HOOFDSTUK

15

Logistische regressie

I N L E I D I N G

De enkelvoudige en meervoudige of multipele lineaire regressiemethoden die we in de hoofdstukken 10 en 11 bestudeerden, worden als model gebruikt voor het verband tussen een te verklaren variabele en een of meer verklarende variabelen. Een belangrijke veronderstelling bij deze modellen is dat de verstoringen normaal verdeeld zijn. In dit hoofdstuk beschrijven we soortgelijke methoden die kunnen worden toegepast wanneer de te verklaren variabele maar twee mogelijke waarden heeft.

- Wat is het verband tussen de dosis van een insecticide en het al dan niet gedood worden van een insect?
- In hoeverre voorspelt het geslacht van een student het antwoord op de vraag of hij of zij een stevige drinker is?
- Geeft hoge bloeddruk een grotere kans op overlijden aan hart- en vaatziekten?

• • •

Veronderstel dat onze afhankelijke variabele maar twee waarden heeft: succes of mislukking, blijven leven of overlijden, accepteren of verwerpen. Als we deze twee waarden 1 en 0 noemen, is het gemiddelde de fractie enen, $p = P(\text{succes})$. In het geval van n onafhankelijke waarnemingen is sprake van de *binomiale situatie* (zie pagina 300). Hier is *nieuw* dat we beschikken over gegevens bij een *verklarende variabele* x. We zijn geïnteresseerd in de vraag hoe p van x afhangt. Veronderstel bijvoorbeeld dat we willen onderzoeken of patiënten die in een ziekenhuis worden opgenomen, blijven leven ($y = 1$) of overlijden ($y = 0$). In dit geval is p de kans dat een patiënt in leven blijft; mogelijke verklarende variabelen zijn (a) of de patiënt in goede of slechte conditie werd opgenomen, (b) de medische reden waarom de patiënt werd opgenomen, (c) de leeftijd van de patiënt. Merk op dat de verklarende variabelen zowel categorisch als kwantitatief kunnen zijn. Logistische regressie[1] is een statistische methode voor het beschrijven van dit soort relaties.

De binomiale verdeling en kansverhoudingen

In hoofdstuk 5 bestudeerden we de binomiale verdeling en in hoofdstuk 8 werd uitgelegd hoe statistische inferentie kan worden toegepast voor een kans p op succes binnen het binomiale model. We beginnen met een korte recapitulatie van de ideeën die we in dit hoofdstuk nodig zullen hebben.

VOORBEELD 15.1

In voorbeeld 8.1 werd een enquête onder 17.096 studenten beschreven die een vierjarige opleiding in de Verenigde Staten volgen. De onderzoekers waren geïnteresseerd in het schatten van het percentage studenten dat als frequente doorzakker kan worden aangemerkt. Een 'frequente doorzakker' werd gedefinieerd als een student die aangaf dat hij of zij gedurende de afgelopen twee weken op drie of meer dagen vijf of meer glazen alcohol had gedronken. In de in hoofdstuk 5 gebruikte notatie is p de fractie frequente doorzakkers in de totale populatie van studenten op scholen in de V.S. met een vierjarige opleiding. Het aantal doorzakkers in een enkelvoudige aselecte steekproef (EAS) met omvang n is binomiaal verdeeld met parameters n en p. De steekproefomvang is $n = 17.096$ en het aantal doorzakkers in de steekproef is 3314. De fractie in de steekproef is dus

$$\hat{p} = \frac{3314}{17.096} = 0{,}1938$$

•

kansverhoudingen
odds

Logistische regressie werkt met *kansverhoudingen* in plaats van met fracties. De kansverhouding, die meestal met het Engelse woord *odds*[2] wordt aangeduid, is eenvoudig de verhouding tussen de fracties bij twee mogelijke uitkomsten. Als $\hat{p}$ de fractie bij de eerste uitkomst is, dan is $1 - \hat{p}$ de fractie bij de tweede uitkomst en

$$\text{ODDS} = \frac{\hat{p}}{1 - \hat{p}}$$

Door p voor $\hat{p}$ te substitueren wordt de waarde voor de kansverhouding binnen de populatie berekend.

VOORBEELD 15.2

De fractie frequente doorzakkers in de steekproef is $\hat{p} = 0{,}1938$. De factie studenten die geen frequente doorzakkers zijn, is dus

$$1 - \hat{p} = 1 - 0{,}1938 = 0{,}8062$$

De kansverhouding voor studenten die frequente doorzakkers zijn, is dus

$$\begin{aligned}\text{ODDS} &= \frac{\hat{p}}{1 - \hat{p}} \\ &= \frac{0{,}1938}{0{,}8062} \\ &= 0{,}24\end{aligned}$$

•

Als men over kansverhoudingen spreekt, wordt vaak afgerond naar gehele getallen of naar breuken. Omdat 0,24 bij benadering gelijk is aan 1/4 zegt men dat de kansverhouding 1 tegen 4 is dat een student een frequente doorzakker is. Omgekeerd is de kansverhouding 4 tegen 1 dat dit *niet* het geval is.

In voorbeeld 8.8 (zie pagina 486) vergeleken we de fracties frequente doorzakkers onder mannelijke en vrouwelijke studenten met behulp van een betrouwbaarheidsinterval. We vonden een fractie van 0,227 (22,7%) voor mannelijke studenten en een fractie 0,170 (17,0%) voor vrouwelijke studenten. Het verschil is 0,057 en het 95%-betrouwbaarheidsinterval is (0,045; 0,069). Dit resultaat kan worden samengevat door de uitspraak: 'De fractie frequente doorzakkers is 5,7% hoger onder mannelijke studenten dan onder vrouwelijke'.

Een andere manier om deze gegevens te analyseren is met behulp van logistische regressie. De verklarende variabele is hier het geslacht,

indicatorvariabele

een categorische variabele. Om deze in een (logistische of anderssoortige) regressie te kunnen gebruiken, is een numerieke codering nodig. Dit wordt gewoonlijk gedaan met behulp van een *indicatorvariabele*. Hier zullen we als indicator gebruiken of de student een man is:

$$x = \begin{cases} 1 & \text{als de student een man is} \\ 0 & \text{als de student een vrouw is} \end{cases}$$

De afhankelijke variabele of responsvariabele is de fractie frequente doorzakkers. Voor het gebruik in een logistische regressie passen we twee transformaties toe op deze variabele. Ten eerste zetten we hem om naar een kansverhouding. Voor mannen krijgen we

$$\begin{aligned} \text{ODDS} &= \frac{\hat{p}}{1-\hat{p}} \\ &= \frac{0{,}227}{1-0{,}227} \\ &= 0{,}294 \end{aligned}$$

Voor vrouwen volgt

$$\begin{aligned} \text{ODDS} &= \frac{\hat{p}}{1-\hat{p}} \\ &= \frac{0{,}170}{1-0{,}170} \\ &= 0{,}205 \end{aligned}$$

Het logistische regressiemodel

Bij enkelvoudige lineaire regressie was de modelveronderstelling voor het gemiddelde μ van de afhankelijke variabele y dat dit een lineaire functie is van de verklarende variabele: $\mu = \beta_0 + \beta_1 x$. Bij logistische regressie zijn we geïnteresseerd in het modelleren van het gemiddelde van de responsvariabele p in termen van de verklarende variabele x. We zouden dit kunnen proberen met de relatie $p = \beta_0 + \beta_1 x$. Helaas is dit geen goed model. Zolang $\beta_1 \neq 0$ geven zeer hoge of lage waarden van x voor $\beta_0 + \beta_1 x$ een waarde die niet in overeenstemming is met het gegeven dat $0 \leq p \leq 1$.

De bij logistische regressie gekozen oplossing voor dit probleem is het transformeren van de kansverhouding $p/(1-p)$ met behulp van de natuurlijke logaritme. We gebruiken voor deze transformatie de term

logaritmische kansverhouding of *log odds*. Deze modelleren we als een lineaire functie van de verklarende variabele:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

In figuur 15.1 is de relatie tussen p en x voor verschillende waarden van β_0 en β_1 grafisch weergegeven. Voor logistische regressie gebruiken we *natuurlijke* logaritmen. Hiervoor bestaan tabellen en veel rekenmachines hebben er een ingebouwde functie voor. Net als bij lineaire regressie noemen we de afhankelijke variabele y. Voor mannelijke studenten volgt

$$y = \log(\text{ODDS}) = \log(0{,}294) = -1{,}23$$

en voor vrouwelijke

$$y = \log(\text{ODDS}) = \log(0{,}205) = -1{,}59$$

In deze expressies gebruiken we y als de waargenomen waarde van de responsvariabele, de logaritmische kansverhouding voor frequente doorzakkers. Nu kunnen we het logistische regressiemodel formuleren.

LOGISTISCH REGRESSIEMODEL

Het *statistische model voor logistische regressie* is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

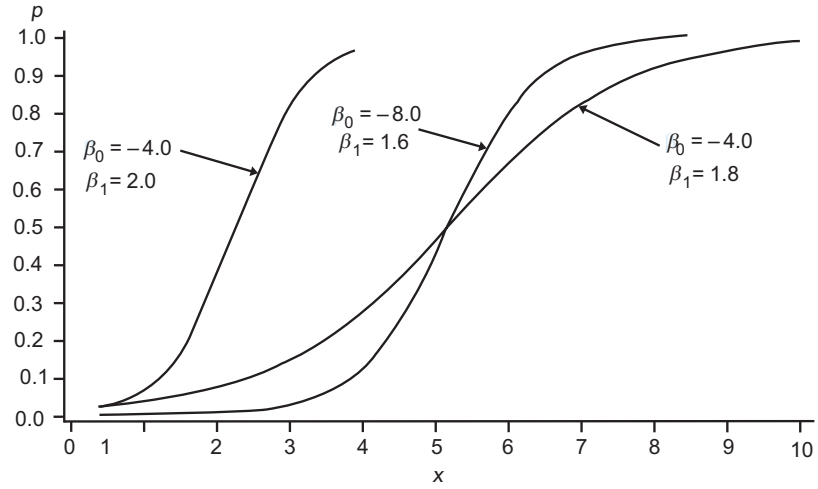
Hierbij is p een binomiale fractie en x de verklarende variabele. De parameters van het logistische model zijn β_0 en β_1.

VOORBEELD 15.3

Bij het voorbeeld van de frequent doorzakkende studenten bevat de steekproef $n = 17.096$ studenten. De verklarende variabele is het geslacht. We codeerden deze met een indicatorvariabele x met de waarde $x = 1$ voor mannen en $x = 0$ voor vrouwen. De responsvariabele is eveneens een indicatorvariabele. Een student is dus ofwel een frequente doorzakker, of is dat niet. Denk aan een proces waarbij een student aselect wordt getrokken en waarbij vervolgens de waarde x wordt vastgelegd en het gegeven of het hier wel of niet een frequente doorzakker betreft. Het model veronderstelt dat de kans (p) dat de student een frequente doorzakker is, afhangt van het geslacht van die student ($x = 1$ of

$x = 0$). Dus heeft p twee mogelijke waarden die we bijvoorbeeld p_{man} en p_{vrouw} kunnen noemen. •

Logistische regressie met een indicator als verklarende variabele is een heel speciaal geval. Het is belangrijk omdat veel meervoudige logistische regressieanalyses zich concentreren op een of meer van dergelijke variabelen als de belangrijkste verklarende variabelen. Hier gebruiken we dit speciale geval om wat meer inzicht in het model te verkrijgen.



Figuur 15.1 Een grafiek van p en x bij verschillende waarden van β_0 en β_1.

Het logistische regressiemodel specificeert de relatie tussen p en x. Omdat x alleen de waarden 1 en 0 kan aannemen, kunnen we twee vergelijkingen uitschrijven. Voor mannen

$$\log\left(\frac{p_{\text{man}}}{1 - p_{\text{man}}}\right) = \beta_0 + \beta_1$$

en voor vrouwen

$$\log\left(\frac{p_{\text{vrouw}}}{1 - p_{\text{vrouw}}}\right) = \beta_0$$

De term β_1 komt alleen in de vergelijking voor mannen voor, omdat daarin $x = 1$. De term ontbreekt in de vergelijking voor vrouwen omdat nu $x = 0$.

Aanpassen en interpreteren van het logistische regressiemodel

In het algemeen zijn de berekeningen voor het bepalen van schattingen b_0 en b_1 voor β_0 en β_1 ingewikkeld en kunnen ze alleen met behulp van een computerprogramma worden uitgevoerd. Wanneer de verklarende variabele echter maar twee mogelijke waarden kan aannemen, zijn de schattingen makkelijk te vinden. Binnen dit eenvoudige raamwerk kan ook duidelijk worden gemaakt wat de betekenis is van de parameters bij logistische regressie.

VOORBEELD 15.4

In het voorbeeld over de frequente doorzakkers vonden we als logaritmische kansverhouding voor mannen

$$y = \log\left(\frac{\hat{p}_{\text{man}}}{1-\hat{p}_{\text{man}}}\right) = \log(0{,}294) = -1{,}23$$

en voor vrouwen

$$y = \log\left(\frac{\hat{p}_{\text{vrouw}}}{1-\hat{p}_{\text{vrouw}}}\right) = \log(0{,}205) = -1{,}59$$

Het logistische regressiemodel voor mannen is

$$\log\left(\frac{p_{\text{man}}}{1-p_{\text{man}}}\right) = \beta_0 + \beta_1$$

en voor vrouwen

$$\log\left(\frac{p_{\text{vrouw}}}{1-p_{\text{vrouw}}}\right) = \beta_0$$

Om de schattingen voor b_0 en b_1 te bepalen, combineren we de mannelijke en vrouwelijke modelvergelijkingen met de bijbehorende gegevensvergelijkingen. We zien dan dat de schatting van de constante term b_0 eenvoudig gelijk is aan de log(ODDS) voor vrouwen:

$$b_0 = -1{,}59$$

terwijl de helling van de regressielijn het verschil is tussen de log(ODDS) voor mannen en de log(ODDS) voor vrouwen:

$$b_1 = -1{,}23 - (-1{,}59) = 0{,}36$$

•

De helling in dit logistische regressiemodel is het verschil tussen de log(ODDS) voor mannen en voor vrouwen. De meeste mensen hebben moeite met het denken in de log(ODDS) schaal. De interpretatie van de resultaten in termen van de helling van de regressielijn is daarom moeilijk. Gewoonlijk wordt een transformatie toegepast die de situatie verduidelijkt. Met enige algebra kan worden aangetoond dat

$$\frac{\text{ODDS}_{\text{mannen}}}{\text{ODDS}_{\text{vrouwen}}} = \text{e}^{0,36} = 1,43$$

odds ratio

De transformatie $\text{e}^{0,36}$ maakt de logaritme ongedaan en transformeert de helling van de regressielijn tot een *odds ratio*. In dit geval betreft het de verhouding tussen de kansverhouding dat de man een frequente doorzakker is en de kansverhouding dat dit voor de vrouw geldt. Dit betekent dat de kansverhouding voor mannen kan worden berekend door de kansverhouding voor vrouwen met de odds ratio te vermenigvuldigen:

$$\text{ODDS}_{\text{mannen}} = 1,43 \times \text{ODDS}_{\text{vrouwen}}$$

Hier is de kansverhouding voor mannen dus 1,43 maal die voor vrouwen.

Merk op dat de codering voor de indicatorvariabelen zodanig werd gekozen dat de helling van de regressielijn positief is. De odds ratio is dan groter dan 1. Hadden we vrouwen als 1 en mannen als 0 gecodeerd, dan waren de tekens van de parameters omgekeerd en was de aangepaste vergelijking $\log(\text{ODDS}) = 1,59 - 0,36x$ geweest. De odds ratio is dan $\text{e}^{-0,036} = 0,70$. De kansverhouding bij vrouwen is 70% van kansverhouding bij mannen.

Logistische regressie met een verklarende variabele met twee mogelijke waarden is een heel belangrijk speciaal geval. Hier volgt een voorbeeld waarin de verklarende variabele kwantitatief is.

VOORBEELD 15.5

De in de data-appendix (zie blz. 377 van het opgavenboek) beschreven gegevensverzameling CHEESE bevat een responsvariabele 'Taste' die de kwaliteit van de kaas meet op grond van verschillende beoordelingen. In dit voorbeeld klassificeren we de kaas als acceptabel (tasteOK = 1) indien Taste $\geq$ 37 en als niet acceptabel (tasteOK = 0) indien Taste $<$ 37. De variabele tasteOK is hier de responsvariabele. De gegevensverzameling bevat drie verklarende variabelen: Acetic (azijnzuurgehalte), H2S (zwavelwaterstofgehalte) en Lactic (melkzuurgehalte). Als we Acetic als verklarende variabele gebruiken, luidt het model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Hierbij is p de kans dat de kaas acceptabel is en x is de waarde van Acetic. Het model voor de met behulp van een computerprogramma bepaalde logaritmische kansverhouding is

$$\log(\text{ODDS}) = b_0 + b_1 x = -13{,}71 + 2{,}25x$$

De odds ratio is $e^{b_1} = 9{,}48$. Dit betekent dat door een verhoging van de hoeveelheid azijnzuur x met één eenheid de kansverhouding voor een acceptabele kaas met een factor 9,5 toeneemt. •

Inferentie bij logistische regressie

Statistische inferentie bij logistische regressie lijkt veel op die bij enkelvoudige lineaire regressie. We berekenen schattingen voor de modelparameters en de standaardfouten bij deze schattingen. Betrouwbaarheidsintervallen worden op de gebruikelijke wijze bepaald, maar we gebruiken standaardnormale z^*-waarden in plaats van kritieke waarden uit t-verdelingen. De basis voor het toetsen van een hypothese is de verhouding tussen de schatting en de standaardfout. De toetsingsgrootheden worden vaak als de kwadraten van deze verhoudingen gegeven en in dit geval worden de P-waarden verkregen uit een chi-kwadraatverdeling met 1 vrijheidsgraad.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-1.5869	0.0267	3520.4040	0.0001	.
X	1	0.3616	0.0388	86.6714	0.0001	1.436

Figuur 15.2 Resultaat van logistische regressie op de gegevens over frequente doorzakkers in voorbeeld 15.6.

We hebben het raamwerk voor het toetsen van hypothesen uitgedrukt in termen van de helling β_1 omdat deze vorm sterk lijkt op de wijze waarop we enkelvoudige lineaire regressie behandelden. Bij veel toepassingen worden de resultaten echter uitgedrukt in termen van de odds ratio. Een helling 0 stemt overeen met een odds ratio 1. De nulhypothese wordt daarom vaak geformuleerd als ‘de odds ratio is 1’. Dit betekent dat de twee kansverhoudingen gelijk zijn en dat de verklarende variabele ons niet helpt bij het voorspellen van de kansverhoudingen.

BETROUWBAARHEIDSINTERVALLEN EN SIGNIFICANTIETOETSEN VOOR LOGISTISCHE REGRESSIEPARAMETERS

Een niveau C betrouwbaarheidsinterval voor de helling β_1 is

$$b_1 \pm z^* \mathrm{SE}_{b_1}$$

De odds ratio is de kansverhouding bij een waarde $x + 1$ van de verklarende variabele gedeeld door de kansverhouding bij een waarde x. Een niveau C betrouwbaarheidsinterval bij de kansratio $e^{\beta 1}$ wordt verkregen door het betrouwbaarheidsinterval bij de helling te transformeren:

$$(e^{b_1 - z^* \mathrm{SE}_{b_1}}, e^{b_1 + z^* \mathrm{SE}_{b_1}})$$

In deze uitdrukkingen is z^* de waarde voor de standaardnormale verdelingsdichtheidscurve met een oppervlak C tussen $-z*$ en z^*. Om de hypothese $H_0 : \beta_1 = 0$ te toetsen, moet de volgende toetsingsgrootheid worden berekend:

$$\mathrm{X}^2 = \left(\frac{b_1}{\mathrm{SE}_{b_1}} \right)^2$$

In termen van een stochastische variabele X^2 die bij benadering een χ^2-verdeling heeft met 1 vrijheidsgraad, is de P-waarde voor een toetsing van H_0 tegen H_a: $\beta_1 \neq 0$ gelijk aan $P(\chi^2 > \mathrm{X}^2)$.

VOORBEELD 15.6

In figuur 15.2 is het resultaat van de logistische procedure in SAS afgedrukt voor het voorbeeld met de frequente doorzakkers. De parameterschattingen zijn $b_0 = -1{,}5869$ en $b_1 = 0{,}3616$. Dit zijn dezelfde waarden als we eerder in voorbeeld 15.4 bij een rechtstreekse berekening vonden, maar met meer significante cijfers. De standaardfouten zijn 0,0267 en 0,0388. Een 95%-betrouwbaarheidsinterval voor de helling is

$$\begin{aligned} b_1 \pm z^* \mathrm{SE}_{b_1} &= 0{,}3616 \pm (1{,}96)(0{,}0388) \\ &= 0{,}3616 \pm 0{,}0760 \end{aligned}$$

Met 95%-betrouwbaarheid ligt de helling tussen 0,2855 en 0,4376. De output levert als odds ratio de waarde 1,436 maar vermeldt geen betrouwbaarheidsinterval. Dit kan echter gemakkelijk uit het betrouwbaarheidsinterval van de

helling worden berekend:

$$(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}}) = (e^{0,2855}, e^{0,4376}) = (1,33; 1,55)$$

Het resultaat is in dit geval de conclusie: 'Mannelijke studenten hebben een hogere kans om frequente doorzakkers te zijn dan vrouwelijke studenten (odds ratio = 1,44. 95% BI is 1,33 tot 1,55). •

In toepassingen zoals deze is het gebruikelijk 95% te kiezen voor de betrouwbaarheidscoëfficiënt. Bij deze conventie levert het betrouwbaarheidsinterval een toetsing van de nulhypothese dat de odds ratio 1 is met een significantieniveau van 0,05. Als het betrouwbaarheidsinterval de waarde 1 niet bevat, wordt de nulhypothese verworpen ten gunste van de alternatieve hypothese dat de kansverhoudingen (odds) voor de twee groepen (mannen en vrouwen) verschillend zijn. Als de H_0 niet kan worden verworpen, leveren de gegevens onvoldoende aanwijzingen om onderscheid tussen beide groepen te kunnen maken.

Het volgende voorbeeld is typerend voor veel toepassingen van logistische regressie. Het gaat hier om een experiment met vijf verschillende waarden voor de verklarende variabele.

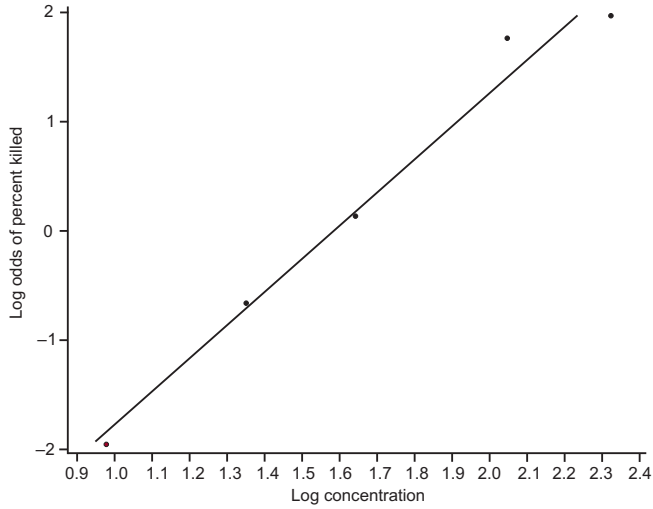
VOORBEELD 15.7

Er werd een experiment opgezet om te onderzoeken hoe goed het insecticide rotenone luizen met de naam *Macrosiphoniella sanborni* doodt die leven op een chrysant[3]). De verklarende variabele is de logaritme uit de concentratie (in milligram per liter) van het insecticide. Telkens werden ongeveer 50 insecten aan een bepaalde concentratie blootgesteld. Een insect werd gedood of niet gedood. We vatten de gegevens samen door een overzicht te geven van de aantallen gedode insecten.

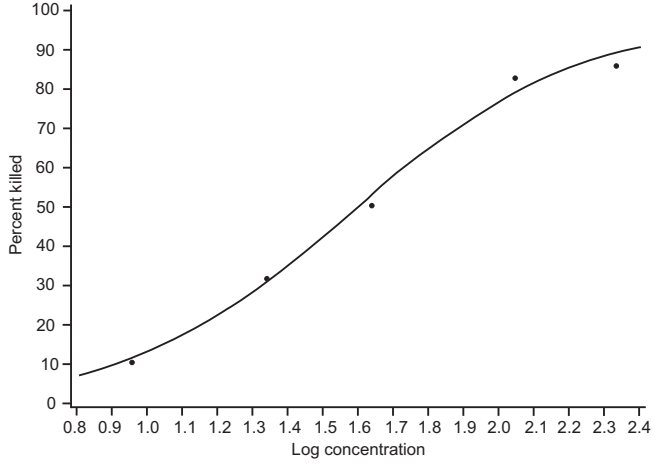
De responsvariabele voor logistische regressie is de logaritmische kansverhouding bij de gedode fractie. Hier volgen de resultaten:

Concentratie (log)	Aantal insecten	Aantal gedood
0,96	50	6
1,33	48	16
1,63	46	24
2,04	49	42
2,32	50	44

Wanneer we de responsvariabele transformeren door de logaritmische kansverhouding te berekenen en als we vervolgens kleinste-kwadratenregressie toepassen, krijgen we de aanpassing die in de grafiek in figuur 15.3 is weergegeven.



Figuur 15.3 Grafiek van de logaritmische kansverhoudingen van de gedode percentages tegen de logaritme van de concentratie van de insecticide met de gegevens uit voorbeeld 15.7.



Figuur 15.4 Grafiek van het gedode percentage tegen de logaritme van de concentratie met logistische aanpassing voor de insecticide gegevens uit voorbeeld 15.7.

De aanpassing bij logistische regressie is in figuur 15.4 weergegeven. Dit is een getransformeerde versie van figuur 15.3, waarbij de aanpassing op grond van het logistische model werd berekend. •

Een van de belangrijkste thema's in dit boek is de weergave van statistische resultaten in de vorm van grafieken. Voor het voorbeeld met het insecticide is dit gebeurd in figuur 15.4 en de resultaten lijken aanvaardbaar. Maar veronderstel eens dat rotenone niet in staat is *Macrosiphoniella sanborni* te doden. Wat is dan de kans dat we een resultaat waarnemen in een experiment dan minstens zo overtuigend is als wat we zien als de veronderstelling wel waar is? Het antwoord is de P-waarde bij de toetsing van de nulhypothese dat de helling van de regressielijn nul is. Als deze P-waarde niet klein is, kan de grafiek misleidend zijn. Statistische inferentie geeft een oplossing voor dit probleem.

VOORBEELD 15.8

De resultaten van de logistische berekening met SAS voor de analyse van de insecticidegegevens zijn in figuur 15.5 afgedrukt. Het model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

De waarden van de verklarende variabele x zijn 0,96; 1,33; 1,63; 2,04; 2,32. De resultaten leveren het volgende aangepaste model

$$\log(\text{ODDS}) = b_0 + b_1 x = -4{,}89 + 3{,}10x$$

Dit is de aanpassing die in figuur 15.4 grafisch is weergegeven. De nulhypothese $\beta_1 = 0$ wordt duidelijk verworpen ($X^2 = 64{,}07$ en $P < 0{,}001$). We berekenen een 95%-betrouwbaarheidsinterval voor β_1 met de schatting $b_1 = 3{,}1035$ en de standaardfout $\text{SE}_{b_1} = 0{,}3877$. Het resultaat is:

$$\begin{aligned} b_1 \pm z^* \text{SE}_{b_1} &= 3{,}1035 \pm (1{,}96)(0{,}3877) \\ &= 3{,}1035 \pm 0{,}7599 \end{aligned}$$

Met 95% betrouwbaarheid ligt de werkelijke waarde van de helling van de regressielijn tussen 2,34 en 3,86.

De odds ratio wordt berekend als 22,277. Een toename van één eenheid in de logaritmische concentratie van de insecticide doet de kansverhouding bij het gedood worden van een insect met een factor 22 toenemen. Het betrouwbaarheidsinterval voor de odds ratio kan worden afgeleid uit het interval bij de helling:

$$\begin{aligned} (e^{b_1 - z^* \text{SE}_{b_1}}, e^{b_1 + z^* \text{SE}_{b_1}}) &= (e^{2{,}34361}; e^{3{,}86339}) \\ &= (10{,}42; 47{,}63) \end{aligned}$$

Merk opnieuw op dat een toets voor de nulhypothese dat de helling nul is, hetzelfde is als een toets voor de nulhypothese dat de kansverhoudingen 1 zijn. Als we het resultaat in termen van kansverhoudingen zouden rapporteren,

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-4.8869	0.6429	57.7757	0.0001	.
LCONC	1	3.1035	0.3877	64.0744	0.0001	22.277

Figuur 15.5 Logistische regressie op de insecticidegegevens in voorbeeld 15.8.

zouden we kunnen zeggen: 'De kansverhouding bij het gedood worden van een insect neemt met een factor 22,3 toe met elke eenheid van toename in de logaritme uit de dosis van het insecticide ($X^2 = 64{,}07$; $P < 0{,}001$; 95% BI is 10,4 tot 47,6). •

In voorbeeld 15.5 bestudeerden we het probleem van het voorspellen van de aanvaardbaarheid van de smaak van kaas wanneer Acetic als verklarende variabele werd gebruikt. We bekijken dit voorbeeld nu opnieuw en laten zien hoe statistische inferentie een belangrijk onderdeel van de conclusie vormt.

VOORBEELD 15.9

In figuur 15.6 is het resultaat weergegeven van een logistische regressie met Acetic (azijnzuurgehalte) als verklarende variabele. Het aangepaste model werd in voorbeeld 15.5 vermeld:

$$\log(\text{ODDS}) = b_0 + b_1 x = -13{,}71 + 2{,}25x$$

Het resultaat geeft aan dat de nulhypothese $\beta_1 = 0$ in dit geval verworpen kan worden omdat $P = 0{,}0285$. De waarde van de toetsingsgrootheid is $X^2 = 4{,}79$ met 1 vrijheidsgraad. We gebruiken de schatting $b_1 = 2{,}2490$ met de standaardfout $\text{SE}_{b_1} = 1{,}0271$ om het 95%-betrouwbaarheidsinterval voor β_1 te berekenen:

$$\begin{aligned} b_1 \pm z^* \text{SE}_{b_1} &= 2{,}2490 \pm (1{,}96)(1{,}0271) \\ &= 2{,}2490 \pm 2{,}0131 \end{aligned}$$

De schatting voor de helling is 2,25 met een betrouwbaarheid van 95% dat de werkelijke waarde tussen 0,24 en 4,26 ligt. De schatting van de odds ratio is 9,48. Het 95%-betrouwbaarheidsinterval hierbij is

$$\begin{aligned} (e^{b_1 - z^* \text{SE}_{b_1}}, e^{b_1 + z^* \text{SE}_{b_1}}) &= (e^{0{,}23588}, e^{4{,}26212}) \\ &= (1{,}27;\ 70{,}96) \end{aligned}$$

•

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-13.7052	5.9319	5.3380	0.0209	.
ACETIC	1	2.2490	1.0271	4.7947	0.0285	9.479

Figuur 15.6 Logistische regressie op de kaasgegevens in voorbeeld 15.9 met Acetic (azijnzuurgehalte) als verklarende variabele.

Onze schatting is dat het verhogen met één eenheid van de hoeveelheid azijnzuur in de kaas, de kansverhouding op acceptabel zijn van de kaas met een factor 9 verhoogt. De gegevens leveren echter geen erg precieze schatting. De odds ratio kan met een betrouwbaarheid van 95% enerzijds slechts weinig groter dan 1 zijn en anderzijds de waarde 71 hebben. We hebben enige aanwijzingen voor de conclusie dat kazen met een hogere concentratie azijnzuur een grotere kans hebben om acceptabel te zijn, maar om het werkelijke verband nauwkeurig vast te leggen zijn meer gegevens nodig.

Meervoudige logistische regressie

Het zo juist behandelde kaasvoorbeeld brengt ons vanzelf bij een volgend onderwerp. De gegevensverzameling bevat drie variabelen: Acetic (azijnzuurgehalte), H2S (zwavelwaterstofgehalte) en Lactic (melkzuurgehalte). We bespraken een model waarin Acetic gebruikt werd om de kansverhouding voor een acceptabele kaas te voorspellen. Bevatten de overige verklarende variabelen extra informatie die ons een betere voorspelling geeft? We gebruiken *meervoudige logistische regressie* om deze vraag te beantwoorden. Het door de computer laten berekenen van de resultaten is eenvoudig. Het is vergelijkbaar met het generaliseren van enkelvoudige lineaire regressie met één verklarende variabele naar meervoudige lineaire regressie met meer dan een verklarende variabele (zie hoofdstuk 11). De statistische begrippen zijn vergelijkbaar, maar de berekeningen zijn wel ingewikkelder. Hier volgt een voorbeeld.

meervoudige logistische regressie

VOORBEELD 15.10

Net als in voorbeeld 15.8 willen we weer de kansverhouding bij een acceptabele kaas voorspellen. De verklarende variabelen zijn nu Acetic, H2S en Lactic. De resultaten zijn in figuur 15.7 afgedrukt. Het aangepaste model luidt

$$\begin{aligned}\log(\text{ODDS}) &= b_0 + b_1\text{Acetic} + b_2\text{H2S} + b_3\text{Lactic}\\ &= -14{,}26 + 0{,}58\text{Acetic} + 0{,}68\text{H2S} + 3{,}47\text{Lactic}\end{aligned}$$

Bij analyse van de gegevens met meervoudige lineaire regressie onderzoeken we eerst de hypothese dat de regressiecoëfficiënten bij de verklarende variabelen nul zijn. Bij logistische regressie doen we dit ook. De hypothese

$$H_0\colon \beta_1 = \beta_2 = \beta_3 = 0$$

wordt getoetst met behulp van een chi-kwadraattoetsingsgrootheid met drie vrijheidsgraden. Deze staat in de computeruitvoer vermeld op de regel die begint met '-2 LOG L' onder het kopje 'Chi-Square for Covariates'. De toetsingsgrootheid is $X^2 = 16{,}33$ en de P-waarde is 0,001. We verwerpen H_0 en concluderen dat het mogelijk is een of meer van de verklarende variabelen te gebruiken om de kansverhouding bij het acceptabel zijn van de kaas te voorspellen. Vervolgens bekijken we de coëfficiënten bij de variabelen afzonderlijk en toetsen de hypothesen dat deze nul zijn. We vinden als P-waarden 0,71 0,09 en 0,19. Geen van de nulhypothesen $H_0\colon < \beta_1 = 0$, $H_0\colon \beta_2 = 0$ en $H_0\colon \beta_3 = 0$ kan worden verworpen. •

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
-2 LOG L	34.795	18.461	16.334 with 3 DF (p=0.0010)

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-14.2604	8.2869	2.9613	0.0853	.
ACETIC	1	0.5845	1.5442	0.1433	0.7051	1.794
H2S	1	0.6849	0.4040	2.8730	0.0901	1.983
LACTIC	1	3.4684	2.6497	1.7135	0.1905	32.086

Figuur 15.7 Resultaten bij logistische regressie op de kaasgegevens in voorbeeld 15.10 met Acetic, H2S en Lactic als verklarende variabelen.

Samenvatting

Als $\hat{p}$ de fractie in de steekproef is, dan wordt de **kansverhouding** of **odds** gedefinieerd als $\hat{p}/(1-\hat{p})$. Dit is de verhouding tussen de fractie van het totaal aantal waarnemingen waarin een gebeurtenis optreedt en de fractie waarin deze niet optreedt.

Het **logistische regressiemodel** legt een verband tussen de logaritme uit de kansverhouding (de log odds) en de verklarende variabele:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 x_i$$

Hierbij zijn de responsvariabelen x_i voor $i = 1, 2, \ldots, n$ onafhankelijke binomiaal verdeelde stochastische grootheden met parameters 1 en p_i. Ze zijn dus onderling onafhankelijk met verdelingen $B(1, p_i)$. De verklarende variabele is x.

De **parameters** van het logistische model zijn β_0 en β_1.

De **odds ratio** is e^{β_1}. β_1 is de helling van het logistische regressiemodel.

Het **niveau *C* betrouwbaarheidsinterval bij de constante** β_0 is

$$b_0 \pm z^* \mathrm{SE}_{b_0}$$

Een **niveau *C* betrouwbaarheidsinterval bij de helling** β_1 is

$$b_1 \pm z^* \mathrm{SE}_{b_1}$$

Een **niveau *C* betrouwbaarheidsinterval voor de odds ratio** e^{β_1} wordt verkregen uit een transformatie van het betrouwbaarheidsinterval bij de helling:

$$(e^{b_1 - z^* \mathrm{SE}_{b_1}}, e^{b_1 + z^* \mathrm{SE}_{b_1}})$$

In deze uitdrukkingen is z^* de waarde met oppervlak C tussen $-z^*$ en z^* onder de standaardnormale verdelingsdichtheid.

Om de hypothese H_0: $\beta_1 = 0$ te toetsen moet de volgende **toetsingsgrootheid** worden berekend:

$$\chi^2 = (b_1 / \mathrm{SE}_{b_1})^2$$

In termen van een stochastische variabele X^2 met een χ^2 verdeling met 1 vrijheidsgraad, is de P-waarde bij toetsing van H_0 tegen de alternatieve hypothese H_a: $\beta_1 \neq 0$ gelijk aan $P(\chi^2 \geq \mathrm{X}^2)$. Dit is hetzelfde als toetsing van de nulhypothese dat de odds ratio de waarde 1 heeft.

Bij **meervoudige logistische regressie** heeft de responsvariabele net als bij enkelvoudige logistische regressie twee mogelijke waarden. Er kan echter sprake zijn van meer dan een verklarende variabele.

OPGAVEN BIJ HOOFDSTUK 15

15.1 Er zijn veel aanwijzingen dat hoge bloeddruk verband houdt met een verhoogd risico op overlijden aan hart- en vaatziekten. Bij het zoeken naar dit verband werden 3338 mannen met hoge bloeddruk en 2676

mannen met lage bloeddruk onderzocht. Tijdens de periode van onderzoek overleden 21 mannen in de groep met lage bloeddruk en 55 in de groep met hoge bloeddruk aan een hart- en vaatziekte.

(a) Bereken de fractie mannen die aan een hart- en vaatziekte overleden in de groep met hoge bloeddruk. Bereken vervolgens de kansverhouding (odds).

(b) Doe hetzelfde voor de groep met lage bloeddruk.

(c) Bereken de odds ratio met de odds bij de groep met hoge bloeddruk in de noemer. Beschrijf het resultaat in woorden.

15.2 In welke mate vertonen grammaticaboeken waarin de structuur van zinnen wordt geanalyseerd een voorkeur voor een mannelijke of vrouwelijke terminologie? Om deze vraag te onderzoeken werden zinnen uit tien teksten in een steekproef opgenomen. Bij het onderzoek werd ondermeer gekeken naar het gebruik van de woorden 'meisje', 'jongen', 'man' en 'vrouw'. We zullen de eerste twee woorden 'jeugdig' noemen en de laatste twee 'volwassen'. Hier volgen de resultaten uit een van de onderzochte teksten. (Zie Monica Macaulay en Colleen Brice, 'Don't touch my projectile: gender bias and stereotyping in syntactic examples', *Language*, 73, no. 4 (1997), pp. 798-825.)

Geslacht	n	X(jeugdig)
Vrouwelijk	60	48
Mannelijk	132	52

(a) Bereken de fractie jeugdige vrouwelijke verwijzingen en transformeer deze naar een kansverhouding (odds).

(b) Doe hetzelfde met de jeugdige mannelijke verwijzingen.

(c) Wat is de odds ratio voor het vergelijken van vrouwelijke verwijzingen naar mannelijke? (Zet de vrouwelijke kansverhouding in de noemer.)

15.3 In oefening 15.1 werd een onderzoek naar het verband tussen hart- en vaatziekten en bloeddrukhoogte beschreven. Een computerberekening met logistische regressieanalyse levert als schatting voor de helling $b_1 = 0{,}7505$ met standaardfout $\mathrm{SE}_{b_1} = 0{,}2578$.

(a) Bereken het 95%-betrouwbaarheidsinterval bij de helling.

(b) Bereken de X^2 toetsingsgrootheid voor het toetsen van de nulhypothese dat de helling nul is. Gebruik tabel F om een benadering voor de P-waarde te vinden.

(c) Beschrijf in het kort de resultaten en de conclusies.

15.4 In oefening 15.2 werd een onderzoek naar voorkeur voor geslacht in grammaticaboeken beschreven. Bij een analyse met logistische regres-

sie werd als schatting voor de helling $b_1 = 1{,}8171$ gevonden en als standaardfout $\mathrm{SE}_{b_1} = 0{,}3686$.

(a) Bereken het 95%-betrouwbaarheidsinterval bij de helling.

(b) Bereken de X^2 toetsingsgrootheid voor het toetsen van de nulhypothese dat de helling nul is. Gebruik tabel F om een benadering voor de P-waarde te vinden.

(c) Beschrijf in het kort de resultaten en de conclusies.

15.5 In oefening 15.3 werden de resultaten van een onderzoek naar het verband tussen bloeddrukhoogte en het overlijden aan hart- en vaatziekten beschreven in termen van het veranderen van de log odds.

(a) Transformeer de helling naar de odds en transformeer het 95%-betrouwbaarheidsinterval bij de helling naar dat bij de odds.

(b) Schrijf een conclusie waarin de odds wordt gebruikt om het resultaat toe te lichten.

15.6 In oefening 15.4 werden de resultaten naar voorkeur voor geslacht in grammaticaboeken op een log odds schaal beschreven.

(a) Transformeer de helling naar de odds en transformeer het 95%-betrouwbaarheidsinterval bij de helling naar dat bij de odds.

(b) Schrijf een conclusie waarin de odds wordt gebruikt om het resultaat toe te lichten.

15.7 Om te kunnen concurreren op de wereldmarkt ondergaan veel bedrijven in de Verenigde Staten omvangrijke reorganisaties. Het gaat hierbij vaak om 'afslanking' (downsizing) en reductie van het aantal personeelsleden (reduction in force of RIF). Volgens federale wetten en de wetten in verschillende staten, mag hierbij geen leeftijdsdiscriminatie worden toegepast. Veel personeel ouder dan 40 heeft echter extra bescherming en veel beschuldigingen over leeftijdsdiscriminatie berusten dan ook op het vergelijken van de behandeling van medewerkers ouder dan 40 met die van hun jongere collega's. Hier volgende gegevens over een recente RIF:

	Ouder dan 40	
Ontslagen	Ja	Nee
Ja	7	41
Nee	504	765

(a) Beschrijf het logistische regressiemodel bij dit probleem met de log odds van de RIF als responsvariabele en een indicator voor ouder dan 40 en niet ouder dan 40 als verklarende variabele.

(b) Licht de veronderstellingen ten aanzien van binomiale verdelingen toe in termen van de variabelen in deze oefening. In hoeverre zijn deze veronderstelling naar uw mening redelijk?

(c) Een computerprogramma levert als schatting voor de helling $b_1 = 1{,}3504$ met standaardfout $\mathrm{SE}_{b_1} = 0{,}4130$. Transformeer de resultaten naar de odds-schaal. Vat de resultaten samen en schrijf een korte conclusie.

(d) Indien er extra verklarende variabelen beschikbaar zouden zijn, bijvoorbeeld een prestatiebeoordeling, hoe zou u deze informatie dan gebruiken om de RIF te onderzoeken?

15.8 Tijdens een onderzoek naar het verband tussen alcoholgebruik en dodelijke fietsongevallen werden gegevens verzameld over een groot aantal dodelijke fietsongevallen. Bij alle slachtoffers werd het geslacht vastgelegd en of de test op alcohol een positieve uitslag had. Hier volgen de gegevens. (Uit Guohua Li en Susan P. Baker, 'Alcohol in fatally injured bicyclists', *Accident Analysis and Prevention*, 26 (1994), pp. 543-548.)

Geslacht	n	X (positieve test)
Vrouw	191	27
Man	1520	515

Pas logistische regressie toe om de vraag te bestuderen of geslacht wel of niet verband houd met alcoholgebruik bij mensen die dodelijk gewond raakten bij fietsongevallen.

15.9 In de voorbeelden 15.5 en 15.9 werden gegevens uit de gegevensverzameling CHEESE geanalyseerd. We gebruikten daarbij Acetic (azijnzuurgehalte) als verklarende variabele. Herhaal de analyse met H2S (zwavelwaterstofgehalte) als verklarende variabele.

15.10 Herhaal oefening 15.9, maar gebruik nu Lactic (melkzuurgehalte) als verklarende variabele.

15.11 Bij de kaasgegevens uit de voorbeelden 15.5, 15.9 en 15.10 en de oefeningen 15.9 en 15.10 zijn drie verklarende variabelen beschikbaar. Het is mogelijk hierop drie verschillende logistische regressies toe te passen met twee van deze verklarende variabelen. Voer deze regressies uit. Geef een overzicht van de resultaten met een, twee en drie verklarende variabelen. Wat is uw conclusie?

De volgende vier opgaven maken gebruik van de gegevensverzameling CSDATA uit de Data-appendix. We onderzoeken modellen die verband leggen tussen het met de GPA (grade point average) gemeten succes en diverse verklarende variabelen. We definiëren hier een indicatorvariabele, HIGPA, die 1 is als de GPA een waarde hoger dan of gelijk aan 3,0 heeft en die 0 is als dit niet het geval is.

15.12 Gebruik logistische regressie om HIGPA te voorspellen met de drie gemiddelde eindexamenuitslagen (HSS, HSM en HSE) als verklarende variabelen.

(a) Geef een overzicht van de resultaten van de toetsing van de hypothese dat de coëfficiënten bij alle drie de verklarende variabelen nul zijn.

(b) Geef de coëfficiënt voor HSM (High School Math, het gemiddelde cijfer voor wiskunde) met een 95%-betrouwbaarheidsinterval. Doe hetzelfde met de twee andere voorspellende grootheden in dit model.

(c) Geef een samenvatting van uw conclusies op grond van de resultaten in (a) en (b).

15.13 Gebruik logistische regressie om HIGPA te voorspellen met de twee SAT-scores (SATM en SATV) als verklarende variabelen.

(a) Geef een overzicht van de resultaten van de toetsing van de hypothese dat de coëfficiënten bij beide verklarende variabelen nul zijn.

(b) Geef de coëfficiënt voor SATM (SAT Math) met een 95%-betrouwbaarheidsinterval. Doe hetzelfde voor SATV (SAT Verbal).

(c) Geef een samenvatting van uw conclusies op grond van de resultaten in (a) en (b).

15.14 Gebruik logistische regressie om HIGPA te voorspellen met de drie gemiddelde eindexamenuitslagen (HSS, HSM en HSE) en de twee SAT-scores (SATM en SATV) als verklarende variabelen. Gevraagd wordt een soortgelijke analyse als in hoofdstuk 11 werd uitgevoerd.

(a) Toets de nulhypothese dat de coëfficiënten bij de drie gemiddelde eindexamenuitslagen nul zijn. Dus toets H_0: $\beta_{\text{HSM}} = \beta_{\text{HSS}} = \beta_{\text{HSE}} = 0$.

(b) Toets de nulhypothese dat de coëfficiënten bij de twee SAT-cores nul zijn. Dus toets H_0: $\beta_{\text{SATM}} = \beta_{\text{SATV}} = 0$.

(c) Wat concludeert u uit de toetsingsresultaten in (a) en (b)?

15.15 In deze oefening onderzoeken we de invloed van geslacht op het behalen van een hoge GPA.

(a) Gebruik geslacht om HIGPA met logistische regressie te voorspellen. Geef een samenvatting van de resultaten.

(b) Voer een logistische regressie uit met geslacht en de twee SAT-scores om HIGPA te voorspellen. Geef een samenvatting van de resultaten.

(c) Vergelijk de resultaten in (a) en (b) met betrekking tot de invloed van geslacht op HIGPA. Geef een samenvatting van de resultaten.

15.16 In voorbeeld 2.32 in hoofdstuk 2 werd een voorbeeld van de paradox van Simpson besproken: *de omkering van een vergelijking of een associatie, wanneer gegevens uit verschillende groepen tot één groep gecombineerd worden*. Het ging in het voorbeeld om twee ziekenhuizen A en B en het wel of niet overlijden van patiënten die een operatie ondergingen. Hier volgen de gegevens:

	Ziekenhuis A	Ziekenhuis B
Overleden	63	16
Niet overleden	2037	784
Totaal	2100	800

En hier volgen de gegevens nogmaals, maar nu onderverdeeld naar de conditie waarin de patiënten zich bevonden toen zij werden opgenomen.

	Goede conditie	
	Ziekenhuis A	Ziekenhuis B
Overleden	6	8
Niet overleden	594	592
Totaal	600	600

	Slechte conditie	
	Ziekenhuis A	Ziekenhuis B
Overleden	57	8
Niet overleden	1443	192
Totaal	1500	200

(a) Gebruik logistische regressie om de kansverhouding (odds) op overlijden te modelleren met ziekenhuis als verklarende variabele. Geef een overzicht van de resultaten en geef een 95%-betrouwbaarheidsinterval voor de odds ratio van ziekenhuis A ten opzichte van ziekenhuis B.

(b) Voer de analyse onder (a) opnieuw uit, maar gebruik nu ziekenhuis en conditie van de patiënt als verklarende variabelen. Geef een overzicht van de resultaten en geef een 95%-betrouwbaarheidsinterval voor de odds ratio van ziekenhuis A ten opzichte van ziekenhuis B.

(c) Verklaar de paradox van Simpson in termen van de resultaten in (a) en (b).

Noten

1. Er zijn ook logistische modellen ontwikkeld voor het algemene geval waarin de responsvariabele meer dan twee waarden kan aannemen. Deze modellen zijn aanzienlijk ingewikkelder en vallen buiten het kader van dit boek. Zie voor meer informatie over logistische regressie: A. Agresti, *An Introduction to Categorical Data Analysis*, Wiley, New York, 1996 en D.W. Hosmer en S. Lemeshow, *Applied Logistic Regression*, Wiley, New York, 1989.
2. Het begrip *odds* wordt soms met *kans* vertaald, maar dit is onjuist omdat de kansverhouding waarden groter dan 1 kan aannemen. [vert.]
3. Dit voorbeeld is ontleend aan een klassieke tekst geschreven door een tijdgenood van R.A. Fisher, de geleerde die veel van de fundamentele ideeën ontwikkelde op het gebied van statistische inferentie, die we heden ten dage nog steeds gebruiken. Zie: D.J. Finney, *Probit Analysis*, Cambridge University Press, Cambridge, 1947. Hoewel we er in de analyse geen gebruik van maakten, is het van belang op te merken dat het experiment een controlegroep bevatte die geen insecticide kreeg toegediend. In deze groep gingen geen luizen dood. We omschrijven hier de respons met ‘dood’. In het boek van Finney wordt deze groep omschreven met ‘kennelijk dood, stervend of zo sterk aangetast dat het niet in staat is meer dan een paar stappen te doen’. Dit is een vroeg voorbeeld van de nauwkeurigheid die nodig is bij het definiëren van variabelen ten behoeve van statistische analyses. Een insect dat ‘niet in staat is meer dan een paar stappen te doen’ zal hoogstwaarschijnlijk maar weinig schade doen aan een chrysant!