

waar men determinanten van gezondheid en ziekte bestudeert, is de uitkomstvariabele vaak: wel of niet ziek. Men vergelijkt dan de kans op ziekte tussen personen met, en personen zonder bepaalde kenmerken. Uitgaande van het voorbeeld van de vorige keer kunnen we onze uitkomstvariabele (het cholesterolgehalte) dichotomiseren in het al dan niet aanwezig zijn van hypercholesterolemie (een te hoog cholesterolgehalte), waarbij we de grens van wat we (te) hoog vinden zelf kunnen definiëren. We vergelijken vervolgens het vóórkomen (de prevalentie) van hypercholesterolemie bij actieven en niet-actieven.

Net als het lineaire regressiemodel bestaat ook het logistisch model uit een lineaire combinatie van parameters en variabelen. Het linkerlid van de vergelijking is nu echter wat anders, namelijk de logaritme van de odds op ziek zijn: $\ln(\text{odds}) = \alpha + \beta_1 x_1$. Over de odds hebben we het al eerder gehad: dit is de kans op een gebeurtenis, gedeeld door de complementaire kans (dat die gebeurtenis niet optreedt). In het model met één x_1 -variabele geeft de β_1 -coëfficiënt het 'ruwe' effect van de primaire determinant aan. In het 'multivariate' geval (ook wel 'multivariabele of multi-pele regressie' genoemd, met x_1 t/m x_n) is het effect 'gecorrigeerd' voor de andere variabelen in het model. De interpretatie van de β_1 -coëfficiënt is vergelijkbaar met die bij lineaire regressie: namelijk de verandering in de afhankelijke variabele, gegeven één eenheid verandering in de onafhankelijke variabele x . Alleen is het zo dat de β_1 -coëfficiënt in het logistisch model niet direct interpreteerbaar is als een verschil tussen gemiddelden of prevalenties, maar eerst nog 'verbouwd' moet worden. In dit geval is de β_1 -coëfficiënt het verschil in $\ln(\text{odds})$ als x_1 een eenheid verschilt, dus bijvoorbeeld 1 in plaats van 0 is. Omdat aftrekken van logaritmen gelijk staat aan het nemen van de logaritme van het quotiënt, is de β_1 -coëfficiënt gelijk aan $\ln(\text{Odds Ratio})$, zoals de afleiding in schema 1 laat zien. Hieruit volgt dat we de β_1 -coëfficiënt tot de e -macht moeten verheffen om de odds ratio (OR) zelf te krijgen.

Schema 1 Afleiding betekenis β_1 -coëfficiënt in de lineaire en de logistische regressievergelijking

	Lineaire regressievergelijking	Logistische regressievergelijking
Regressievergelijking (algemeen)	$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots$	$\ln(\text{odds}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots$
Regressievergelijking voor blootgestelden: $x_1=1$	$y_1 = \alpha + \beta_1 + \beta_2 x_2 + \dots$	$\ln(\text{odds}_1) = \alpha + \beta_1 + \beta_2 x_2 + \dots$
Regressievergelijking voor niet-blootgestelden: $x_1=0$	$y_0 = \alpha + 0 + \beta_2 x_2 + \dots$	$\ln(\text{odds}_0) = \alpha + 0 + \beta_2 x_2 + \dots$
Verskil tussen wel- en niet-blootgestelden met overigens dezelfde kenmerken	$y_1 - y_0 = (\alpha + \beta_1 + \beta_2 x_2 + \dots) - (\alpha + \beta_2 x_2 + \dots) = \beta_1$	$\ln(\text{odds}_1) - \ln(\text{odds}_0) = (\alpha + \beta_1 + \dots) - (\alpha + \dots) = \beta_1$. Voorts is: $\ln(\text{odds}_1) - \ln(\text{odds}_0) = \ln(\text{odds}_1 / \text{odds}_0) = \ln(\text{Odds Ratio})$. Hieruit volgt dat: $\ln(\text{Odds Ratio}) = \beta_1$, en dus dat $\text{OR} = \exp(\beta_1)$. *
Interpretatie van een schatting (steekproefuitkomst) van een β_1 -coëfficiënt	Verskil in gemiddelde waarde van de afhankelijke variabele (y) tussen wel- en niet-blootgestelden.	Na verheffing tot de e -macht van de gevonden coëfficiënt hebben we een schatting van de Odds Ratio. Alleen indien de prevalentie laag is (<10%), is deze bij benadering gelijk aan de relatieve 'kans' op ziekte voor blootgestelden ten opzichte van niet-blootgestelden.

* De notatie $\exp(\beta)$ betekent hetzelfde als e^β .

INTERPRETATIE

Het logistische regressiemodel geeft de mogelijkheid om de OR als effectmaat te corrigeren voor confounders. Wanneer de primaire determinant ook een dichotome variabele is en de absolute kansen niet groter dan 10% zijn, dan is de interpretatie redelijk gemakkelijk. De OR is dan namelijk een goede schatting van het relatieve risico op ziek zijn voor blootgestelden, c.q. degenen met een bepaald kenmerk ($x_1 = 1$), ten opzichte van niet-blootgestelden, c.q. degenen zonder dat kenmerk ($x_1 = 0$).

Zouden in ons voorbeeld de lichamelijk niet-actieven ($x_1 = 1$) een kans van 9% hebben op hypercholesterolemie en de wel-actieven ($x_1 = 0$) een kans van 6% dan zou de ongecorrigeerde β_1 -coëfficiënt 0,438 zijn, hetgeen na logaritmische transformatie een OR van 1,549 oplevert (immers $1,549 = (9/91)/(6/94) = \text{odds ratio}$). Na correctie voor leeftijd en eventueel andere confounders kan de β_1 -coëfficiënt veranderen en bijvoorbeeld 0,095 worden, hetgeen overeenkomt met een OR van 1,10. We mogen dan concluderen dat de kans op een verhoogd cholesterol voor niet-actieven van dezelfde leeftijd niet ca 1,5 keer zo groot, maar slechts 1,1 maal zo groot is. Let op dat deze benadering slechter wordt bij prevalenties groter dan 10% (zie aflevering 4, toelichting bij schema 3).

Net als bij het lineaire model is het belangrijk om aan te geven hoe de onafhankelijke variabelen (primaire determinant en confounders) gedefinieerd zijn. In ons voorbeeld was de primaire determinant, net als de uitkomstmaat, een dichotome variabele met de waarden 0 en 1 voor respectievelijk af- en aanwezigheid van het kenmerk 'niet lichamelijk actief zijn'. Hadden we echter een continue variabele als determinant gehad, bijvoorbeeld leeftijd in hele jaren, dan was de interpretatie van de e -macht van de β -coëfficiënt het relatief risico *per leeftijdsjaar* geweest.

MODELKEUZE

Wat zijn nu de verschillen tussen een gewoon lineair model en het logistische model? Zoals gezegd wordt voor de uitkomstvariabele een andere kansverdeling verondersteld: de normale verdeling voor een continue variabele (bij gewone regressie) versus de binomiale verdeling voor een dichotome variabele (bij logistische regressie). Maar ook de resulterende effectmaat is verschillend. We hebben al eerder betoogd dat voor de praktijk absolute risico's en risicoverschillen vaak veel meer zeggingskracht hebben dan een relatieve maat zoals de odds ratio of de prevalentie ratio. Dit zou een goede reden kunnen zijn om toch de voorkeur te geven aan gewone regressie boven logistische regressie. De combinatie van een binaire uitkomstmaat (wel/niet ziek), de neiging om een verband in een relatieve maat uit te drukken en de beschikbaarheid van het logistisch model hebben ertoe geleid dat ook in artikelen over dwarsdoorsnede-onderzoek vaker gecorrigeerde OR's dan bijvoorbeeld gecorrigeerde verschillen in percentages worden gepresenteerd. Auteurs zouden in zo'n geval ook moeten overwegen of zij niet meer geïnteresseerd zijn in verschillen tussen prevalenties, en daarom beter voor gewone lineaire regressie kunnen kiezen. Immers, een percentage is niets anders dan het gemiddelde (dé uitkomstmaat van een gewone lineaire regressie) van een dichotome variabele, met de waarden 0 en 1. Weliswaar is de uitkomstvariabele niet normaal verdeeld, maar dat maakt niet uit voor de interpretatie van de β -coëfficiënt. Waar we wel voorzichtig mee moeten zijn is met de p -waarden en het betrouwbaarheidsinterval die het model geeft.

Deze zijn niet geheel correct omdat de verdeling van de (binaire) variabele afwijkt van wat verondersteld wordt in het gekozen model (namelijk een continue afhankelijke variabele die normaal verdeeld is).

Hadden we in het voorbeeld over hypercholesterolemie dus niet het logistisch model gebruikt maar het gewone lineaire model, dan zou de ruwe β_1 -coëfficiënt 0,03 (3%) zijn geweest en direct, zonder 'verbouwing', interpreteerbaar als het verschil in percentage personen met hypercholesterolemie tussen niet-actieve en wel-actieve personen. Na toevoeging van leeftijd als confounder zou de β_1 -coëfficiënt bijvoorbeeld 0,007 (0,7%) kunnen zijn geworden. Dit is dan het voor leeftijd gecorrigeerde verschil in percentage personen met hypercholesterolemie tussen niet- en wel-actieven.

Bij dezen roepen we auteurs op om in de toekomst wat vaker het lineaire regressiemodel te gebruiken als verschillen in percentages relevant zijn, bijvoorbeeld in de tijd, tussen (etnische) groepen of geografische gebieden. Zo krijgen lezers, zeker indien het absolute percentage van de referentiecategorie is gegeven, direct een goed beeld van de 'maatschappelijke relevantie' van de gegeven verschillen.

CORRESPONDENTIEADRES

Dr. Willy-Anne H.J. van Stiphout, vrijgevestigd sociaal geneeskundige en docent epidemiologie, www.decijfersdebaas.nl of e-mail: stip.info@xs4all.nl

A LETTER FROM WASHINGTON, DC

Medicaid hervormingen: flexibiliteit om te komen tot betaalbaarheid

Een collega op de gang werkt als vrijwilliger. Ze deelt regelmatig 's avonds en 's nachts condooms uit aan mensen die behoren tot risicogroepen. Voor Nederlandse begrippen een weinig schokkende activiteit, maar hier in de VS vindt ongeveer de helft van de bevolking het nogal ongewoon.

Jenny, want zo heet ze, werkt aan het dossier Medicaid. Dat is de publieke verzekering voor Amerikanen met een laag inkomen die 55 miljoen mensen voorziet van gezondheidszorg en langdurige zorg. Om het beeld compleet te maken: van de 300

miljoen burgers regelen ongeveer 160 miljoen hun zorgverzekering via een werkgever, 43 miljoen ouderen zijn verzekerd via Medicare, een paar miljoen mensen hebben individuele dekking en er zijn 46 miljoen mensen onverzekerd – al maken sommige onverzekerden wel weer gebruik van Medicaid.

Als we inzoomen, dan zitten in Medicaid enerzijds erg veel kinderen (27 miljoen) en verder de meest arme en zieke mensen van het land, inclusief gehandicapten en chronisch zieken. Armen kunnen geen private verzekering veroorloven, zieken worden gewoonweg geweigerd (risicoselectie). De totale uitgaven van het programma bedragen \$300 miljard, een zesde van de totale zorguitgaven in de VS. De federale overheid stelt basisvereisten op wie in aanmerking komen en financiert gemiddeld 57%. De staten – die het programma daadwerkelijk uitvoeren – financieren de rest en zij hebben de mogelijkheid de dekking uit te breiden: meer mensen of een breder pakket. Er is daardoor een grote variatie tussen staten onderling. Nog