Logistische regressie analyse: een handleiding

Inge Sieben[1]
Liesbeth Linssen

Inhoudsopgave:

Wat is logistische regressie analyse?

Wanneer je de invloed wilt nagaan van een of meerdere onafhankelijke variabelen (X) op een afhankelijke variabele (Y), dan kom je al snel bij lineaire regressie analyse uit. Zo'n regressie model gaat er van uit dat de afhankelijke variabele continue van aard is, dus gemeten is op interval- of rationiveau. Het komt echter regelmatig voor dat de afhankelijke variabele van een ander meetniveau is, bijvoorbeeld dat er sprake is van een nominale variabele met slechts enkele categorieën. Lineaire regressie-analyse is dan niet mogelijk. Om toch de invloed van allerlei onafhankelijke variabelen op een nominale variabele na te kunnen gaan, zijn er verschillende analyse-technieken ontwikkeld. Een uitgebreide bespreking van deze technieken vind je in Lammers e.a. (2007). De techniek die het meest bij lineaire regressie analyse aansluit, is *logistische regressie analyse*. Logistische regressie analyse is geschikt voor een afhankelijke variabele die dichotoom van aard is: er zijn maar twee categorieën.

Hoe ziet het logistische model eruit? Dit laat zich het makkelijkst vertellen aan de hand van een voorbeeld. Stel dat we willen nagaan welke kenmerken van invloed zijn op het wel of niet gaan stemmen bij Tweede Kamer verkiezingen. We zijn dus geïnteresseerd in de voorspelling (door onafhankelijke variabelen) van de kans dat een persoon in de categorie 'wel gaan stemmen' of in de categorie 'niet gaan stemmen' valt. Een 'gewone' lineaire regressie analyse zal over het algemeen wel de juiste richting van de b-coëfficiënten opleveren. Maar de schatting is niet helemaal correct, omdat enkele belangrijke regressie assumpties geschonden worden, zoals de normaliteitsassumptie en de assumptie van homoscedasticiteit. Het grootste probleem is evenwel dat de door lineaire regressie voorspelde kansen groter kunnen zijn dan 1 en kleiner dan 0. Dergelijke kansen zijn niet te interpreteren. Het is daarom aan te raden om logistische regressie analyse te gebruiken wanneer je te maken hebt met een dichotome afhankelijke variabele.

Zoals gezegd gaat het logistische model uit van kansen, of beter gezegd van kansverhoudingen: odds. De odds in het voorbeeld is de kans om wel te gaan stemmen (p_{wel}) gedeeld door de kans om niet te gaan stemmen(p_{niet}). Een odds heeft een bereik van 0 (de kans om te gaan stemmen is nul) tot oneindig (de kans om te gaan stemmen is één). Omdat we liever met een variabele

[1] Met dank aan Paul de Graaf, John Hendrickx, Gerbert Kraaykamp en Jan Lammers.

rekenen die loopt van min oneindig naar plus oneindig, wordt de natuurlijke logaritme[2] van de odds genomen. Deze wordt de log odds of logit genoemd. Kans, odds en logit zijn dus eigenlijk drie manieren om hetzelfde te zeggen. Als we de onafhankelijke variabelen X_1, X_2 enz. noemen, dan ziet het logistische model er in formulevorm als volgt uit:

$$\ln \frac{p_{wel}}{p_{niet}} = a + b_1X_1 + b_2X_2 + \ldots.$$

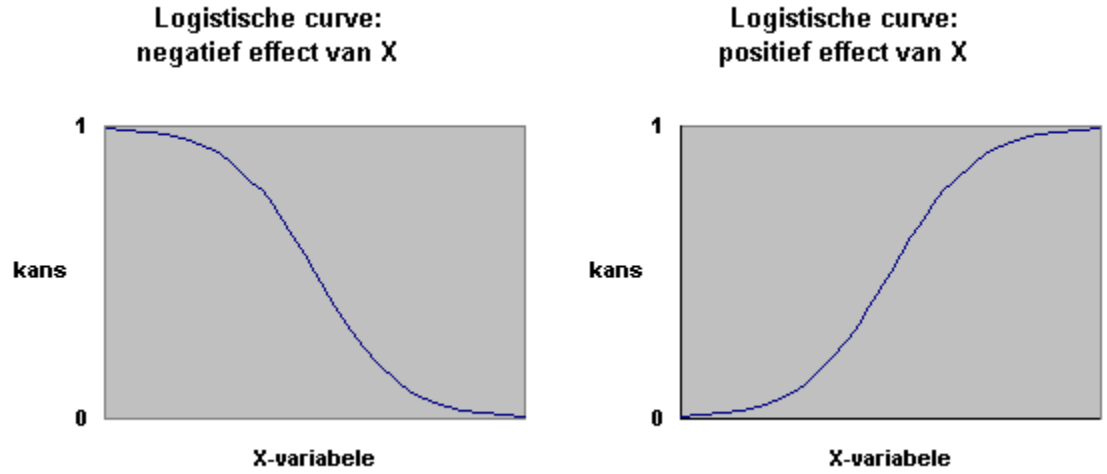
Dit model lijkt sterk op een gewoon regressie model: a is het intercept, b_1 is de parameter die het effect van X_1 aangeeft, b_2 de parameter die het effect van X_2 aangeeft enz. Hoe deze parameters geïnterpreteerd moeten worden, wordt later besproken. We kunnen het logistische model ook omzetten in een kans model. De kans dat iemand wel gaat stemmen bij Tweede Kamer verkiezingen is:

$$p_{wel} = \frac{e^{(a+b_1X_1+b_2X_2+\ldots)}}{e^{(a+b_1X_1+b_2X_2+\ldots)}+1}$$

En de kans dat iemand niet gaat stemmen bij Tweede Kamer verkiezingen is:

$$p_{niet} = \frac{1}{e^{(a+b_1X_1+b_2X_2+\ldots)}+1}$$

Aan deze formules is af te lezen dat de kansen p_{wel} en p_{niet} bij elkaar opgeteld gelijk zijn aan één. Verder is te zien dat de kansen p_{wel} en p_{niet} afhankelijk zijn van de variabelen X_1, X_2 enz., maar dat deze afhankelijkheid niet lineair is. Een logistische regressielijn ziet er dus niet als een rechte lijn uit, maar als een S-vormige curve. Hieronder zijn twee logistische curve getekend: links voor een negatief effect van de onafhankelijke variabele X, rechts voor een positief effect van X. Deze curves beschrijven het niet-lineaire verband tussen de kans en de onafhankelijke variabele X. We zien dat - bij een positief effect van X - de kans bij lage waarden van X niet zo veel stijgt, daarna sterk toeneemt, om vervolgens bij hoge waarden van X weer minder snel te stijgen. Oftewel: het effect van de variabele X op de kans dat Y voorkomt is het grootst bij de middenwaarden van X.



[2] De natuurlijke logaritme is de logaritme met als grondgetal e (e=2,71828…): 'elog' of 'ln' .

Hoe stuur je logistische regressie analyse in SPSS aan?
We zullen nu de invloed van geslacht, leeftijd, politieke voorkeur en opleiding nagaan op de kans om wel of niet te gaan stemmen bij de Tweede Kamer verkiezingen. De gegevens komen uit de Familie-enquete Nederlandse bevolking 1992-1993; voor onze analyse zijn 887 respondenten beschikbaar.

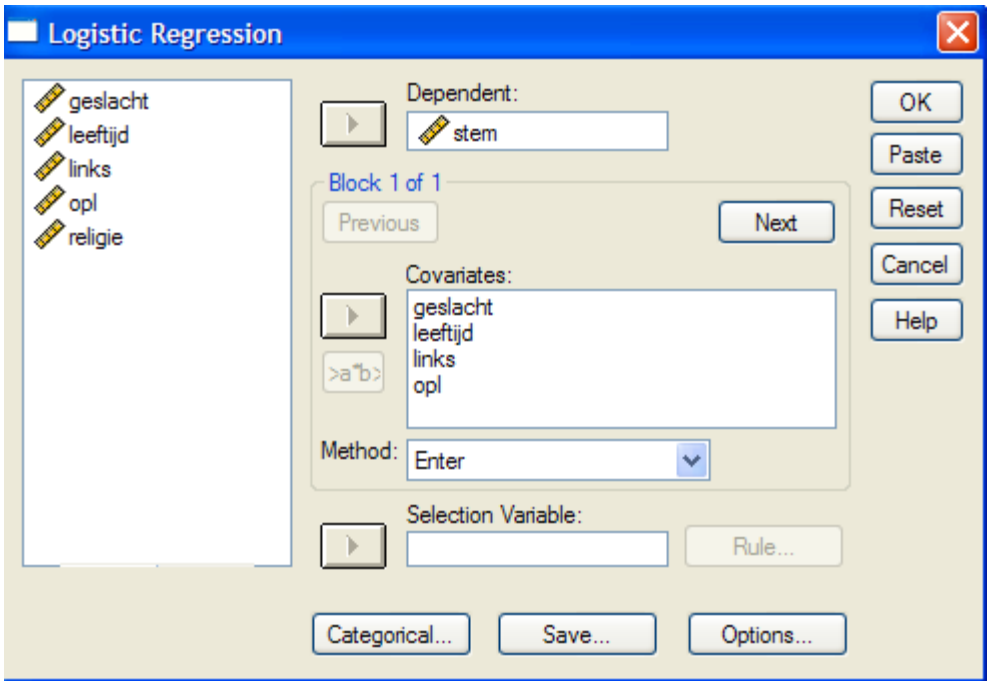
De variabelen zien er als volgt uit:
stem: gaan stemmen bij Tweede Kamer verkiezingen (0=’niet’; 1=’wel’)[3]
geslacht: geslacht van de respondent (0='man'; 1='vrouw')
leeftijd: leeftijd van de respondent (19 tot en met 75 jaar)
links: links/rechts zelfplaatsing (0='rechts'; 1='links')
opl: hoogst voltooide opleiding (1='minder dan lagere school' tot en met 10= 'postacademisch onderwijs')

In onderstaand kader kun je zien hoe deze variabelen verdeeld zijn:

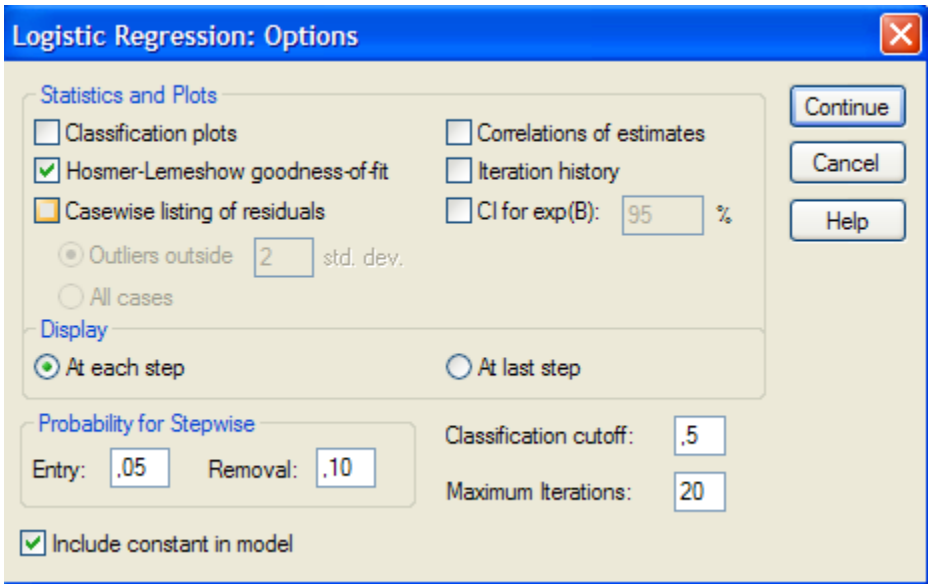
stem	niet: 87 (9.8%); wel: 800 (90.2%)
geslacht	man: 442 (49.8%); vrouw: 445 (50.2%)
leeftijd	gemiddelde: 40.8; standaard afwijking: 11.4
links	rechts: 392 (44.2%); links: 495 (55.8%)
opl	gemiddelde: 5.2; standaard afwijking: 2.3

In SPSS kan logistische regressie analyse op twee manieren aangestuurd worden: via de menu's en via de syntax. Kijken we eerst naar de menu's. We vinden logistische regressie analyse via: '**Analyze**' → '**Regression**' → '**Binary Logistic**'. Dit laatste menu ziet er als volgt uit:

[3] Logistische regressie analyse vereist dat de afhankelijke variabele uit twee categorieën bestaat. Hercodeer dus indien nodig je afhankelijke variabele. SPSS geeft een foutmelding wanneer je probeert om een logistische regressie analyse uit te voeren met een afhankelijke variabele die meer dan twee categorieën heeft. Bij een afhankelijke variabele met andere categoriewaarden dan 0 en 1 kent het programma 'intern' - dus alleen om mee te rekenen - de waarden 0 en 1 aan de categorieën toe. Het is echter aan te raden om dit niet aan SPSS over te laten, maar de hercodering zelf uit te voeren. Dan weet je zeker dat de voor jou meest interessante categorie (bv. wel gaan stemmen) de waarde 1 krijgt.



Hierin kunnen bij '**Dependent**' de afhankelijke variabele en bij '**Covariates**' de onafhankelijke variabelen worden opgegeven. Als '**Method**' wordt er standaard '**Enter**' opgegeven. Onder '**Options**' staan een aantal extra's. Voor ons is de Hosmer and Lemeshow Goodness-of-Fit Test het meest interessant. Hiermee kan worden nagegaan of het model goed bij de data past. Bij de bespreking van de output komen we hier op terug. We vinken deze optie aan:



De overige opties worden hier niet besproken; ze zijn terug te vinden in Norušis (2005) en onder HELP → Help Topics → Algorithms → Regression Models Option → Logistic Regression → Logistic Regression Options.

Dezelfde aansturing via de syntax ziet er als volgt uit:

```
LOGISTIC REGRESSION VARIABLES  stem
  /METHOD = ENTER geslacht leeftijd links opl
  /PRINT = GOODFIT.
```

Achter het commando **LOGISTIC REGRESSION** komt de afhankelijke variabele (**VAR=stem)**. Als **METHOD** geven we **ENTER** op, gevolgd door alle onafhankelijke variabelen. De Hosmer en Lemeshow Goodness-of-Fit Test verkrijgen we door het commando **PRINT=GOODFIT**.
Deze aansturing levert de volgende output op:

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	37,308	4	,000
	Block	37,308	4	,000
	Model	37,308	4	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	531,883(a)	,041	,087

a Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2,631	8	,955

Contingency Table for Hosmer and Lemeshow Test

		stem stemmen tweede kamer verkiezingen = 0 niet stemmen		stem stemmen tweede kamer verkiezingen = 1 wel stemmen		
		Observed	Expected	Observed	Expected	Total
Step 1	1	20	20,193	69	68,807	89
	2	14	14,722	75	74,278	89
	3	15	12,041	74	76,959	89
	4	10	9,992	79	79,008	89

	5	7	8,330	83	81,670	90
	6	7	6,885	82	82,115	89
	7	4	5,585	85	83,415	89
	8	4	4,376	85	84,624	89
	9	3	3,163	87	86,837	90
	10	3	1,713	81	82,287	84

Classification Table(a)

			Predicted		
			stemmen tweede kamer verkiezingen		
	Observed		0 niet stemmen	1 wel stemmen	Percentage Correct
Step 1	stemmen tweede kamer verkiezingen	0 niet stemmen	0	87	,0
		1 wel stemmen	0	800	100,0
	Overall Percentage				90,2

a The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	geslacht	-,323	,234	1,899	1	,168	,724
	leeftijd	,033	,011	8,846	1	,003	1,033
	links	-,542	,247	4,827	1	,028	,581
	opl	,265	,059	20,354	1	,000	1,303
	Constant	,198	,595	,111	1	,739	1,220

a Variable(s) entered on step 1: geslacht, leeftijd, links, opl.

Interpretatie van de fitmaten

Eerst kijken we of het model dat we geschat hebben goed bij de data past. De meest belangrijke toets hiervoor is de Chi^2-toets (**Chi-Square**), die in de output te vinden is onder de **Omnibus Tests of Model Coefficients**. Deze Chi^2-toets vergelijkt de aannemelijkheidratio van het geschatte model (**-2 Log Likelihood**, hier gelijk aan 531,883) met de aannemelijkheidsratio van een model met alleen maar een constante. Het verschil tussen deze twee aannemelijkheidsratio's is de **Chi-square** (hier gelijk aan 37,308). Het aantal vrijheidsgraden bij deze Chi2 is 4 (zie kolom **df**), we hebben immers te maken met vier variabelen. Een Chi2 van 37,308 is significant bij 4 vrijheidsgraden. Hetgeen betekent dat ons model met de variabelen *geslacht*, *leeftijd*, *links* en *opl* beter bij de data past dan een model zonder deze variabelen.

Logistische regressie analyse geeft geen proportie verklaarde variantie (R^2), zoals die voor interval of ratio variabelen in een lineair model gedefinieerd is. Wel bestaan er verschillende pseudo R^2-maten, die vergelijkbaar zijn met de R^2 uit lineaire regressie analyse. De SPSS output geeft twee van zulke maten. De Cox & Snell R^2 (hier gelijk aan 0,041) wordt niet zo vaak gebruikt, omdat deze maat nooit de waarde één kan bereiken. De R^2 van Nagelkerke (hier gelijk aan 0,087) kan dit wel. Een andere pseudo R^2-maat is de McFadden R^2. Deze wordt niet in de output gegeven, maar is eenvoudig te berekenen door de aannemelijkheidsratio van het model

zonder onafhankelijke variabelen (G0) af te trekken van de aannemelijkheidsratio van het model met onafhankelijke variabelen (Gm), en dit te delen door de eerste aannemelijkheidsratio (G0). In ons voorbeeld is dat dus:

$$\text{McFadden } R^2 = \frac{G_0 - G_m}{G_0} = \frac{37,308}{531,883 + 37,308} = 0,066$$

Opgemerkt moet worden dat alle drie pseudo R^2-maten over het algemeen lage waarden aannemen. Er is nog geen consensus over welke maat de 'beste' is.

Een andere manier om de fit van het model te bepalen is door te kijken naar **de Hosmer en Lemeshow Goodness-of-Fit Test** (Hosmer & Lemeshow, 1989). Daartoe worden de individuen geordend naar oplopende voorspelde slaagkans en aan de hand daarvan in 10 even grote groepjes ingedeeld. Vervolgens wordt er voor elk van die 10 groepjes gekeken hoeveel individuen er een 0 en 1 hebben gescoord en hoeveel je er op basis van het model zou verwachten. De verschillen worden met een Chi-2 toets getoetst[4]. In ons voorbeeld zijn deze verschillen niet significant (**Significance** is 0,955), zodat we kunnen concluderen dat het model goed bij de data past. De Hosmer en Lemeshow test moet voorzichtig gebruikt worden bij kleine steekproeven. De test zal in zo'n geval meestal aangeven dat het model past, ook wanneer dit niet het geval is. Bij heel grote steekproeven is het omgekeerde het geval: de test geeft dan vaak aan dat er significante verschillen zijn, terwijl het model toch goed bij de data past.

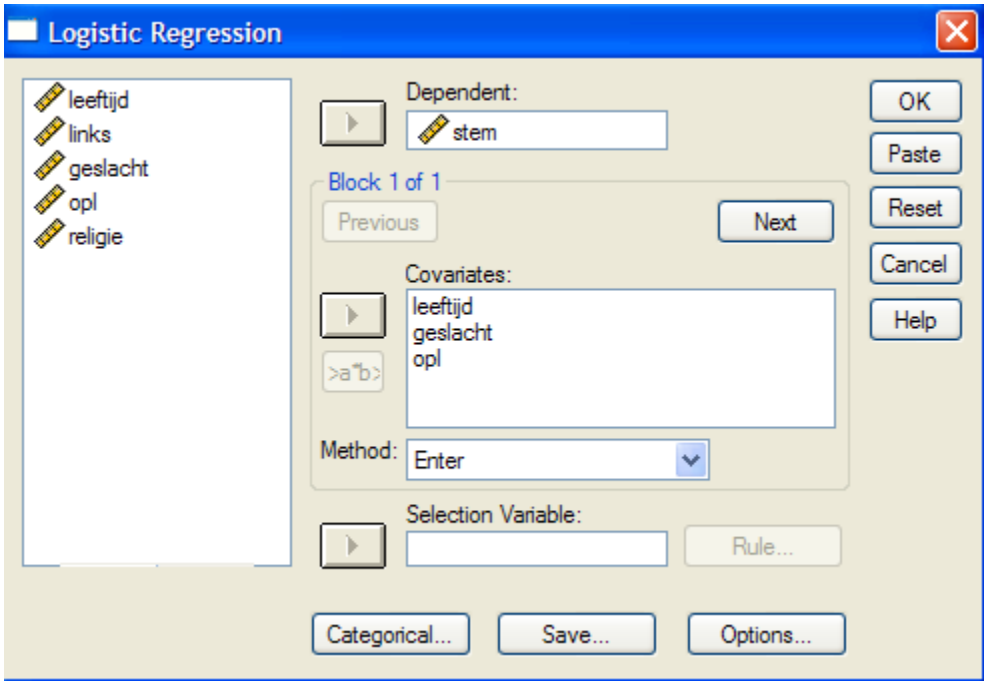
In de output treffen we ook nog een **Classification Table** aan, maar over het algemeen wordt deze tabel niet erg waardevol gevonden. Een bespreking van de classificatie-tabel en van alle fit-maten is te vinden in Lammers e.a. (2007). Overigens: het feit dat er sprake is van een goed passend model betekent niet dat er geen beter passende modellen zijn!

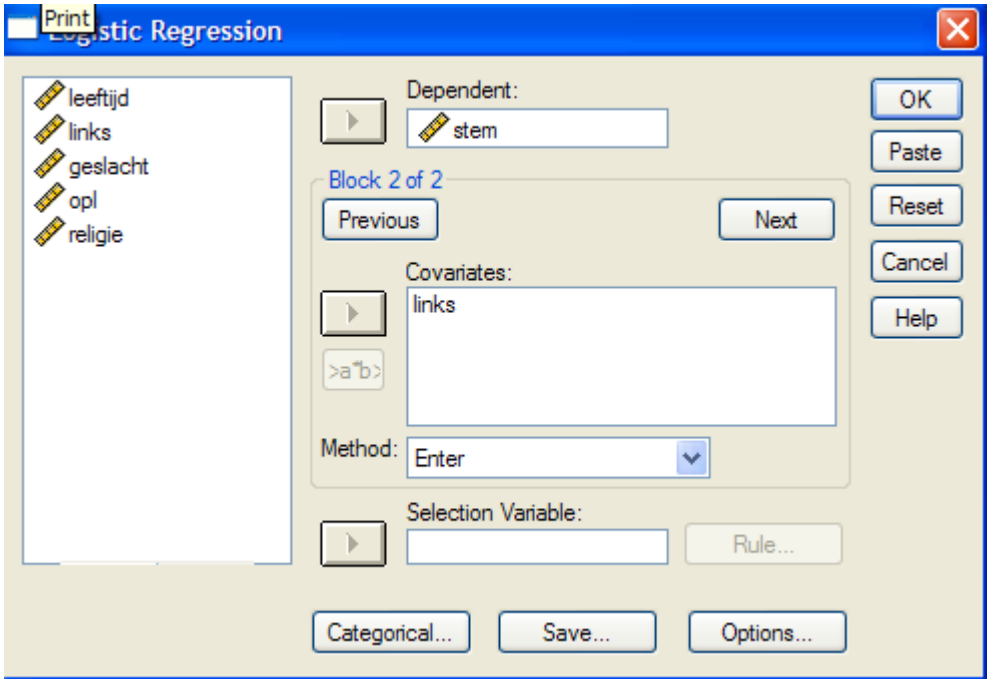
Interpretatie van de effecten

Nu we de fit van het model bepaald hebben, gaan we over naar de effecten van de onafhankelijke variabelen, die als laatste in de SPSS output verschijnen. Hier gaat het ons tenslotte om! Laten we eerst kijken welke variabelen een significante invloed hebben op de kans om wel versus niet te gaan stemmen bij Tweede Kamer verkiezingen. Dit wordt getest met behulp van de Wald-statistic (de waarden hiervan staan in de kolom **Wald**); deze is gelijk aan het kwadraat van (**B/S.E.**). In de kolom **Sig** kun je zien welke effecten significant zijn. In ons voorbeeld blijken *leeftijd*, *links* en *opleiding* een significante invloed te hebben op de kans om wel versus niet te gaan stemmen; het effect van *geslacht* is niet significant.

Er zit wel een nadeel aan de Wald-statistic: als de waarde van de regressie-coëfficiënt **B** groot is, dan kan het zo zijn dat je op basis van de Wald-statistic besluit dat er geen significant effect is, terwijl dat er eigenlijk wel is. Daarom de volgende tip: als het effect van een variabele net niet significant is, doe dan een logistische analyse met en zonder deze variabele en vergelijk de Chi^2 van de twee modellen. Wanneer we in ons voorbeeld zouden twijfelen over de significantie van het effect van de variabele *links*, vraag dan de Chi^2 op van een model zonder *links* en vergelijk deze met die van het model met *links*. Kies daartoe in het menu van logistische regressie de variabelen *geslacht, leeftijd* en *opl* in block 1. Door vervolgens op het vakje **NEXT** te klikken kun je de variabele *links* selecteren in block 2.

[4] Zie Hosmer en Lemeshow (1989) pag. 140.





In syntax:

```
LOGISTIC REGRESSION VARIABLES  stem
  /METHOD = ENTER geslacht leeftijd opl  /METHOD = ENTER links
  /PRINT = GOODFIT.
```

Dit geeft als output:

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	5,045	1	,025
	Block	5,045	1	,025
	Model	37,308	4	,000

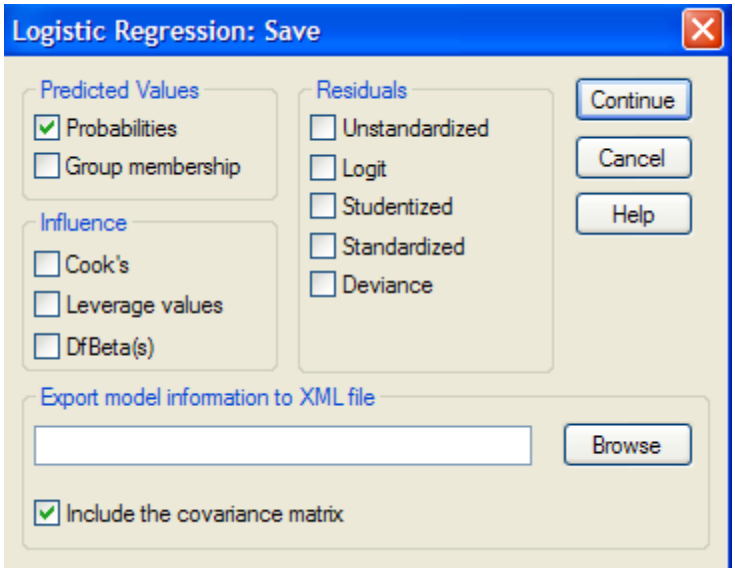
De Chi^2 bij de **Omnibus Tests of Model Coefficients** in Block 2 is 5,045 en is significant. De variabele *links* heeft dus inderdaad een significant effect op de kans om wel versus niet te gaan stemmen.

In kolom **B** staan de geschatte effecten op de logit (of log odds): de natuurlijke logaritme van de kansverhouding om wel versus niet te gaan stemmen. Hoe groter het getal, hoe groter het effect op deze logit. Net als bij lineaire regressie betekent een positief getal een positief effect, en een negatief getal een negatief effect. Het effect van leeftijd bijvoorbeeld is gelijk aan .033, wat betekent dat met ieder jaar de logit om wel versus niet te gaan stemmen toeneemt met .033. Omdat we meestal liever praten in termen van kansverhoudingen (odds) dan in logits, kijken we naar de kolom **Exp(B)**. Hierin zien we dat met ieder jaar de kans om wel te gaan stemmen versus de kans om niet te gaan stemmen met een factor 1.033 ($= e^{.033}$) groter wordt. Anders gezegd, ieder jaar neemt de odds om te gaan stemmen met 3,3% ((1.033 - 1) x 100%) toe. Bij een positief effect is de waarde van de **Exp(B)** groter dan 1, bij een negatief effect ligt de waarde tussen de 0 en de 1. Zo is de kans om wel versus niet te gaan stemmen .581 keer zo klein voor linkse mensen als voor rechtse mensen.

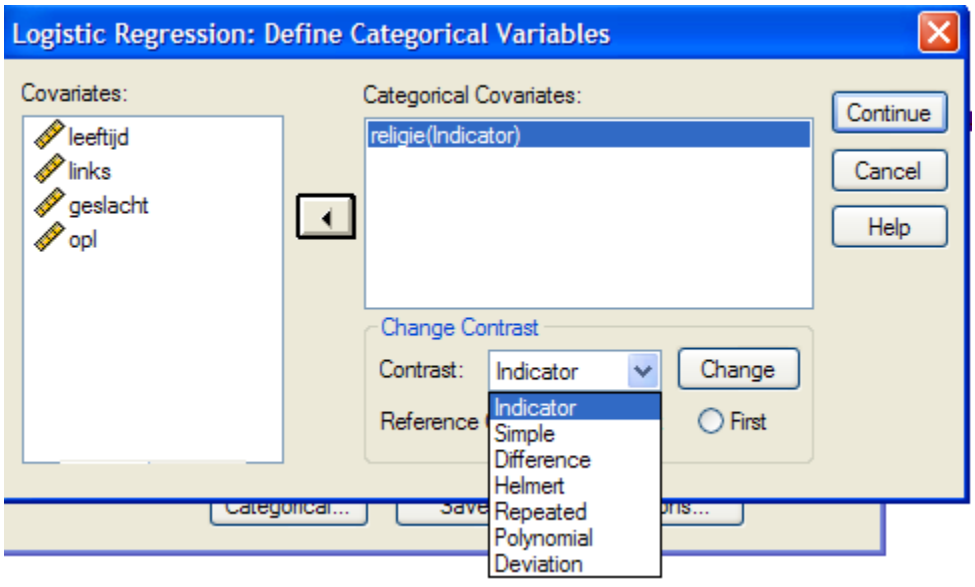
Aan de hand van de eerder gegeven kansformule, kunnen we nu ook de kans voorspellen dat iemand gaat stemmen. Willen we bijvoorbeeld weten wat de kans is dat een man (*geslacht*=0) van 30 jaar (*leeftijd*=30) die links is (*links*=1) en alleen lagere school heeft (*opl*=2) gaat stemmen, dan vullen we deze waarden in de kansformule in:

$$p_{wel} = \frac{e^{.198-.323*0+.033*30+.265*2-.542*1}}{e^{.198-.323*0+.033*30+.265*2-.542*1}+1} = 76,4\%$$

De kans voor deze 30-jarige, linkse, laag opgeleide man om wel te gaan stemmen bij de Tweede Kamer verkiezingen is dus 76,4%. De kans om niet te gaan stemmen is 23,6% (100,0%-76,4%). De kansen per individu kunnen in SPSS opgevraagd worden door in het menu onder **Save** te kiezen voor **Predicted values → Probabilities**.



Hoe kun je nominale variabelen in het model opnemen?
Natuurlijk kun je ook in een logistische regressie analyse onafhankelijke variabelen opnemen die nominaal zijn. In feite hebben we dit al gedaan door de variabelen *geslacht* en *links* in het model te plaatsen. Net als bij een lineaire regressie analyse moeten net zoveel dummy-variabelen gemaakt worden als er categorieën zijn, die alle op één na in het model worden opgenomen. Een gunstige eigenschap van logistische regressie is dat de dummies niet zelf gemaakt hoeven te worden. Door in het menu van logistische regressie de optie '**Categorical**' te kiezen, kunnen we aangeven hoe de dummies gemaakt moeten worden. Als voorbeeld nemen we hier de variabele *religie*, die vijf categorieën kerklidmaatschap kent: 1='rooms-katholiek', 2='nederlands hervormd', 3='gereformeerd', 4='anders' en 5='geen kerklid'.



In het menu moet bij de '**Categorical Covariates**' de nominale onafhankelijke variabele worden ingevuld. Bij de optie '**Contrast**' kunnen we kiezen op welke manier de dummies gemaakt moeten worden. Standaard wordt het indicator contrast gebruikt. Dit is misschien wel het meest

bekende contrast: één van de categorieën wordt als referentiecategorie gebruikt. Logistische regressie kiest standaard voor de laatste categorie, in ons voorbeeld is dat de categorie 'geen kerklid'. Wil je de eerste categorie als referentie nemen, dan klik je bij **Reference Category** de optie '**First**' aan. Een beschrijving van andere manieren om dummy variabelen te maken, kun je vinden door op **Help** te drukken.

In de syntax ziet deze aansturing er als volgt uit:

```
LOGISTIC REGRESSION VARIABLES  stem
  /METHOD = ENTER leeftijd links geslacht opl religie
  /CONTRAST (religie)=Indicator(5)
  /PRINT = GOODFIT.
```

Achter het **METHOD** commando, volgt nu een **CONTRAST** commando, waar tussen haakjes de onafhankelijke variabele genoemd wordt. Dan volgt het gekozen contrast, in ons geval het indicator-contrast. Tussen haakjes kan dan worden opgegeven welke categorie als referentiecategorie gebruikt moet worden; hier is voor de laatste gekozen. Als er meerdere nominale onafhankelijke variabelen zijn, dan moeten die allemaal apart worden opgegeven met een nieuw **CONTRAST** commando.

Met deze aansturing wordt de volgende output verkregen:

Categorical Variables Codings

			Parameter coding			
		Frequency	(2)	(3)	(4)	(1)
religie kerklidmaatschap	1 rooms-katholiek	337	1,000	,000	,000	,000
	2 nederlands hervormd	102	,000	1,000	,000	,000
	3 gereformeerd	36	,000	,000	1,000	,000
	4 anders	31	,000	,000	,000	1,000
	5 geen kerklid	381	,000	,000	,000	,000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	44,985	8	,000
	Block	44,985	8	,000
	Model	44,985	8	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	524,205(a)	,049	,104

a Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	leeftijd	,035	,011	9,870	1	,002	1,036
	links	-,589	,254	5,360	1	,021	,555
	geslacht	-,300	,236	1,625	1	,202	,741
	opl	,260	,059	19,345	1	,000	1,297
	religie			6,880	4	,142	
	religie(1)	-,366	,263	1,938	1	,164	,694
	religie(2)	,161	,443	,132	1	,717	1,175
	religie(3)	1,223	1,039	1,385	1	,239	3,399
	religie(4)	-,896	,508	3,115	1	,078	,408
	Constant	,288	,620	,216	1	,642	1,334

a Variable(s) entered on step 1: leeftijd, links, geslacht, opl, religie.

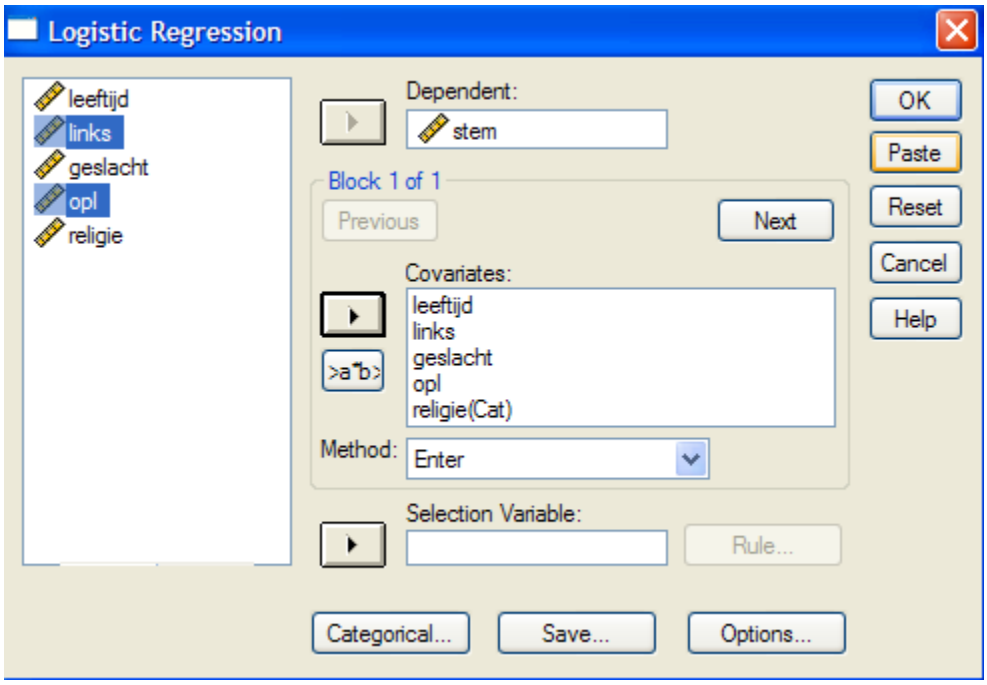
De interpretatie van de fit-maten is hetzelfde als bij onze eerste analyse. Door *religie* in het model op te nemen wordt de Chi^2 gelijk aan 44,985. Vergelijking met de Chi^2 van ons vorige model laat zien dat het verschil (44,985 - 37,308 = 7,677) niet significant is[5]. Het aantal vrijheidsgraden dat bij deze toets hoort is 4 (8 - 4; zie kolom kolommen **df**), hetgeen gelijk is aan het verschil in geschatte parameters: voor iedere opgenomen categorie van religie één. De verbetering van de pseudo R^2-maten is overigens ook gering.

Dat de variabele *religie* niet significant bijdraagt, blijkt ook uit de Wald-statistic. Deze is hier 6,880 en niet significant. Wanneer we naar de parameters voor de afzonderlijke categorieën van religie kijken, zien we dat de kans om wel versus niet te gaan stemmen voor rooms-katholieken .694 (kolom **exp(B)**) kleiner is dan voor de referentie-categorie 'geen kerklid'. Voor nederlands hervormden en gereformeerden is deze kans groter dan voor onkerkelijken, en voor anders gelovigen is de kans kleiner. Al deze verschillen blijken echter niet significant te zijn (kolom **Sig**).

In ons voorbeeld zijn zowel de Wald-statistic voor de variabele *religie* als geheel, als ook de Wald-statistic voor de afzonderlijke dummies niet significant. Net als in lineaire regressie analyse kan het echter voorkomen dat de variabele als geheel wel en de afzonderlijke categorieën niet significant zijn. Door op een andere manier dummies te maken (bijvoorbeeld door een andere referentie-categorie te kiezen of door een ander contrast te gebruiken), kunnen er wel significante verschillen tussen categorieën gevonden worden. Ook het omgekeerde kan het geval zijn: de variabele als geheel is niet significant maar één of meer dummies wel.

[5] Dit is in SPSS op te vragen door *religie* in Block 2 op te geven.

Hoe kun je interactietermen in het model opnemen?
Ook interactietermen zijn in logistische regressie mogelijk. In ons voorbeeld willen we nagaan of het effect dat opleiding heeft op de kans om wel of niet gaan stemmen bij Tweede Kamerverkiezingen, anders is voor linkse dan voor rechtse mensen. We kunnen dus een nieuwe variabele *links*opl* maken (bijvoorbeeld met het SPSS commando **COMPUTE**) en die opnemen in ons model. Maar logistische regressie analyse heeft ook een mogelijkheid om interactietermen te maken. Hiervoor selecteren we de twee variabelen die de interactie gaan vormen (hier dus *links* en *opl*) en drukken daarna op de '**a*b**'-knop. SPSS maakt dan de interactieterm *links*opl* aan.



De aansturing in de syntax is als volgt:

```
LOGISTIC REGRESSION VAR=stem
/METHOD=ENTER geslacht leeftijd links opl religie links*opl
/CONTRAST (religie)=Indicator(5)
/PRINT=GOODFIT.
```

En de bijbehorende output:

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	49,109	9	,000
	Block	49,109	9	,000
	Model	49,109	9	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	520,081(a)	,054	,114

a Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	geslacht	-,342	,238	2,073	1	,150	,710
	leeftijd	,035	,011	9,746	1	,002	1,036
	links	-1,670	,602	7,691	1	,006	,188
	opl	,096	,096	,997	1	,318	1,101
	religie			6,887	4	,142	
	religie(1)	-,345	,263	1,720	1	,190	,708
	religie(2)	,180	,445	,164	1	,686	1,197
	religie(3)	1,213	1,041	1,357	1	,244	3,362
	religie(4)	-,929	,511	3,305	1	,069	,395
	links by opl	,247	,121	4,174	1	,041	1,280
	Constant	1,052	,731	2,069	1	,150	2,862

a Variable(s) entered on step 1: geslacht, leeftijd, links, opl, religie, links * opl .

Vergelijking van de Chi^2-waarden laat zien dat dit model een significant betere fit heeft dan het voorgaande model. We concentreren ons bij de bespreking van de resultaten op de interactieterm *links*opl*. De interpretatie van een interactieterm in logistische regressie verloopt op dezelfde manier als in lineaire regressie analyse. In de output zien we dat de B-coëfficiënt voor opleiding gelijk is aan .096. Dit betekent dat voor respondenten met een waarde nul op de variabele *links* (dit zijn mensen die zichzelf als rechts beschouwen), het effect van opleiding erg klein en zelfs niet significant is. Voor linkse mensen daarentegen is het effect van opleiding beduidend groter, namelijk (.096 + .247 =) .343. Er is dus sprake van een significant verschillend effect van opleiding voor rechtse en voor linkse mensen (kolom **Sig**). Wanneer we weer in termen van kansverhoudingen willen praten, dan kijken we naar de kolom **exp(B)**. Hierin vinden we de waarde 1.101 als het effect van opleiding op de kans om wel versus niet te gaan stemmen voor rechtse mensen. Voor linkse mensen moeten we het effect van opleiding op de kansverhouding

zelf uitrekenen, dit is gelijk aan $e^{.343} = 1.409$.

Hoe presenteer je de resultaten?
Er zijn verschillende manieren om de resultaten te presenteren. Analoog aan een lineaire regressie analyse kun je de geschatte parameters opnemen in een tabel. Minimaal moet je de parameters B (of de Exp(B)) en de significantie van deze parameters vermelden. Ook is het goed om een fit-maat te vermelden, bijvoorbeeld de Chi^2 of een van de pseudo R^2-maten. In Tabel 1 wordt een voorbeeld gegeven, waarbij de B-coëfficiënten en de pseudo R^2 van Nagelkerke vermeld zijn.

Tabel 1: Geschatte parameters van logistische regressie-modellen voor de logit om te gaan stemmen bij Tweede Kamer verkiezingen.

	Model A		Model B		Model C	
	B-coëfficient	s.e.	B-coëfficient	s.e.	B-coëfficient	s.e.
constante	.198	.595	.288	.620	1.052	.731
geslacht	-.323	.234	-.300	.236	-.342	-.342
leeftijd	.033**	.011	.035**	.011	.035**	.035
linkse politieke voorkeur	-.542*	.247	-.589*	.254	-1.670**	-1.670
opleiding	.265**	.059	.260**	.059	.096	.096
religie (ref: geen kerklid)						
rooms katholiek			-.366	.263	-.345	-.345
nederlands hervormd			.161	.443	1.213	1.041
gereformeerd			1.223	1.039	.180	.445
anders			-.896	.508	-.929	-.511
opleiding * links					.247*	.121
Nagelkerke pseudo R^2	.087		.104		.114	

* significant (0.01<p<0.05) ** significant (p<0.01)

Een ander voorbeeld staat in Tabel 2. Hierin wordt de nadruk niet gelegd op de logit, maar op de kansverhouding om wel versus niet te gaan stemmen. We hebben hierbij een trucje toegepast voor de waarden van de exp(B). Omdat bij negatieve effecten een exp(B) hoort die kleiner is dan één, lijkt het op het eerste gezicht alsof deze effecten kleiner zijn dan de positieve effecten. Door nu voor alle negatieve effecten de exp(B) te noteren als $(1/\exp(B))^{-1}$, wordt dit 'gezichtsbedrog' weggenomen. Bijvoorbeeld: het effect van geslacht op de odds om te gaan stemmen bij Tweede Kamer verkiezingen is $(1/.71)^{-1} = 1.41^{-1}$. Merk trouwens op dat de constante geen exp(B) heeft. Achter iedere parameter staat de bijbehorende Wald-statistic. Sommige onderzoekers geven de voorkeur aan deze statistic, omdat ze uitdrukt hoe sterk de bijdrage van de parameter is. Tenslotte hebben we in plaats van een pseudo R^2 maat, de Chi^2 van het model vermeld met het bijbehorend aantal vrijheidsgraden.

Tabel 2 Uitkomsten van een logistische regressie analyse voor de kans om wel versus niet te gaan stemmen

	exp(B)	Wald
geslacht	1.41^{-1}	2.07
leeftijd	1.04 **	9.75
linkse politieke voorkeur	5.32^{-1} **	7.69
opleiding	1.10	1.00
religie (ref: geen kerklid)		6.89
rooms katholiek	1.41^{-1}	1.72
nederlands hervormd	3.36	.16
gereformeerd	1.20	1.36
anders	2.53	3.31
opleiding * links	1.28*	4.17
Chi^2 (df=9)	49,109	

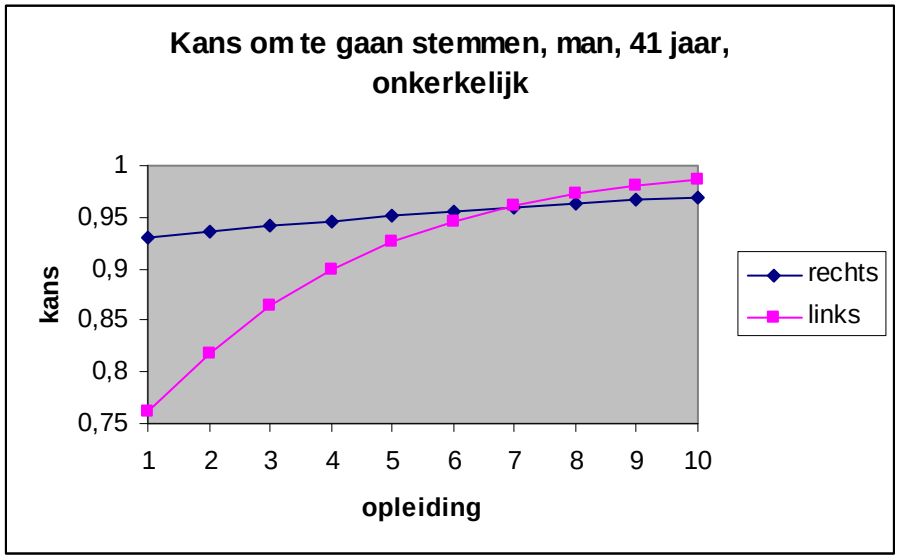
* .01<p<.05 ** p<.05

Wanneer je speciaal geïnteresseerd bent in het effect van één onafhankelijke variabele, bijvoorbeeld opleiding, op de kans om te gaan stemmen bij Tweede Kamer verkiezingen, dan kun je dit effect het beste weergeven in een grafiek. Hiervoor maken we gebruik van de kansformule:

$$p_{wel} = \frac{e^{1.052-.342*geslacht+.035*leeftijd-1.67*links+.096*opleiding-.345*rk+.180*nh+1.213*geref-.929*anders+.247*linksopl}}{e^{1.052-.342*geslacht+.035*leeftijd-1.67*links+.096*opleiding-.345*rk+.180*nh+1.213*geref-.929*anders+.247*linksopl}+1}$$

Zoals we al eerder gezien hebben, moeten er voor de variabelen waarden ingevuld worden. We willen hier twee effecten in een grafiek laten zien: het effect van opleiding voor linkse mensen en het effect van opleiding voor rechtse mensen. Dit betekent dat we voor *opl* één voor één de opleidingscategorieën (1 t/m 10) invullen, en voor *links* 0 (rechts) of 1 (links). Voor de overige variabelen uit ons model moeten we een keuze maken. Voor *leeftijd* vullen we 41 jaar in; dit is de gemiddelde leeftijd van de respondenten. We kiezen er verder voor om de kansen van mannen (*geslacht*=0) te laten zien. De variable geslacht heeft geen significant effect op het wel versus niet gaan stemmen bij Tweede Kamer verkiezingen, dus voor vrouwen zijn de kansen vrijwel hetzelfde als voor mannen. Wat religie betreft kiezen we voor de referentiecategorie 'onkerkelijk'. Dit vereenvoudigt de berekening omdat onkerkelijken op de vier dummyvariabelen van religie (*rk, nh, geref* en *anders*) een nul scoren. Zowel voor linkse als voor rechtse mensen beschikken we nu over tien kansen: één voor iedere opleidingscategorie. Deze kansen worden uitgezet in onderstaande grafiek[6].

[6] Grafiek is gemaakt met Excel.



Hieruit blijkt dat een hogere opleiding de kans om te gaan stemmen bij Tweede kamer verkiezingen verhoogt. Ook wordt duidelijk dat dit vooral het geval is voor mensen met een linkse politieke voorkeur. Volgens het model hebben linkse mensen met minder dan lagere school een kans van 76,1% om te gaan stemmen, terwijl deze kans voor linkse mensen met een universitaire opleiding tot maar liefst 98,6% is gestegen. De kans stijgt niet zo sterk voor mensen met een rechtse politieke voorkeur. Zij hebben - ongeacht hun opleidingsniveau - een zeer hoge kans om te gaan stemmen bij Tweede Kamer verkiezingen (93% tot 96,9%). We hadden al eerder gezien dat voor mensen met een rechtse politieke voorkeur het effect van opleiding niet significant is.

Wat is er verder nog van belang bij logistische regressie analyse?

Afhankelijke variabele met meer dan twee categorieën

Logistische regressie is geschikt voor de analyse van een nominale afhankelijke variabele met twee categorieën. Als er meer dan twee categorieën zijn, kun je proberen - op inhoudelijke of empirische gronden - de afhankelijke variabele te hercoderen naar twee categorieën. Een andere mogelijkheid is om meerdere logistische regressie analyses uitvoeren, waarbij je steeds twee categorieën met elkaar vergelijkt. Bijvoorbeeld bij een variabele met drie categorieën a, b en c kun je drie logistische regressie analyses uitvoeren: één voor de kans a versus b; één voor de kans a versus c; en één voor de kans b versus c. Het nadeel hiervan is dat de resultaten van de verschillende analyses meestal niet dezelfde uitkomsten geven. De beste oplossing is dan ook om een zogenaamde multinomiale logistische regressie analyse toepassen. Deze techniek is een uitbreiding van de gewone logistische regressie analyse en wordt hier verder niet behandeld. Ze staat beschreven in Lammers e.a. (2007).

Selectie van predictoren (forward, backward)

Meestal zul je besluiten om predictoren in het model op te nemen vanwege hun theoretisch belang. Wil je echter exploratief te werk gaan, dan kun je de selectie van onafhankelijke variabelen van de empirie laten afhangen, en gebruik maken van de voorwaartse (forward) of achterwaartse (backward) selectieprocedure van logistische regressie. De forward selectieprocedure start met een model zonder predictoren. De onafhankelijke variabelen worden

één voor één aan het model toegevoegd, waarbij telkens gekeken wordt of het model hierdoor verbetert. Na iedere opname wordt getoetst of de opgenomen variabelen weer verwijderd mogen worden. Bij de backward selectieprocedure worden eerst alle predictoren opgenomen, waarna bekeken wordt of er een variabele verwijderd kan worden. Daarna volgt weer het beurtelings opnemen en verwijderen van variabelen, net als in de forward selectieprocedure. Als criterium van selectie wordt vaak de aannemelijkheidsratio (**LR**) aangeraden. Beide selectieprocedures zijn te vinden in het menu van logistische regressie, onder de knop '**method**'. In plaats van '**Enter**' kun je nu dus kiezen voor '**Forward:LR**' of '**Backward:LR**'. In de syntax komt dit overeen met de commando's '**/METHOD=FSTEP(LR)**' voor een forward selectieprocedure, en '**/METHOD=BSTEP(LR)**' voor een backward selectieprocedure. Een uitgebreidere beschrijving van deze procedures kun je vinden in Lammers e.a. (2007) of in Norušis (2005).

Collineariteit
Wanneer de onafhankelijke predictoren sterk samenhangen, kunnen er -net als in lineaire regressie analyse- vreemde resultaten ontstaan. Logistische regressie analyse kent geen opties om collineariteit vast te stellen. Als je collineariteit vermoedt, dan kun je het beste een lineaire regressie doen en de collinearity diagnostics (zoals de condition index) opvragen. Gebruik daarbij dezelfde onafhankelijke variabelen als in de logistische regressie analyse en een willekeurige afhankelijke variabele.

Residuenanalyse
Ook in logistische regressie kan een residuen analyse uitgevoerd worden. Het gaat hierbij niet zozeer om het controleren van allerlei assumpties omtrent de errorterm, zoals in lineaire regressie analyse. Veel van die assumpties zijn immers niet van toepassing in logistische regressie. Wel kan het, vooral bij kleine steekproeven, van belang zijn om te bekijken of de residuen niet al te groot zijn, of dat er sprake is van outliers. Net als in lineaire regressie analyse, staan hiervoor verschillende (tijdelijke) variabelen ter beschikking. Logistische regressie maakt deze aan met de optie '**Save**' in het menu of het commando '**/CASEWISE**' in de syntax. Enkele voorbeelden zijn: RESID, ZRESID, LEVER, COOK en DFBETA. Deze variabelen worden beschreven in Lammers e.a. (2007).

Missing values
In logistische regressie analyse worden alle cases met missings als listwise behandeld, hetgeen betekent dat ze niet worden meegenomen in de analyse. Het is niet mogelijk om een zogenaamde pairwise analyse te doen. Wel kan het commando '**/MISSING=INCLUDE**' gegeven worden, waardoor je een bepaalde categorie, die van te voren als missing gedefinieerd is, toch in de analyse kunt meenemen. Dit kan handig zijn wanneer je ook het effect wilt weten van een bepaalde categorie als "weet niet", die als missing gedefinieerd is. De optie '**/MISSING=INCLUDE**' is niet terug te vinden in het menu van logistische regressie.

Waar kan ik meer literatuur over logistische regressie analyse vinden?

- Jan Lammers, Ben Pelzer, John Hendrickx, Rob Eisinga (2007*). Categorische Data Analyse met SPSS, Inleiding in loglineaire analysetechnieken*. Assen, Van Gorcum.
- Marija J. Norušis (2005). *SPSS 14.0 Statistical Procedures Companion*. Chicago, SPSS Inc.
- Scott Menard (1995). *Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage.
- David W. Hosmer & Stanley Lemeshow (1989). *Applied Logistic Regression*. New York: Wiley.

- Alfred DeMaris (1995). A Tutorial in Logistic Regression. *Journal of Marriage and the Family*, 57: 956-968.

Verder wordt er bij de vakgroep Methoden een tweetal cursussen gegeven waarin logistische regressie analyse behandeld wordt. 'Categorische data-analyse' (MTB9028) wordt verzorgd door Ben Pelzer en ‘Regressie Analyse voor Sociologie’ (MTB2022) door Ben Pelzer en Manfred te Grotenhuis. Voor meer informatie over deze cursussen kun je terecht in de studiegids of bij het secretariaat van Methoden.

Laatst gewijzigd in februari 2009, RTOG
http://www.ru.nl/socialewetenschappen/rtog/