Fixed versus Random

Bas Engel

Dit artikel is gebaseerd op een voordracht gehouden op de themadag "Fixed of random; pragmatisch of principieel?", georganiseerd door LBS, ANed en MBS. Papier is geduldiger dan transparanten. Hier en daar, en vooral in het uitgewerkte voorbeeld aan het eind, zijn daarom wat details toegevoegd. In grote lijnen is de tekst echter gelijk gebleven aan die van de voordracht.

De bedoeling was aan te geven dat de keuze voor het fixed of random opnemen van een factor in een model soms anders uitpakt dan je op het eerste gezicht zou verwachten. De reden hiervoor is dat naast de "conventionele" argumentatie (is er sprake van een onderliggende populatie en zijn uitspraken op populatieniveau gewenst?) ook andere overwegingen een rol kunnen spelen. Overwegingen die te maken hebben met de specifieke probleemstelling en randvoorwaarden met betrekking tot de structuur van de data.

Bas Engel, DLO-Groep Landbouwwiskunde (GLW-DLO),
Postbus 100, 6700 AC Wageningen
Tel: 08370 74687. E-mail: B.Engel@SC.AGRO.NL

Met dank aan Henri Woelders (ID-DLO, Zeist) voor het gebruik van de data over beoordeling van spermakwaliteit, aan Willem Buist (ID-DLO, Zeist) voor assistentie bij de analyse van deze data en aan Saskia Burgers (GLW-DLO) voor kritisch commentaar op een eerdere versie van de tekst.

1 Inleiding

Het laatste hoofdstuk van "Linear models" van S.R. Searle (1971) is lange tijd de enig beschikbare samenvatting van technieken voor het schatten van variantiecomponenten geweest. In dit boek zegt Searle het volgende over fixed en random effecten:

> *"...when inferences are going to be confined to the effects in the model, the effects are considered fixed; and when inferences will be made about a population of effects from which those in the data are considered to be a random sample then the effects are considered as random..."*

Wanneer de niveaus van een factor (ik beperk me gemakshalve in dit verhaal hoofdzakelijk tot factoren) beschouwd kunnen worden als trekkingen uit een populatie, en de interesse gaat uit naar karakteristieken van die populatie, dan wordt de factor als random opgenomen. Het gaat dan niet om de niveaus in het experiment, maar om de kansverdeling waaruit die niveaus afkomstig zijn. In veel toepassingen wordt aangenomen dat de niveaus uit een normale verdeling worden getrokken. In dat geval gaat het om de variantie van die normale verdeling.

De vraag of, bijvoorbeeld, individuen random of fixed moeten worden opgenomen kun je vaak beantwoorden door de volgende vraag te stellen:

> zou het experiment herhaald worden, betreft die herhaling dan dezelfde individuen, of andere individuen?

Gaat het om andere individuen in de (vaak denkbeeldige) herhaling, dan is er doorgaans sprake van een populatie van individuen. In dat geval is het meestal terecht individuen random te nemen. Hier lijkt inderdaad uit te volgen dat individuen zelf niet van belang zijn. Toch is dat niet altijd waar, zoals we dadelijk zullen zien.

In "Variance components" van Searle, Casella en McCulloch (1992) staat over fixed en random effecten ondermeer het volgende:

> *"...thus the distinction between fixed and random effects centers on whether we are willing to assume that the levels of a factor are sampled randomly from a distribution, not whether we are specifically interested in the levels of that factor..."*

en

> *"...thus it is that the situation to which a model applies is the deciding factor in determining whether effects are to be considered as fixed or random..."*

Andere uitspraken over fixed en random effecten in het boek sluiten wel zo'n beetje aan bij de "definitie" van 20 jaar eerder. Maar in aanvulling daarop is duidelijk wat water bij de wijn gedaan. Kennelijk kunnen we nu een factor als random beschouwen en toch ook in de niveaus van die factor geïnteresseerd zijn. We zullen twee voorbeelden bekijken waar dat inderdaad het geval is: één uit de veefokkerij en één uit de geostatistiek. Met name het laatste voorbeeld geeft aanleiding tot enige controverse. Maar ook het fokkerij model zal misschien niet ieders eerste keus zijn. We zullen zien dat de specifieke probleemstelling de keuze voor random effecten motiveert, ondanks het feit dat het idee van een onderliggende populatie een vrij gekunsteld karakter heeft.

Alvorens naar de voorbeelden te gaan, kijken we eerst naar verschillen tussen het schatten of voorspellen van effecten wanneer deze fixed of random worden opgenomen. Na de

voorbeelden gaan we hier nog kort op door en geven we wat verbindingen met ridge regressie, Bayesiaanse statistiek en het gebruik van penalty functies aan. We eindigen met een praktijkvoorbeeld. Hoewel we ons in dit artikel voornamelijk tot normale verdelingen beperken, gaan de mogelijkheden thans al veel verder. Het praktijkvoorbeeld illustreert dit, daar het hier een analyse van percentages betreft.

2 Het IQ van Ronnie Fysher

Dit is een voorbeeld uit Searle et al. (1992). Het is wat flauw, maar geschikt als eenvoudig illustratiemateriaal. Ronnie Fysher is een eerstejaars student. Hij heeft een aantal toetsen afgelegd. Het resultaat y van een toets kan beschouwd worden als een schatting van Ronnies IQ u:

$$y = u + e \tag{1}$$

Voor het gemak nemen we aan dat $e \sim N(0, \sigma_e^2)$. Toetsen worden onafhankelijk verondersteld. Een voor de hand liggende schatter voor u op basis van n toetsen is het gemiddelde $\bar{y}$:

$$E(\bar{y}) = u \quad \text{met} \quad E(u-\bar{y})^2 = Var(\bar{e}) = \sigma_e^2 / n.$$

IQ's zijn misschien niet bedacht om tussen mensen te discrimineren, maar ze worden er wel voor gebruikt. We zouden Ronnie kunnen beschouwen als één element uit een populatie van eerstejaars studenten. Stel dat we (met voldoende nauwkeurigheid) kunnen stellen dat de verdeling van het IQ in deze populatie normaal is met gemiddelde m en variatie σ_u^2. Neem nu eens aan dat u in (1) random is. In dat geval beschikken we naast $\bar{y}$ over een tweede schatting voor u, namelijk: m met $E(u-m)^2 = Var(u) = \sigma_u^2$. We combineren nu de twee schatters, waarbij we wegen met de inverses van de varianties σ_e^2 / n en σ_u^2:

$$\tilde{u} = [\sigma^{-2}_u / (\sigma^{-2}_u + n\sigma^{-2}_e)]\, m + [(n\sigma^{-2}_e) / (\sigma^{-2}_u + n\sigma^{-2}_e)]\, \bar{y}. \tag{2}$$

Voor grote n is het verschil tussen $\tilde{u}$ en $\bar{y}$ klein. Maar, wanneer we over minder informatie beschikken, zal $\tilde{u}$ meer op het overall gemiddelde m gaan lijken. $\tilde{u}$ is de *Best Linear Unbiased Predictor* (BLUP). *Best* in de zin van:

$$E(\tilde{u}-u)^2 \text{ minimaal}$$

en *unbiased* in de zin van:

$$E(\tilde{u}-u) = 0.$$

Omdat $\tilde{u}$ een stukje in de richting van m gaat, en het verschil met m als het ware is gekrompen, is $\tilde{u}$ een voorbeeld van een *shrinkage estimator*. Merk op dat voor gegeven u:

$$E(\tilde{u}-u \mid u) \neq 0,$$

in die zin is de BLUP niet zuiver. Het predicaat *best* beperkt zich tot lineaire functies van de data, onder aanname van random u. $\tilde{u}$ is de regressie van u op de waarnemingen y.

Onder normaliteit is $\tilde{u} = E(u \mid \bar{y})$. Veel meer over BLUP vindt u in een mooi artikel van Robinson (1991).

Of Ronnies IQ nu moet worden bepaald door $\bar{y}$ of door $\tilde{u}$ hangt van de context af. Voor selectiedoeleinden is $\tilde{u}$ een geschikte grootheid, zoals we straks zullen zien. Althans vanuit het oogpunt van degene die de selectie moet uitvoeren. Ronnie heeft in feite geen boodschap aan de resultaten, vooral de slechte resultaten, van medestudenten, en vindt $\bar{y}$ mogelijk een geschiktere schatting.

3 Melkproduktiegegevens en stiervaders

Hieronder is een dataset(je) weergegeven. De (gefingeerde) gegevens zijn afkomstig uit Robinson (1991). Het gaat om melkproduktiegegevens van dochters van stiervaders (A, B, C en D) op 3 bedrijven (1, 2 en 3). In werkelijkheid kan zo'n dataset miljoenen waarnemingen betreffen, hier zijn het er maar 9. Het gaat er om de beste vaderdieren te selecteren.

Bedrijf	Vader	Productie dochter
1	A	110
1	D	100
2	B	110
2	D	100
2	D	100
3	C	110
3	C	110
3	D	100
3	D	100

Het model luidt als volgt:

$$y_{ijk} = b_i + u_j + e_{ijk},$$

waarin y de produktie, b het bedrijfsniveau, u de bijdrage van de stiervader en e de restbijdrage weergeeft. Stel: $e_{ijk} \sim N(0, \sigma_e^2)$. Indices i, j en k nummeren de bedrijven, vaders en dochters binnen vaders. De e_{ijk} worden onderling onafhankelijk verondersteld.

Nemen we vaders fixed, dan geeft variantie-analyse (met als bijvoorwaarde dat vaderbijdragen tot 0 optellen):

$$\hat{u}_A = 2.50,\ \hat{u}_B = 2.50,\ \hat{u}_C = 2.50 \text{ en } \hat{u}_D = -7.50.$$

Voor de bedrijven volgt: $\hat{b}_1 = \hat{b}_2 = \hat{b}_3 = 107.5$. De vergelijking tussen de vaders onderling en tussen de bedrijven loopt in beide gevallen via vader D.

Nemen we vaders random op, met $u_j \sim N(0, \sigma_u^2)$ en u_j's onderling onafhankelijk en nemen we aan dat $\sigma_u^2 / \sigma_e^2 = 0.1$, dan volgt voor de BLUPs voor de vaders:

$$\tilde{u}_A = 0.40,\ \tilde{u}_B = 0.52,\ \tilde{u}_C = 0.76 \text{ en } \tilde{u}_D = -1.67$$

en voor de bedrijven: $\hat{b}_1 = 105.64$, $\hat{b}_2 = 104.28$, $\hat{b}_3 = 105.46$ (over de berekening meer in § 5).

Voor selectiedoeleinden heeft BLUP een aantal aantrekkelijke eigenschappen. Laten we voor het gemak even aannemen dat alle bedrijfsniveaus gelijk aan 0 zijn. Niet alleen het dochtergemiddelde maar ook de bijbehorende variantie σ_e^2 / n, waarin n het aantal dochters is, speelt een rol. Wanneer het dochtergemiddelde van een stier een beetje lager is dan van een andere stier, maar de bijbehorende standaardafwijking is veel lager, dan zouden we geneigd zijn op zeker te spelen en voor het nauwkeurigste gemiddelde te kiezen. De BLUPs zijn zowel op de dochtergemiddelden als op de bijbehorende standaardafwijkingen gebaseerd. Een stier met een gering aantal dochters zal een BLUP hebben die dichter bij 0 ligt dan een stier met een vergelijkbaar dochtergemiddelde $\bar{y}$, maar gebaseerd op een groter aantal dochters:

$$\tilde{u} = [\sigma_u^2 / (\sigma_u^2+\sigma_e^2/n)]\ \bar{y},$$

zoals volgt uit (2) voor $m = 0$. Voor een groot aantal dochters ($n \rightarrow \infty$) geldt: $\tilde{u} \approx \bar{y}$. Voor $n = 0$ volgt dat $\tilde{u} = 0$. Het dochtergemiddelde van vaders A, B en C is gelijk. De BLUP voor vader C valt hoger uit dan voor vaders A en B, omdat C een dochter meer heeft. Formele argumenten voor het random opnemen van de stieren worden gegeven in Cochran (1951), Searle (1974) en Portnoy (1982).

Er is in de loop van de tijd een tweede argument voor random genetische bijdragen bijgekomen. De dieren zijn onder toenemende selectie allang geen onafhankelijke trekkingen uit een kansverdeling meer. Maar, door de familieverbanden via covarianties tussen de verschillende u's op te nemen, worden de selectie-effecten op praktisch afdoende wijze weggenomen. Dit is het geval zolang alle in het selectieproces betrokken dieren met hun waarnemingen in de dataset zijn opgenomen. Uitspraken hebben dan betrekking op de uitgangspopulatie bij aanvang van het selectieproces.

Ondanks het feit dat het hier om een beperkt aantal vaderdieren gaat waarvan we het genetisch potentieel willen achterhalen en we in het geheel niet in andere vaderdieren zijn geinteresseerd, worden de vaderbijdragen toch random genomen. De probleemstelling (selectie) en randvoorwaarden (familieverbanden) zijn bepalend. Het is vooral het laatste punt, familieverbanden via covarianties, wat de link vormt met het volgende voorbeeld. Overigens wordt tegenwoordig in plaats van een *sire model* meestal een *animal model* gebruikt. In dit laatste model stelt u de random genetische bijdrage van bijvoorbeeld één van de dochters aan haar productie voor. Familieverbanden worden weer via covarianties meegenomen.

4 Op zoek naar delfstoffen

Voor dit voorbeeld maken we dankbaar gebruik van een artikel van de Gruijter en ter Braak (1990). Zie verder ook de Gruijter (1991). Stel u staat aan de rand van een groot terrein. Voor het gemak denken we dit terrein als een schaakbord ingedeeld in een zeer groot aantal vakken. In ieder vak bevindt zich in de bodem een hoeveelheid delfstof. In sommige vakken niets, in andere een beetje of misschien juist heel veel. Aan u om uit te zoeken hoeveel delfstof er per oppervlakte-eenheid (zeg maar per vak) in de grond zit. Daar kunt u exact achterkomen door in alle vakken een meting te verrichten. Dat is echter te duur. U mag maar een beperkt aantal vakken "openmaken" om te kijken wat er in zit. Hoe kiest u die vakken?

We kijken naar twee mogelijkheden voor een model op grond waarvan we een keuze voor de steekproef maken. Onder de grond ligt een "patroon" van delfstof. Dit patroon kunnen we fixed of random veronderstellen.

Bij een fixed patroon maken we in feite geen enkele aanname omtrent het patroon. Omdat we geen aanname maken, is er geen reden om voor een bepaalde systematische structuur voor de steekproef te kiezen. Kiezen we de vakjes in de steekproef aselect, dan zijn de bijbehorende waarnemingen onafhankelijk. Het gemiddelde is een zuivere schatter voor de hoeveelheid delfstof per oppervlakte-eenheid. Dit is de zogenaamde *design-based* aanpak.

Nemen we aan dat het patroon random is, dan moeten we een proces specificeren dat dit patroon en andere patronen kan genereren. Parameters van de kansverdelingen die daarbij een rol spelen moeten dan worden gekozen of geschat uit de data. Dit is de *model-based* aanpak. We kiezen de vakken nu niet willekeurig, maar juist systematisch op een manier die wordt gedicteerd door het gestelde model. Dit model heeft ontegenzeggelijk iets gekunstelds. Het gaat ons in feite uitsluitend om dit ene patroon en van een populatie van patronen is eigenlijk geen sprake. Zouden we toch besluiten om het gehele terrein af te graven dan weten we precies wat er in de grond zit, maar de model parameters voor het genereren van patronen zijn nog steeds niet exact bekend.

Toch lijkt er een voorkeur voor de model-based aanpak te zijn. Een voorkeur die gepaard lijkt te gaan met de nodige misverstanden omtrent de kwaliteiten van de design-based oplossing. De Gruijter en ter Braak geven een aantal citaten en dit is er één van (Barnes, 1988):

> *"...The classic development of nonparametric tolerance intervals begins with an assumption of independent identically distributed random variables. This is unrealistic for the geologic environment - in general, geologic site characterization data are not independent...".*

Onafhankelijkheid onder het fixed patroon is géén aanname, maar een gevolg van de aselecte keuze van de meetplaatsen. Barnes kiest voor een model-based benadering en bekritiseert van daar uit de design-based aanpak. De twee benaderingen liggen echter op een verschillende "golflengte" en kriteria geldig voor random patronen, zijn niet noodzakelijk van toepassing binnen een model met een fixed patroon. Het misverstand komt waarschijnlijk voort uit voorkennis die niet in het fixed maar wel in het random model is opgenomen. We weten dat wanneer zich op een bepaalde plaats veel delfstof bevindt er op naburige plaatsen ook nog wel het een en ander in de grond zal zitten. Met een random patroon kunnen we daar rekening mee houden door de correlatie tussen metingen op verschillende plaatsen af te laten hangen van de afstand tussen die plaatsen. Hoe dichter meetplaatsen bij elkaar liggen, hoe hoger de correlatie. Het is dan niet verstandig om naburige meetplaatsen te kiezen. Dit maakt een systematische keuze wellicht aantrekkelijker dan een aselecte keuze van meetplaatsen. Barnes komt intuïtief kennelijk vanzelf bij de model-based aanpak terecht. De aantrekkelijke mogelijkheid om correlaties op te nemen tussen meetplaatsen voor een random patroon is in feite vergelijkbaar met de keuze voor random genetische-effecten in het animal model als middel om familieverbanden mee te nemen. De fysieke afwezigheid van een populatie van terreinen of patronen schept hier echter veel meer ruimte voor controverse dan in het fokkerij voorbeeld het geval is. Ook omdat de consequenties voor het nemen van een steekproef zo groot zijn.

5 Ridge-regressie, Bayesiaanse statistiek en penalty functies

Voor we afsluiten met een praktijkvoorbeeld knopen we hier nog wat losse einden een beetje aan elkaar vast.

Wat formeler dan we tot nu toe hebben gedaan schrijven we het model nu als volgt:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}, \qquad (3)$$

de fixed effecten (bijvoorbeeld behandelingsverschillen) zitten in vector β, de random effecten in vector **u**. Bijbehorende design matrices zijn **X** en **Z**. Verder zijn elementen van **u** en **e** onafhankelijk en normaal verdeeld met varianties σ^2_u en σ^2_e. Dit is een zogenaamd *mixed model*. *Mixed* omdat er een mix van fixed en random effecten in het model zit.

Nemen we de elementen van **u** toch fixed, dan kunnen we β en **u** uit de *normaal vergelijkingen* schatten:

$$\begin{bmatrix} \mathbf{X}'X & X'Z \\ Z'X & Z'Z \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}.$$

Hierbij worden bijvoorwaarden opgelegd om een unieke oplossing te verzekeren. Nemen we de elementen van u random dan kunnen we een schatting voor β en de BLUP voor **u** oplossen uit de zogenaamde *mixed model vergelijkingen*:

$$\begin{bmatrix} \mathbf{X}'X & X'Z \\ Z'X & Z'Z+\lambda I \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix},$$

hierin is $\lambda = \sigma^2_e / \sigma^2_u$. De overeenkomst tussen de stelsels vergelijkingen is duidelijk: de mixed model vergelijkingen volgen uit de normaalvergelijkingen door λ op te tellen bij de diagonaalelementen voor de randomeffecten links in het stelsel. Zie voor details bijvoorbeeld Engel (1990). Hier ligt een link met *ridge regressie* (zie bijvoorbeeld Draper en Smith, 1981, p313), waar in feite een soortgelijke truc wordt uitgehaald om aan een slecht geconditioneerd probleem toch nog informatie te ontfutselen. Voor het voorbeeld met de melkproduktiegegevens zien de mixed model vergelijkingen er als volgt uit:

$$\begin{bmatrix} 2 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 4 & 0 & 0 & 2 & 2 \\ 1 & 0 & 0 & 11 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 11 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 12 & 0 \\ 1 & 2 & 2 & 0 & 0 & 0 & 15 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 210 \\ 310 \\ 420 \\ 110 \\ 110 \\ 220 \\ 500 \end{bmatrix}.$$

Het verschil tussen random en fixed effecten verdwijnt wanneer we het mixed model (3) op z'n Bayesiaans bekijken. De normale verdeling voor de elementen van **u** wordt dan een *a priori* verdeling. Ook de fixed effecten hebben een a priori verdeling. Die verdeling is alleen vaak niet informatief en heeft dan de vorm van een oneindig brede homogene verdeling. Dit brengt overigens conceptueel wel wat problemen met zich mee (zie bijvoorbeeld Box en Tiao, 1992, p25), maar daar gaan we niet verder op in. Het opnemen van voorkennis past heel natuurlijk bij het gebruik van a priori verdelingen. De noodzaak tot het veronderstellen van een onderliggende populatie vervalt.

De mixed model vergelijkingen volgen, onder normaliteit, door de afgeleiden van de simultane *a posteriori* verdeling van $\boldsymbol{\beta}$ en **u** nul te stellen. Dit komt op hetzelfde neer als het nul stellen van de afgeleiden van de simultane verdeling van waarnemingen **y** en random effecten **u**:

$$\partial f(\mathbf{y},\mathbf{u}) / \partial \boldsymbol{\beta} = \mathbf{0} \quad \text{en} \quad \partial f(\mathbf{y},\mathbf{u}) / \partial \mathbf{u} = \mathbf{0}.$$

De simultane verdelingsdichtheid kunnen we ook schrijven als:

$$f(\mathbf{y},\mathbf{u}) = g(\mathbf{y} \mid \mathbf{u})\, h(\mathbf{u}),$$

waarin $g(. \mid .)$ de dichtheid van de kansverdeling van **y** gegeven **u** weergeeft (dus alsof **u** fixed was) en h(.) de normale dichtheid van **u** is. Het komt er dan op neer dat het maximum moet worden bepaald van:

$$\log(g(\mathbf{y} \mid \mathbf{u})) - \tfrac{1}{2} \Sigma\, u_k^2 / \sigma_u^2.$$

In feite hebben we dan aan de loglikelihood onder fixed u een kwadratische *penalty functie* in termen van de elementen u_k van **u** toegevoegd. Er staat als het ware een "boete" op het gebruik van waarden voor u_k die "ver" van 0 af liggen. Zie voor het gebruik van penalty functies bijvoorbeeld Eilers (1989). Schattingsmethoden voor mixed modellen voor niet-normale verdelingen, bijvoorbeeld voor 0 / 1 gegevens, zogenaamde *gegeneraliseerde lineaire mixed modellen*, geven voor $\boldsymbol{\beta}$ en **u** soortgelijke *penalized quasi-deviances* (zie bijvoorbeeld Engel, Buist en Visscher, 1995).

De schatting voor de variantie component σ_u^2 wordt ontleend aan de *marginale posteriori verdeling* van die component. Dit is de kansverdeling van σ_u^2 gegeven de waarnemingen (deze is *marginaal* in de zin dat alle overige parameters er "uit zijn geïntegreerd"). Met een oneindig brede homogene a priori verdeling voor de fixed effecten volgt de *restricted maximumlikelihood* (REML) schatter. REML is momenteel de populairste methode voor het schatten van variantiecomponenten (zie Engel, 1990). Met een puntmassa a priori verdeling voor de fixed effecten vinden we daarentegen de conventionele maximum likelihood (ML) schatter. Zo'n puntmassa verdeling houdt in feite in dat je de componenten schat alsof fixed effecten bekend zijn. REML schatters zijn (in frequentistische zin) zuiverder dan ML schatters, omdat beter rekening wordt gehouden met de aanwezigheid van de fixed effecten in het model die ook geschat moeten worden.

6 De kwaliteit van stieresperma

Nu een recent voorbeeld uit mijn eigen praktijk. De keuze voor fixed en random effecten zal niet zo verrassend zijn. Dit voorbeeld geeft echter wel een indruk wat voor gevolgen

het fixed of random nemen van een factor kan hebben voor het trekken van conclusies. Bovendien maakt het duidelijk dat een random model met zorg moet worden gekozen.

Het voorbeeld betreft onderzoek van het DLO-Instituut voor Dierhouderij en Diergezondheid (ID-DLO). Het gaat hier om het effect van verschillende manieren van behandelen van sperma van stieren op de kwaliteit van dit sperma.

Sperma van 5 stieren is verdund met 5 verschillende media (A, B, C, D en E) en vervolgens ingevroren volgens 4 invriesprotocollen (1, 2, 3 en 4). Dit invriezen gebeurt in porties (rietjes), die in praktijk voor kunstmatige inseminatie zouden worden gebruikt. Van iedere combinatie van stier, medium en protocol zijn 2 rietjes ontdooid. Ieder rietje is 2 maal beoordeeld. Deze beoordelingen zijn steeds door dezelfde persoon verricht. Een beoordeling houdt in dat 100 spermatozoën op hun kwaliteit worden beoordeeld onder de microscoop. Eén van de kwaliteitskriteria is het aantal niet gekleurde cellen (aan een monster wordt een kleurstof toegevoegd die in de kapot gevroren cellen dringt).

De waarnemingen zijn percentages niet gekleurde cellen. In de analyse werken we met fracties. Verwachte fracties zijn kansen p op kleuring.
Kansen liggen tussen 0 en 1. Onder- en bovengrenzen zijn lastig bij het modelleren, daarom "rekken" we de kansen op en modelleren we de logits van de kansen:

$$\text{logit}(p) = \log(p/(1-p)).$$

Hoofdeffecten en interacties voor factoren S (stier, 5 niveaus), M (media, 5 niveaus), P (protocol, 4 niveaus) en R (rietje, 2 niveaus binnen iedere combinatie van S, M en P) introduceren we nu op logitschaal:

$$\text{logit}(p) = S + M + P + R + S.M + S.P + M.P + S.M.P.$$

Hierin geeft S het hoofdeffect voor stieren weer, S.M de interactie tussen stieren en media, enzovoorts.

Het gaat ons om eventuele verschillen tussen de combinaties van media en invriesprotocollen. Uiteraard willen we geen uitspraak doen die beperkt blijft tot de 5 stieren in de proef. De dieren in de proef beschouwen we als een representatieve steekproef uit een onderliggende populatie. Effecten die stieren betreffen zijn daarom als random effecten opgenomen. Om dezelfde reden worden rietjes ook random genomen. Er is maar weinig informatie beschikbaar voor de variantiecomponent voor het stier-hoofdeffect. Dit hoofdeffect nemen we daarom fixed op. De analyse werkt dan alleen met binnen-stier contrasten. Interacties met stieren nemen we wel random op. Conclusies hebben dan nog steeds betrekking op de onderliggende stierpopulatie.

Zonder extra random effecten zouden we de data met een logistisch regressiemodel hebben geanalyseerd. Het logistische regressiemodel behoort tot de klasse van gegeneraliseerde lineaire modellen (GLMs) (Dobson, 1990). Omdat er naast fixed effecten ook random effecten op logitschaal voorkomen, hebben we een combinatie van een GLM en een mixed model nodig. Een geschikte klasse van modellen zijn de al genoemde gegeneraliseerde lineaire mixed modellen (GLMMs). Voor meer informatie over GLMMs zie bijvoorbeeld Engel en Keen (1994), Engel en Buist (1995) en Engel, Buist en Visscher (1995). Hier zullen we verder geen aandacht aan GLMMs besteden. Voor de berekeningen is gebruik gemaakt van een procedure (Keen, 1994), geschreven in Genstat 5 (1993).

Gegeven stieren en rietjes gaan we er vanuit dat de variantie kan worden weergegeven als een veelvoud van de binomiale variantie:

ϕ p(1-p) / 100.

Hierin is ϕ een parameter die over- of misschien onderdispersie vertegenwoordigt ten opzichte van binomiale variatie. Deze onbekende dispersieparameter wordt ook uit de data geschat en speelt binnen een GLMM de rol van restvariantie.

In Tabel 1 zijn de schattingen voor de variantiecomponenten weergegeven (tussen haakjes de bijbehorende standaardafwijkingen).

Tabel 1. *Schattingen voor de variantiecomponenten.*

Component	Schatting (s.a.)
stier x medium	0.005 (0.005)
stier x protocol	0.023 (0.012)
stier x medium x protocol	0.023 (0.007)
rietjes	negatief
restvariantie (= ϕ)	1.02 (0.08)

De negatieve schatting voor de variantiecomponent voor rietjes is in de analyse vervangen door een promille van de schatting van de restvariantie ϕ. De schatting $\hat{\phi}$ = 1.02 suggereert dat een binomiale verdeling de variatie op het "residuele" niveau afdoende kan weergeven. Ter illustratie zijn in de eerste kolom van Tabel 2 de schattingen voor de media gemiddelden op logitschaal weergegeven.

Tabel 2. *Gemiddelden op logitschaal en standaardafwijkingen voor verschillen tussen deze gemiddelden (s.e.d.'s).*

Medium	model met random stierinteracties en random rietjes	model met fixed stierinteracties en random rietjes	model met fixed stier interacties en geen effecten voor rietjes = gewone logistische regressie met overdispersie
A	0.49	0.48	0.49
B	1.09	1.08	1.10
C	1.00	1.00	1.02
D	0.90	0.90	0.92
E	0.94	0.94	0.95
s.e.d.			
media	0.08	0.04	0.04
media & prot.	0.13	0.07	0.07

Eigenlijk zijn we in de combinaties van media en protocollen geïnteresseerd, maar het nu volgende punt kunnen we ook met het geringere aantal media gemiddelden illustreren. In dezelfde tabel zijn ook resultaten opgenomen van een analyse met fixed stierinteracties en

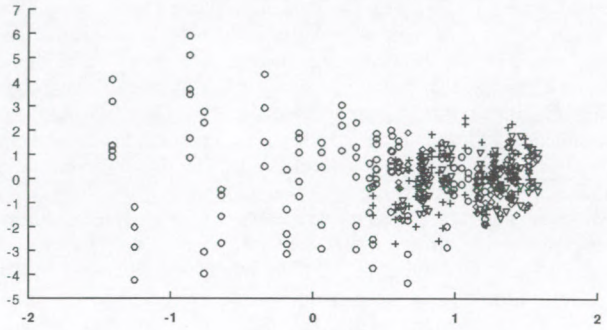
van een gewone logistische regressie waarin stiereffecten fixed zijn opgenomen en rietjes zijn genegeerd. Ook deze modellen bevatten een extra multiplicatieve parameter ϕ in de variantie. Schattingen voor deze parameter (uit Pearsons chi-kwadraat: $\Sigma(y-\hat{p})^2/\{\hat{p}(1-\hat{p})\}$) bedragen in beide gevallen 1.05. De gemiddelden ontlopen elkaar nauwelijks onder de drie modellen. Wat dat betreft zou je denken dat het allemaal de moeite niet waard is. Maar, niets is minder waar. De verschillen zitten niet zozeer in de gemiddelden (die in het algemeen vaak op elkaar lijken), maar in de standaardafwijkingen die we nodig hebben voor het paarsgewijs vergelijken van de gemiddelden. Deze standaardafwijkingen zijn niet voor alle paren gemiddelden gelijk. In Tabel 2 is de gemiddelde standaardafwijking weergegeven (s.e.d. = standard error of difference). Voor de volledigheid is in de laatste rij van de tabel ook de gemiddelde standaardafwijking voor het vergelijken van gemiddelden van combinaties van media en protocollen weergegeven. De standaardafwijkingen in het model met random stierinteracties zijn bijna twee keer zo groot als die in de modellen met fixed stierinteracties. Op het eerste gezicht lijkt dit misschien tegen de intuïtie in te gaan. Wanneer er meer informatie in een model wordt gestopt, hier in de vorm van een kansverdeling voor de stierinteracties, verwacht je te worden beloond met lagere standaardafwijkingen. Echter, wanneer effecten random in plaats van fixed worden genomen, brengt dit een essentiële wijziging in de aard van deze effecten met zich mee. De vraagstelling verandert ook. We zijn dan geïnteresseerd in uitspraken over een populatie van stieren en niet alleen over de 5 stieren in de proef. In het eerste geval zijn we veel ambitieuzer dan in het laatste geval. Het gevolg is dat uitspraken voorzichtiger en standaardafwijkingen van verschillen groter worden.

Protocol 4 wijkt nogal af van de overige protocollen in die zin dat de invriessnelheid veel hoger ligt (-350 °C/min. tegen -40, -100 en -150 °C/min. voor respectievelijk protocol 1, 2 en 3). Ter controle van de aanname van constante dispersieparameters over de combinaties van media en protocollen is de analyse een keer herhaald zonder protocol 4. De schattingen voor de variantiecomponenten staan in de linkerkolom van Tabel 3.

***Tabel 3**. Variantiecomponenten zonder en met protocol 4 in het model, met in het laatste geval extra random effecten voor protocol 4.*

Component	Model zonder protocol 4	Model met protocol 4 en extra random effecten voor protocol 4
Stier x medium	0.0032 (0.0031)	0.0042 (0.0039)
Stier x protocol	0.0023 (0.0027)	0.0223 (0.0112)
Stier x medium x protocol	0.0056 (0.0037)	0.0130 (0.0050)
Rietjes	negatief	negatief
Restvariantie(=ϕ)	0.65 (0.06)	0.64 (0.06)
Extra component protocol 4	-	0.08 (0.015)

Opvallend is de lage waarde voor $\hat{\phi}$. De schatting is nu veel lager dan 1! Een analyse van kwaliteitsbeoordelingen na verdunnen en vóór invriezen, leverde een vergelijkbare lage waarde op. Een residuenplot van residuen op logitschaal (Pearsonresiduen op de oorspronkelijke schaal) voor het complete model geeft aan dat voor protocol 4 relatief te veel extreme residuen voorkomen (zie Figuur 1).



***Figuur 1**. De residuen op logit-schaal uitgezet tegen de aangepaste waarden (schattingen fixed effecten + voorspellingen bijdragen random interacties en rietjes). Residuen voor protocol 4 zijn met rondjes (o) aangegeven.*

Op de logitschaal nemen we daarom voor protocol 4 een extra set restbijdragen op met een bijbehorende extra variantiecomponent. Eleganter zou zijn geweest om een afwijkende waarde voor ϕ voor protocol 4 toe te laten, maar dat was op korte termijn veel moeilijker numeriek te verwezenlijken. Daar de "binomiale totalen" hier constant zijn, komt het bijna op het zelfde neer. De rechterkolom in de tabel geeft de nieuwe resultaten weer voor het volledige model met de extra restbijdragen voor protocol 4. Deze resultaten blijken de lage dispersieparameter te combineren met schattingen voor de interactiecomponenten die redelijk overeenstemmen met de vorige tabel.

Overdispersie is geen ongewoon verschijnsel, maar onderdispersie wel. Onderdispersie zal ook zelden extreem zijn (zie Engel en te Brake, 1993). Inhomogeniteit van het monster onder de microscoop kan onderdispersie veroorzaken, maar nader onderzoek zal moeten uitwijzen of dit een afdoende verklaring is voor de toch wel erg lage uitkomst voor $\hat{\phi}$. Ter geruststelling voor diegenen die liever de som van de twee herhaalde waarnemingen, dus met een totaal van 200, hadden geanalyseerd: op papier is eenvoudig plausibel te maken dat dit vergelijkbare conclusies voor de combinaties van media en protocollen geeft (daadwerkelijke berekening bevestigde dit gelukkig).

Voor de uiteindelijke verslaggeving is het geen probleem om protocol 4 met z'n veel slechtere resultaten weg te laten. Voor een indruk van de optimale invriessnelheid bij ieder medium ligt dat anders. De invriessnelheden voor protocollen 1, 2 en 3 liggen vrij dicht bij elkaar en voor het aanpassen van een kwadratische curve hebben we protocol 4 echt nodig. De waarde van deze berekeningen is uiteraard beperkt, maar de orde van grootte van de schattingen kan toch informatief zijn. In de berekeningen behouden we de factoriële structuur voor de random effecten, maar vervangen we het hoofdeffect voor de

invriesprotocollen door een kwadratische curve in termen van de invriessnelheden. Resultaten in Tabel 4 geven een indruk van het effect van de extra random effecten voor protocol 4. De schattingen zijn weer ongeveer gelijk, maar de bijbehorende standaardafwijkingen zijn duidelijk kleiner als gevolg van het feit dat nu voor protocollen 1, 2 en 3 sprake is van onderdispersie (factor 0.64), terwijl voor protocol 4 een overdispersiefactor geldt (ongeveer een factor 2).

***Tabel 4**. Optimale invriessnelheden zonder en met de extra restbijdragen op logitschaal voor protocol 4 in het model.*

Medium	zonder extra effecten	met extra effecten
A	117.4 (14.4) °C/min.	117.5 (10.2) °C/min.
B	143.2 (56.9)	143.2 (41.2)
C	141.2 (72.8)	140.6 (52.1)
D	105.3 (33.1)	105.4 (23.3)
E	26.2 (125.5)	27.4 (88.9)

7 Slot

Ook wanneer er formeel geen sprake is van een onderliggende populatie, kan het aantrekkelijk zijn om effecten random en niet fixed op te nemen. Bijvoorbeeld wanneer het gaat om selectie. Zeker in Bayesiaanse zin ligt daar geen probleem. Bij random effecten is het eenvoudig om samenhang tussen de waarnemingen (familieverbanden, naburige geografische ligging) in de vorm van covarianties mee te nemen.

In experimenten met een groot aantal factoren op veel niveaus, waar conclusies doorgaans gebaseerd worden op tabellen die vaak maar twee en hoogstens drie factoren per tabel betreffen, kan het aantrekkelijk zijn hogere orde interacties random op te nemen. Zoals de link met penalty functies aangeeft: het random opnemen van een effect houdt in zekere zin in dat schattingen "gladgestreken" worden. Dit kan voor hogere orde interacties, die doorgaans als ondergeschikt aan hoofdeffecten en tweede-orde-interacties worden beschouwd, een aantrekkelijke strategie zijn. Zeker wanneer het interacties met *nuisance* factoren betreft. Variantiecomponenten geven bovendien een snelle en compacte indruk van het relatief belang van verschillende interacties in het model.

Verwacht mag worden dat mixed modellen, zowel voor continue als voor discrete data, in toenemende mate aftrek zullen vinden. Niet alleen als gevolg van de toenemende beschikbaarheid van faciliteiten in statistische pakketten, maar ook omdat de eigenschappen van predicties van random effecten soms aantrekkelijker zijn dan de eigenschappen van schattingen voor fixed effecten.

Verwijzingen

Barnes, R.J. (1988). Bounding the required sample size for geologic site characterization. *Mathematical Geology* **20**, 477-490.

Box, G.E.P., Tiao, G.C. (1992). *Bayesian inference in statistical analysis.* Wiley Classics Library Edition. John Wiley, New York.

Cochran, W.G. (1951). Improvement by means of selection. *Proceedings of the 2nd Berkeley Symposium.* (J. Neyman, editor), 499-470. University of California Press, Los Angeles.

Dobson, A.J. (1990). *An introduction to generalized linear models.* Chapman and Hall, Londen.

Draper, N., Smith, H. (1981). *Applied regression analysis.* 2e editie. John Wiley, New York.

Eilers, P.H.C. (1989). Smoothing en interpolatie met gegeneraliseerde lineaire modellen. *Kwantitatieve Methoden* **30**, 33-46.

Engel, B. (1990). The analysis of unbalanced linear models with variance components. *Statistica Neerlandica* **44**, 195-219.

Engel B, Te Brake J (1993) Analysis of embryonic development with a model for under- or over-dispersion relative to binomial variation. *Biometrics* **49**, 269-279.

Engel, B., Buist (1995). Analysis of a generalized linear mixed model: a case study and simulation results. *Biometrical Journal* **37**, 00-00 (in druk).

Engel, B., Buist, W., Visscher, A. (1995). Inference for threshold models with variance components from the generalized linear mixed model perspective. *Genetics Selection Evolution* **27**, 15-32.

Engel, B., Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statististica Neerlandica* **48**, 1-22.

Genstat 5 committee (1993) *Genstat 5 Release 3 Reference Manual.* R.W. Payne (Voorzitter) en P.W. Lane (Secretaris). Clarendon Press, Oxford.

Gruijter, de J.J., ter Braak, C.J.F. (1990). Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology* **22**, 407-415.

Gruijter, de J.J. (1991). Klassieke steekproeftheorie in de ruimtelijke statistiek. *Kwantitatieve Methoden* **37**, 51-63.

Keen, A. (1994). *Procedure IRREML.* In: *Genstat 5 GLW-DLO Procedure Library Manual Release 3[1].* (P.W. Goedhart en J.T.N.M. Thissen, editors), Rapport LWA-94-17, DLO-Groep Landbouwwiskunde (GLW-DLO), Wageningen.

Portnoy, S. (1982). Maximizing the probability of correctly ordering random variables using linear predictors. *Journal of Multivariate Analysis* **12**, 256-269.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15-51.

Searle, S.R. (1971). *Linear models.* John Wiley, New York.

Searle, S.R. (1974). Prediction, mixed models and variance components. In: *Reliability and Biometry* (F. Proschan en R.J. Serfling, editors), 229-266. Society for Industrial and Applied Mathematics, Philadelphia.

Searle, S.R., Casella, G., McCulloch, C.E. (1992). *Variance components.* John Wiley, New York.

Ontvangen: 15-3-1995 Geaccepteerd: 16-6-1995