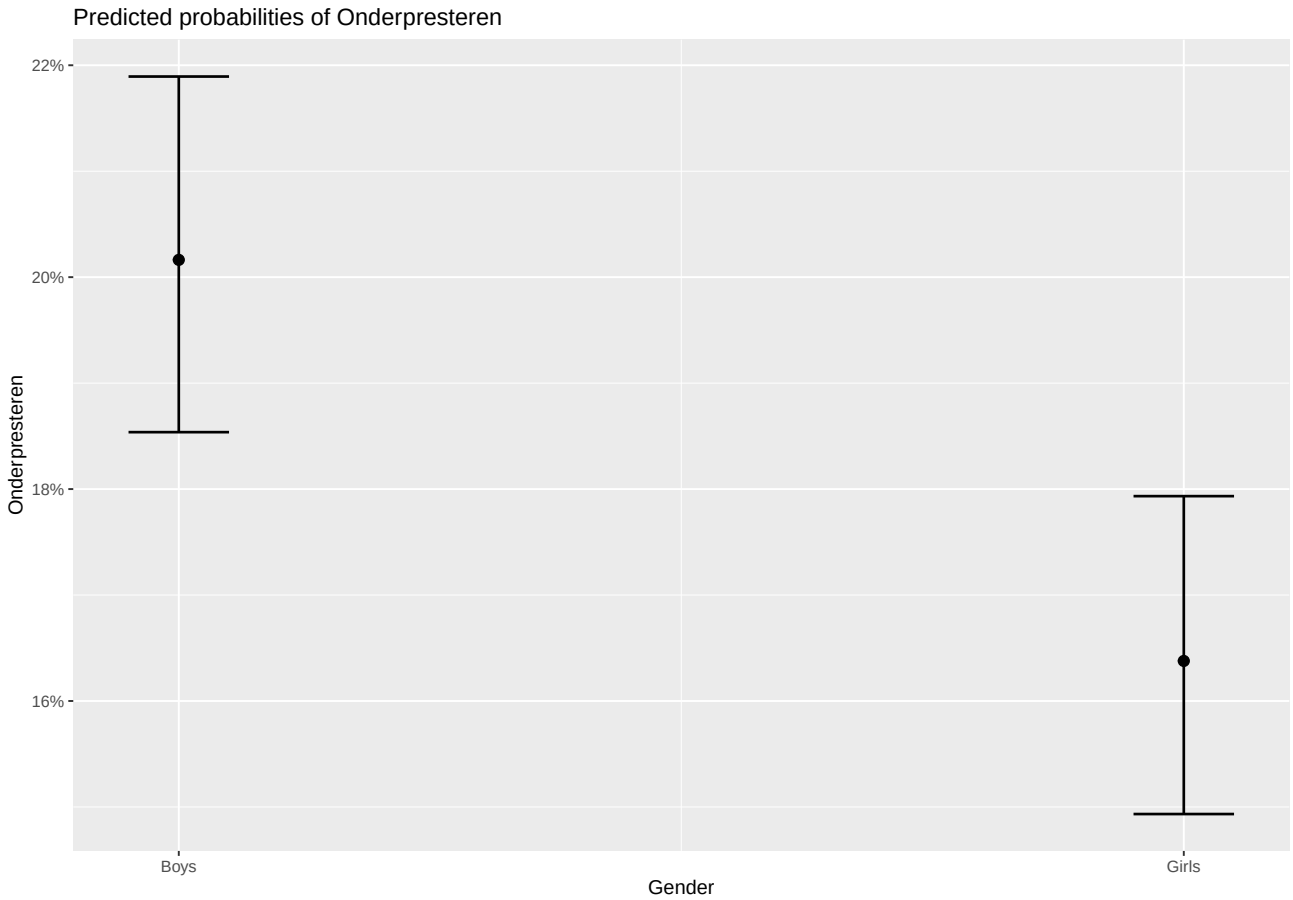
Effect van 'Ouders_GraagLezenZ'





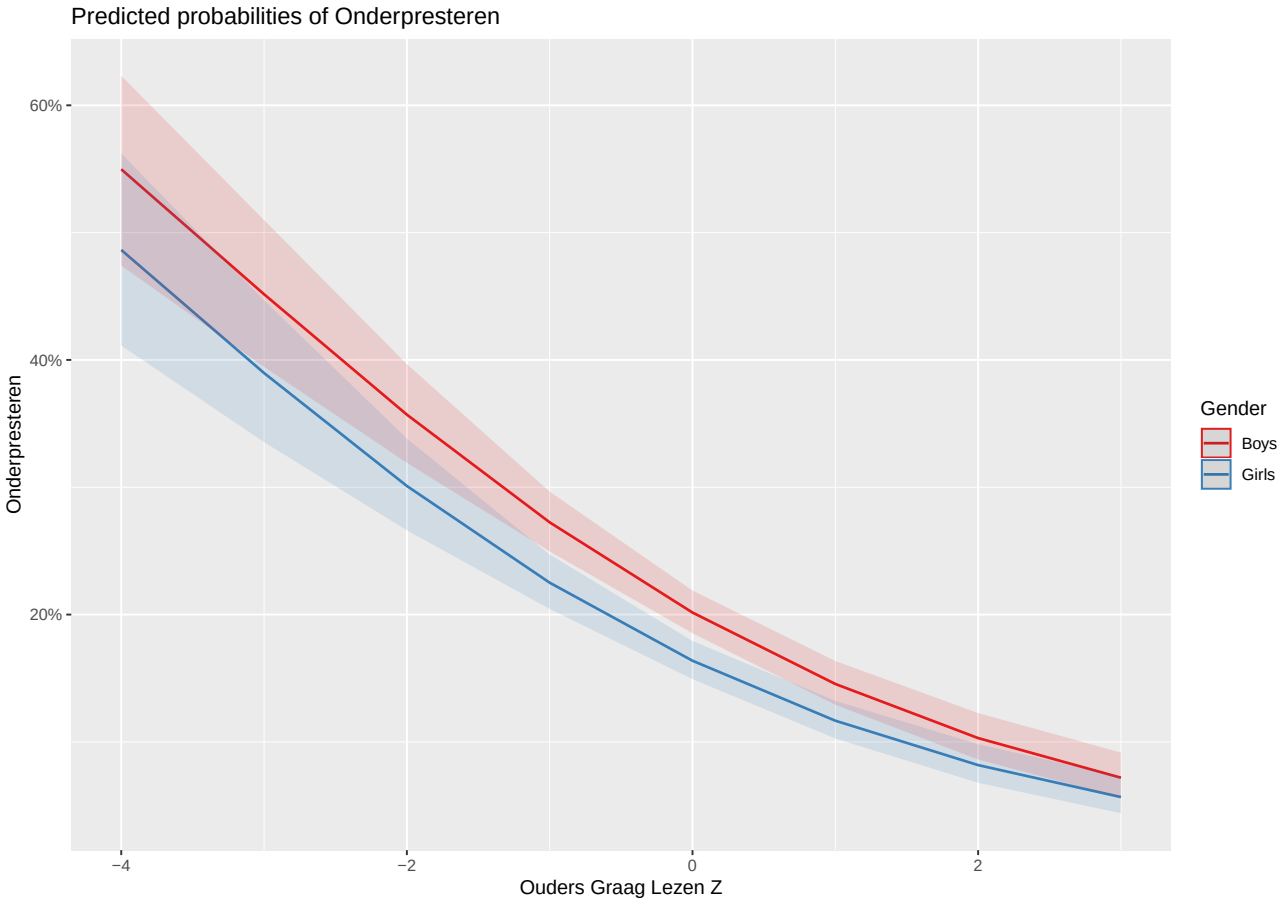
Effect van 'Gender'

```
library(sjPlot)
plot_model(M1_PIRLS, transform = NULL, type = "eff", terms = c("Gender"))
```



Beide effecten samen?

```
plot_model(M1_PIRLS, transform = NULL, type = "eff", terms = c("Ouders_GraagLezenZ", "Gender"))
```



Let's talk in Odds

Parameters als Odds interpreteren (1)

$$Logit(Onderpr. = 1) =$$

$$-1.376 + (-0.254 * GenderGirl) + (-0.394 * OudersLezenZ)$$

```
exp(-1.376)
```

```
[1] 0.2525869
```

Voor jongens wiens ouders gemiddeld graag lezen is de kans om te behoren tot de onderpresteerders 0.25 keer groter (of 1/0.25 = 4 keer kleiner) dan de kans om niet tot de onderpresteerders te behoren

Parameters als Odds interpreteren (2)

$$Logit(Onderpr. = 1) =$$

$$-1.376 + (-0.254 * GenderGirl) + (-0.394 * OudersLezenZ)$$

```
exp(-0.254)
```

```
[1] 0.7756918
```

```
1/exp(-0.254)
```

```
[1] 1.289172
```

Voor meisjes wiens ouders gemiddeld graag lezen is de kansverhouding om te behoren tot de onderpresteerders eerder dan tot de 'niet onderpresteerders' 0.776 keer groter (of 1/0.776 = 1.289 keer kleiner) dan voor jongens

0.776 is een **Odds Ratio**

Parameters als Odds interpreteren (3)

	Est.		*p*
	Logits	Odds	
Intercept	-1.376	0.253	< 0.001 ***
Meisje	-0.254	0.756	< 0.001 ***
Ouders_GraagLezenZ	-0.394	0.674	< 0.001 ***

=exp(logits)

Voorspelde Odds voor meisjes:

```
## Voorspelde Odds meisjes:
exp(-1.376) * exp(-0.254)
```

```
[1] 0.1959296
```

Odds zijn **multiplicatief**

Hoe goed zijn modellen?

Geen R^2

Bij gewone regressie-analyse hebben we een geschat residu:

$$Score_i = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \epsilon_{ij}$$

Gewone regressieanalyse: *Ordinary Least Squares* (OLS) schattingen

Schattingen die de afstand van de regressielijn met de residuen minimaliseert!

Maximum Likelihood

Bij logistische regressieanalyse hebben we geen geschat residu!

$$Logit(X = 1) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ...$$

Logistische regressieanalyse: Maximum Likelihood (ML) schattingen

Maximum Likelihood

Likelihood = functie van parameterwaarden (gegeven de data)!

Doel: die combinatie van parameterwaarden waarvoor de Likelihood zo hoog mogelijk is (=Maximaal)

Hoe?

- Likelihood wordt eerst log-getransformeerd
- Via 'afgeleiden' van Log-likelihood functie parameterwaarden waarvoor de log-likelihood maximaal is

$\rightarrow$ Voor een model krijgen we ook een Log-likelihood (LL) waarde (= indicatie van FIT!)

Modellen vergelijken

2 concurrerende modellen, welk model zou je weerhouden?

$\rightarrow$ Model met hoogste waarde voor LL!

Nulmodel als start

Nulmodel = model zonder voorspellers

```
M0_PIRLS <- glm(Onderpresteren ~ 1,
                data = Vlaanderen_1_2_3, family = binomial())
summary(M0_PIRLS)
```

```
Call:
glm(formula = Onderpresteren ~ 1, family = binomial(), data = Vlaanderen_1_2_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6725  -0.6725  -0.6725  -0.6725   1.7875

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.37145    0.03452  -39.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5236.4  on 5197  degrees of freedom
Residual deviance: 5236.4  on 5197  degrees of freedom
AIC: 5238.4

Number of Fisher Scoring iterations: 4
```

Vergelijking met Model1

```
logLik(M0_PIRLS)
```

```
'log Lik.' -2618.19 (df=1)
```

```
logLik(M1_PIRLS)
```

```
'log Lik.' -2188.681 (df=3)
```

In onderzoek wordt -2 keer LL gehanteerd (= **-2LL** of **Deviance**)

```
deviance(M0_PIRLS)
```

```
[1] 5236.379
```

```
deviance(M1_PIRLS)
```

```
[1] 4377.362
```

Via anova()

```
anova(M0_PIRLS, M1_PIRLS)
```

→ Error in anova.glmlist(c(list(object), dotargs), dispersion = dispersion, : models were not all fitted to the same size of dataset

Vergelijking modellen (invloed van 'missing values')

Modellen kunnen enkel vergeleken worden als ze geschat zijn op dezelfde dataset (en dus ook op evenveel observatie-eenheden)!

```
nrow(M0_PIRLS$model)
```

```
[1] 5198
```

```
nrow(M1_PIRLS$model)
```

```
[1] 4635
```

$\rightarrow$ Nulmodel herschatten op enkel de 4635 observaties om model te kunnen vergelijken

Vergelijking modellen

Missing values verwijderen:

```
Dat_analyse <- na.omit( Vlaanderen_1_2_3[ , c("Onderpresteren", "Gender", "Ouders_GraagLezenZ")] )
```

Modellen herschatten:

```
M0_PIRLS <- glm(Onderpresteren ~ 1,
                data = Dat_analyse, family = binomial())

M1_PIRLS <- glm(Onderpresteren ~ Gender + Ouders_GraagLezenZ,
                data = Dat_analyse, family = binomial())
```

Modellen vergelijken:

```
anova(M0_PIRLS , M1_PIRLS, test="Chi")
```

```
Analysis of Deviance Table

Model 1: Onderpresteren ~ 1
Model 2: Onderpresteren ~ Gender + Ouders_GraagLezenZ
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      4634     4496.7
2      4632     4377.4  2   119.32 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vergelijking modellen

Stappenplan:

- Nadenken over welke modellen je gaat schatten (gegeven je OV)
- Data-object maken zonder missings voor alle variabelen `na.omit()`
- Alternatieve modellen schatten op aangemaakt data-object
- Modellen vergelijken
- Beste model weerhouden en **herschatten op je originele dataset**
- Interpretatie (Nadenken over tabellen en figuren)