Multivariate analyse

Logistische regressie is een statistische analysemethode waarmee het gelijktijdige effect van een aantal factoren op een uitkomst kan worden gemeten. De methode wordt gebruikt om simultaan rekening te houden met een groot aantal expositiefactoren en mogelijke confounders. Logistische regressie wordt vooral gebruikt bij observationele onderzoeken. Bij het onderzoeken van interventies is en blijft randomisering de gouden standaard voor het opheffen van confounding.

Simone Buitendijk

In deel IV van deze serie zijn de verstorende variabelen (confounders) aan de orde geweest.[1] De beste manier om geen last te hebben van confounders, bestaat uit het randomiseren van de onderzoeksgroep; het lot bepaalt dan welk deel van de groep wordt blootgesteld aan een bepaalde factor, en welk deel niet. Veel 'blootstellingen' zijn echter niet geschikt voor onderzoek in een gerandomiseerde opzet. Zo zou het niet erg ethisch zijn (en dus ook niet erg haalbaar) om voor een onderzoek naar het effect van alcoholgebruik door zwangeren een gerandomiseerde opzet toe te passen; dergelijke exposities kunnen alleen in een observationele onderzoeksopzet worden onderzocht. In zo'n opzet is het veel lastiger om te corrigeren voor het effect van confounders, vooral als het aantal mogelijke confounders groot is.

Mw.dr. S.E. Buitendijk,
TNO Preventie en Gezondheid,
Postbus 2215,
2301 CE Leiden;
se.buitendijk@pgo.tno.nl

Mogelijke confounders

In het voorbeeld van alcoholgebruik tijdens de zwangerschap is roken een voor de hand liggende confounder. Als dat de enige zou zijn, zou het voor de hand liggen om de onderzoeksgroep te stratificeren naar wel/niet roken. Je kan dan het effect van alcoholgebruik onderzoeken in een groep niet-rooksters en in een groep vrouwen die allen ongeveer evenveel roken.

Het wordt ingewikkelder als je daarnaast ook rekening wilt houden met het mogelijk verstorende effect van drugsgebruik, voedingsgewoonten, leeftijd van de moeder, etniciteit en sociaal-economische status – stuk voor stuk factoren die te maken kunnen hebben met zowel rookgedrag als negatieve zwangerschapsuitkomsten. Nog veel lastiger wordt het als je eigenlijk ook zou willen weten (in hetzelfde onderzoek) wat het effect is van drugsgebruik of voedingsgewoonten op de uitkomst.

Als je alleen het effect van alcoholgebruik zou willen weten, kan je in theorie het effect van wel of geen alcoholgebruik nagaan binnen een homogene subgroep met bijvoorbeeld de volgende kenmerken: niet roken, geen drugsgebruik, gezond dieet, jonger dan 30 jaar, van Nederlandse afkomst en hoog opgeleid. Vervolgens kan je hetzelfde doen binnen alle andere homogene subgroepen. Het effect van alcoholgebruik in al die subgroepen (strata) kan je dan als het ware bij elkaar optellen om zo een *totaaleffect* te krijgen van alcoholgebruik op de uitkomst die je onderzoekt. Alleen weet je dan nog niet wat het *afzonderlijke* effect is van bijvoorbeeld drugsgebruik of voedingsgewoonten.

Voor dit soort situaties wordt vaak multivariate analyse gebruikt: in de eerste plaats omdat dit gemakkelijker is dan eindeloos stratificeren op alle verschillende niveaus van de confounders, en in de tweede plaats omdat je met behulp van multivariate analyse ook het afzonderlijke effect van een aantal andere expositiefactoren kan nagaan.

Multivariate analyse

Multivariate analyse is een statistische techniek die wordt gebruikt na afloop van de dataverzameling. Het is een manier om in de analysefase van het onderzoek te kijken naar het effect op een uitkomst van een aantal mogelijke factoren tegelijk.

De uitkomst waarin de onderzoeker geïnteresseerd is, wordt wel afhankelijke variabele genoemd (*dependent variable*), en de expositiefactoren onafhankelijke variabelen (*independent variables*). Wat het statische programma in feite doet, is het effect van alle factoren laten zien, rekening houdend met het effect van de andere factoren.

Voor bovenstaand voorbeeld kijkt het programma dus naar het effect van alcoholgebruik, gecorrigeerd voor roken en alle andere factoren in het model. Als bijvoorbeeld blijkt dat moeders die veel alcohol gebruiken een 1,5 maal grotere kans hebben op een kind met groeivertraging, dan is daarbij al rekening gehouden met het feit dat deze vrouwen mogelijk ook veel roken. Roken kan dan niet het effect van alcohol 'wegverklaren'.

Voorwaarde is dat alle expositiefactoren en mogelijke confounders gemeten zijn. Voor een confounder die niet gemeten is, kan niet worden gecorrigeerd. Als dus geen informatie aanwezig is over voedingsgewoonten, kan daar in de analyse geen rekening mee worden gehouden. Ook onvolledig of verkeerd gemeten confounders kunnen de conclusies in belangrijke mate verstoren. Multivariate analyse is dus geen gemakkelijk trucje om te zorgen dat onderzochte groepen volstrekt vergelijkbaar zijn.

Logistische regressie

Er zijn diverse vormen van multivariate analyse. De bekendste zijn multiple lineaire regressie en logistische regressie.

- *Multiple lineaire regressie* wordt gebruikt om het effect van een aantal onafhankelijke variabelen op een continue uitkomstvariabele weer te geven. De naam geeft aan dat de uitkomstvariabele in die analyse een *lineaire* functie is van de expositievariabelen.
- *Logistische regressie* wordt gebruikt wanneer de periode van observatie gelijk is voor alle deelnemers aan het onderzoek en de uitkomst *dichotoom* is (ja/nee, levend/dood, groeivertraagd/niet groeivertraagd). Een logistisch regressiemodel is geen lineair model, maar een log-lineair model (vandaar de naam). De kans op de uitkomst wordt in dit model uitgedrukt in een odds ratio, die een functie is van alle expositievariabelen bij elkaar. Dit risico ligt altijd tussen 0 en 1.

Veel uitkomsten die in de verloskunde een rol spelen, zijn dichotoom.

De expositievariabelen die bij logistische regressie worden gebruikt kunnen continue of categorische variabelen zijn, mits de uitkomst dichotoom is. Het model geeft als het ware het effect van elke factor afzonderlijk op de kans op de verkregen uitkomst, na correctie voor alle andere factoren. Logistische regressie laat dus ook zien welke factor het meest bepalend is voor een bepaald ziekterisico.

Ook kunnen in een logistisch regressiemodel *interactietermen* worden opgenomen. Dat is aan de orde als het effect van een bepaalde expositie heftiger is in combinatie met een andere expositie. Wat dat betekent, blijkt uit het volgende voorbeeld:

1 Roken verdubbelt de kans op groeivertraging (odds ratio voor rooksters 2,0 ten opzichte van niet-rooksters).
2 Een bepaalde genetische factor verdubbelt de kans op groeivertraging.
3 Een vrouw die én rookt (1) én de desbetreffende genetische factor heeft (2), heeft een odds ratio van 6,0 op groeivertraging.

Dit is dus meer dan het effect van de twee afzonderlijke risicofactoren bij elkaar opgeteld of met elkaar vermenigvuldigd. In een logistisch regressiemodel kan voor de aanwezigheid van dergelijke interactie-effecten worden getoetst.

mulitvariate analyse is geen gemakkelijk trucje

Tabel

Niet-gecorrigeerde en gecorrigeerde odds ratio's voor zeer laag geboortegewicht (<1500 g) en laag geboortegewicht (1500-2499 g) in relatie tot de maternale leeftijd (n=171.619)

Maternale leeftijd (jaren)	<1500 g (niet gecorr. oddsratio)	<1500 g (gecorr. oddsratio. 95%-betrouwbaar-heidsinterval)*	1500-2499 g (niet-gecorrigeerde oddsratio)	1500-2499 g (gecorr. oddsratio. 95%-betrouwbaar-heidsinterval)*
20-24	1.0	1.0	1.0	1.0
25-29	0.9	1.0 (0.9-1.1)	1.0	1.0 (1.01-1.1)
30-35	1.2	1.2 (1.03-1.5)	1.02	1.4 (1.3-1.5)
35-39	2.2	1.9 (1.5-2.4)	1.6	1.7 (1.5-1.9)
≥40	2.5	1.8 (1.04-3.0)	2.0	2.0 (1.5-2.5)

* Gecorrigeerd voor opleiding van de moeder, samenwoning met de vader van het kind, roken (moeder), infertiliteit, hypertensieve aandoeningen, diabetes en bloedverlies ante partum.

Bron De tabel is ontleend aan *Cnattingius et al.* en *Zaadstra.*[2,3]

Voorbeeld

De *tabel* laat het resultaat zien van een logistische regressie-analyse. De Zweedse onderzoekers *Cnattingius et al.* onderzochten het effect van de leeftijd van de moeder op de kans op een laag geboortegewicht van de baby. Daarbij werd rekening gehouden met een aantal mogelijke confounders: opleiding moeder, samenwonen met de vader van het kind, roken moeder, infertiliteit, hypertensieve aandoeningen, diabetes mellitus en bloedverlies ante partum. Daarbij gaat het dus om factoren die zowel kunnen verschillen tussen oudere en jongere moeders als het risico op groeivertraging kunnen beïnvloeden.

In de eerste kolom zien we het niet-gecorrigeerde risico op een kind met een geboortegewicht <1500 gram. Daaruit blijkt bijvoorbeeld dat het risico voor vrouwen tussen de 35 en 39 jaar 2,2 maal zo groot is als dat voor vrouwen tussen 20 en 24 jaar (de referentiecategorie). In de bepaling van deze odds ratio is echter geen rekening gehouden met de mogelijke confounders. Stel dat oudere moeders meer roken, dan kan het zijn dat niet de leeftijd, maar het roken het verhoogde risico veroorzaakt. Het kan ook zijn dat als oudere moeders juist minder roken, de eigenlijke odds ratio groter is dan 2,2. Wat dus meer informatie geeft is, de kolom met gecorrigeerde odds ratio's.

In de tweede kolom zien we dat, als rekening wordt gehouden met alle mogelijke confounders, het relatieve risico op een kind met een geboortegewicht <1500 gram 1,9 bedraagt. Anders geformuleerd: als je de oudere groep 'gelijk maakt' aan de jongere wat betreft opleiding, roken, infertiliteit en zelfs hypertensie en diabetes mellitus, dan bestaat nog steeds een bijna 2 keer zo hoog risico.

Stapsgewijze logistische regressie

Je kunt je afvragen of het hier wel nodig is om ook voor hypertensie en diabetes mellitus te corrigeren. Dit kunnen immers ook factoren zijn die een eventueel verhoogd risico bij oudere moeders verklaren. Dan zijn het meer factoren in het causale pad dan confounders per se. Blijkbaar zijn er dus nog andere factoren die niet in de tabel zijn opgenomen, die het hogere risico bij oudere vrouwen zouden verklaren. We stuiten hier op een belangrijk probleem: als in een logistisch regressiemodel te veel factoren tegelijk worden opgenomen, kan daardoor het effect van de belangrijkste factor verdwijnen.

Een manier om dit probleem op te lossen en om meer inzicht te krijgen in het effect van de verschillende expositiefactoren, is het gebruik van verschillende modellen, waarbij per model steeds een aantal expositiefactoren wordt toegevoegd. Dit heet *stapsgewijze* logistische regressie. In het voorbeeld zouden dan eerst alleen roken, opleiding en samenwoonsituatie in het model worden opgenomen ('echte' confounders); in een volgend model zouden bijvoorbeeld hypertensie en diabetes kunnen worden toegevoegd. In beide modellen kan je dan zien wat het effect is van de leeftijd van de moeder.

Van belang is dat de aantallen onderzochte personen groot genoeg zijn om een aantal confouders tegelijk te kunnen onderzoeken. Hoe groter het aantal expositiefactoren dat tegelijkertijd in een model wordt gestopt, des te groter in feite het aantal subgroepen dat wordt geanalyseerd. Per subgroep moeten dan wel voldoende gegevens beschikbaar zijn. ●

Literatuur

1 Buitendijk SE, De Miranda E. Confounding. Tijdschrift voor Verloskundigen 2001;26(7/8):597-601.
2 Cnattingius S, Forman MR, Berendes HW, Isotalo L. Delayed childbearing and risk of adverse perinatal outcome. JAMA 1992;268:886-90.
3 Zaadstra BM. Kind in de toekomst, kind van de rekening? Tijdschrift voor Verloskundigen 1999;24(2): 103-9.