

STATISTISCHE ASPECTEN VAN CALIBRATIE^{**)}

A.A.M. Jansen^{*)}

Samenvatting

Deze tekst bevat de inhoud van een voordracht voor de Landbouwkundige Sectie van de Vereniging voor Statistiek/Studiekring voor Statistische Techniek van het Koninklijk Genootschap voor Landbouwwetenschap op 2 mei 1986. Er wordt vooral aandacht besteed aan de betekenis van de parameters en de relaties daartussen bij de klassieke en de inverse schatters en aan proeftechnische aspecten van calibratieproblemen. De indeling van de tekst is als volgt:

- 1 Inleiding.
- 2 Het model.
- 3 Schatten van de parameters van het model.
- 4 Gebruiken van het model voor het meten.
- 5 Proeftechnische aspecten.
 - 5.1 Schattingsonnauwkeurigheid van de parameters.
 - 5.2 Keuze van de instelwaarden van x of van y.
 - 5.3 Aantal y-waarden groter dan het aantal punten in de oefenset.
 - 5.4 Fouten in de x-waarden.
 - 5.5 Herhaalde meting van y.

APPENDIX.

REFERENTIES.

*) Instituut voor Toegepaste Informatica, afdeling Statistiek-Landbouw,
Postbus 100, 6700 AC Wageningen, tel.: 08370 -19100

**)Tevens verschenen als TNO-rapport 86 ITI A 24, september 1986

1 Inleiding

Aan een calibratieprobleem kunnen de volgende aspecten worden onderscheiden:

- het model voor de relaties tussen de beschouwde grootheden,
- het schatten van de parameters van het model,
- het gebruiken van het model voor het meten,
- proeftechnische aspecten.

Veel historische controverse onder statistici over het calibratieprobleem is terug te voeren tot onvolledige formuleringen van model en probleemstelling. Met name het schatten van parameters en het gebruiken van het model voor het vaststellen van nieuwe metingen zijn aspecten die beter onafhankelijk van elkaar kunnen worden behandeld. Dit zal in het volgende worden gedaan.

2 Het model

Bij calibratieproblemen is er altijd sprake van functionele relaties tussen twee stellen grootheden die men wil gebruiken voor het vaststellen van het ene stel (x) op basis van metingen aan het andere stel (y). Deze functionele relaties tussen y en x hoeven niet lineair te zijn, maar in eerste instantie zal ik mij beperken tot lineaire relaties. Behalve de functionele relaties dienen ook de stochastische componenten van de meet-situatie te worden gemodelleerd. Ook hier zal ik mij in eerste instantie beperken tot de meest eenvoudige veronderstellingen: alleen de gemeten grootheden y bevatten meetfouten, met een covariantiestructuur die onafhankelijk is van de functionele relatie. Zonodig worden de meetfouten normaal verdeeld verondersteld. Ik zal dus, kortom, het klassieke regressiemodel als uitgangspunt beschouwen:

$$y = \alpha + \beta x + e \quad Ee^2 = \sigma^2 \quad (1)$$

$$\text{of } y = \alpha + B'x + e \quad Eee' = \Sigma \quad (2)$$

Het eerste model heeft betrekking op enkelvoudige variabelen y en x .

In het tweede model zijn y , α en e vectoren met lengte q ; x is een vector met lengte p ; B' is een matrix van $q \times p$ en Σ is een matrix van $q \times q$.

Voor calibratie van het model c.q. het schatten van de parameters van het model is het niet van belang op welke wijze de waarden van de zogenaamde instelvariabelen x gekozen zijn; het is niet nodig onderscheid te maken tussen situaties waarin de x -variabelen ingesteld zijn en situaties waarbij de x -variabelen op een andere wijze, mits zonder meetfout, zijn verkregen. Voor latere beschouwingen zal ik nodig hebben dat in de praktijk de te meten waarden van x uit een relevante populatie afkomstig kunnen zijn en dan een verdeling hebben. Deze verdeling zal ik steeds kiezen als een normale verdeling met verwachting μ_x en met variantie σ_x^2 , dan wel met covariantiematrix D .

3 Schatten van de parameters van het model

Wanneer gegevens beschikbaar zijn van een voldoende omvangrijke 'trainingsset' waarvoor zowel y als x is vastgesteld kan men met lineaire regressietheorie schattingen voor alle parameters (α ; β , B of σ^2 of Σ) berekenen en de precisie daarvan vaststellen. Ik ga hier niet op in. In het volgende zal worden aangenomen dat alle parameters met een zodanige nauwkeurigheid zijn vastgesteld dat zij als bekende grootheden kunnen worden beschouwd. Later kom ik hierop terug.

4 Gebruiken van het model voor het meten

Een voor de hand liggende methode om x vast te stellen als y gemeten is volgt uit de theorie voor lineaire regressie. Neem model (2), nu met x als te schatten parametervector en B' als matrix met bekende, dat wil zeggen in de trainingsfase vastgestelde, elementen:

$$y = \alpha + B'x + e \quad (2)$$

De schatter voor x is:

$$\bar{x}_{GLS} = (B\Sigma^{-1}B')^{-1} B\Sigma^{-1}(y - \alpha) \quad (3)$$

De letters GLS staan voor "generalised least squares". Voor model (1) reduceert dit tot

$$\bar{x}_{GLS} = \frac{y - \alpha}{\beta} \quad (4)$$

De schatter $\tilde{x}_{GLS}$ is de beste lineaire zuivere schatter, en deze staat bekend als de klassieke schatter. Daarvoor geldt:

$$E\tilde{x}_{GLS}|x = x \quad (5)$$

$$\text{Var}(\tilde{x}_{GLS}|x) = (B\Sigma^{-1}B')^{-1}. \quad (6)$$

Voor model (1) reduceert de laatste uitdrukking tot

$$\text{Var}(\tilde{x}_{GLS}|x) = \sigma^2/\beta^2.$$

Zolang naast de meting y alleen het model wordt gehanteerd als bron van informatie over x geeft $\tilde{x}_{GLS}$ de beste voorspeller voor x die voorhanden is bij de gemeten waarde van y .

Als men zou weten dat de meting y behoort bij een onbekende x -waarde, die echter aselekt getrokken is uit een bekende verdeling van x -waarden, kan men een andere lineaire voorspeller bedenken die "beter" is dan de bovenstaande. Voor model (1) is dat deze:

$$\tilde{x}_{BLP} = (1 - \rho^2) \mu_x + \rho^2 \tilde{x}_{GLS} \quad (7)$$

waarin $\rho^2 = \beta^2\sigma_x^2/(\beta^2\sigma_x^2 + \sigma^2)$. Men kan $\tilde{x}_{BLP}$ opvatten als een gewogen gemiddelde van $\tilde{x}_{GLS}$ en μ_x met als gewichten de inversen van σ^2/β^2 en σ_x^2 . In deze opvatting wordt μ_x beschouwd als een "schatter" voor x met mean square error (MSE) gelijk aan σ_x^2 . De overeenkomstige uitdrukking voor model (2) wordt (zie appendix):

$$\begin{aligned} \tilde{x}_{BLP} &= (D^{-1} + B\Sigma^{-1}B')^{-1}[D^{-1}\mu_x + B\Sigma^{-1}B'\tilde{x}_{GLS}] \\ &= [I - DB(B'DB + \Sigma)^{-1}B'] \mu_x + DB(B'DB + \Sigma)^{-1}B' \tilde{x}_{GLS} \\ &= (I - A) \mu_x + A\tilde{x}_{GLS} \end{aligned} \quad (8)$$

$\tilde{x}_{BLP}$ staat bekend onder de naam "inverse schatter" vanwege de verwant-

schap tussen $\tilde{x}_{BLP}$ en de regressie van x op y . In feite kan men $\tilde{x}_{BLP}$ algemener via de regressie van x op y definiëren. Immers voor $p > q$ bestaat wel de regressie, maar niet $\tilde{x}_{GLS}$. Door in (8) de term $B\Sigma^{-1}B'\tilde{x}_{GLS}$ te vervangen door $B\Sigma^{-1}(y-\alpha)$ krijgt men direct de gedaante van de regressie van x op y .

Als men de covariantiematrix van y en x in de populatie opstelt op basis van de modelparameters vindt men voor model (1) en model (2) respectievelijk:

	x	y		x	y
x	σ_x^2	$\beta\sigma_x^2$	x	D	DB
y	$\beta\sigma_x^2$	$\beta^2\sigma_x^2 + \sigma^2$	y	B'D	B'DB + Σ

Hieruit kan men gemakkelijk (7) en (8) afleiden (zie appendix). Voorts dat geldt bij model (1):

$$E(\tilde{x}_{BLP}|x) = \rho^2 x + (1 - \rho^2) \mu_x = x - (1 - \rho^2)(x - \mu_x) \quad (9)$$

$$EMSE(\tilde{x}_{BLP}|x) = (1 - \rho^2)\sigma_x^2 = \rho^2\sigma^2/\beta^2 \leq \text{Var}(\tilde{x}_{GLS}|x) = \sigma^2/\beta^2. \quad (10)$$

Met andere woorden: $\tilde{x}_{BLP}$ is geen zuivere schatter voor de werkelijke waarde x , maar ligt gemiddeld over alle waarden van x in de populatie dichter bij x dan $\tilde{x}_{GLS}$. Ditzelfde geldt voor de multivariate situatie, waarvoor overeenkomstige uitdrukkingen kunnen worden afgeleid (zie appendix). Daarom is het aantrekkelijk $\tilde{x}_{BLP}$ te gebruiken in al die gevallen waarin x beschouwd kan worden als een aselekt gekozen representant van een goed gedefinieerde populatie. Merk op, dat de mean square error van $\tilde{x}_{BLP}$ gelijk is aan de variantie rond de regressie van x op y , c.q. de variantie-covariantiematrix na multivariate regressie van x op y .

De keuze tussen $\tilde{x}_{GLS}$ en $\tilde{x}_{BLP}$ hangt overigens niet alleen af van de mogelijkheid een zinvolle populatie van x -waarden te definiëren en de parameters daarvan te bepalen. Het is duidelijk dat $|\tilde{x}_{BLP} - x|$ groter wordt naarmate x verder van μ_x ligt. Boven een zekere grens zal men beter $\tilde{x}_{GLS}$ kunnen gebruiken. Dit is zeker het geval als men in bepaalde excentrische trajecten van x -waarden een grotere precisie wenst dan elders. Een

combinatie van de beide schatters ligt dan ook voor de hand (Naes, 1985a). Tenslotte dient nog te worden opgemerkt dat het verschil tussen $\bar{x}_{BLP}$ en $\bar{x}_{GLS}$ klein wordt als ρ^2 groot is, dat wil zeggen als σ^2 klein is ten opzichte van $\beta^2 \sigma_x^2$. Overeenkomstige uitspraken kunnen worden geformuleerd voor de multivariate situatie. Aan de hand van enkele eenvoudige voorbeelden zal ik aan enkele details aandacht schenken.

Voorbeeld 1 (één y-variabele en één x-variabele)

Voor het vaststellen van het vleespercentage (x) van varkenskarkassen meet men de spekdikte (y). Een trainingssset kan men krijgen door van een aselechte steekproef uit de populatie zowel y als x te meten. De regressie van x op y is in dit geval de relevante voorspeller. Als men dit als uitgangspunt kiest kan men de precisie van de voorspellers nog opvoeren door de karkassen waarvoor men de moeilijke meting x uitvoert te selecteren aan de hand van y. Het is bekend dat dit geen invloed heeft op de regressie van x op y, tenminste als men bij de geselecteerde y-waarden aselekt karkassen kiest voor het meten van x.

Voorbeeld 2 (uit Oman & Wax, 1984; twee y-variabelen en één x-variabele)

Voor het vaststellen van de leeftijd van een foetus gebruikt men als voorspellende variabelen schedelomvang en femurlengte. Omdat voor de zich aandienende gevallen niet op zinvolle wijze een verdeling van leeftijden kan worden gedefinieerd is de klassieke schatter $\bar{x}_{GLS}$ de voor de hand liggende keuze voor de voorspeller. In dit voorbeeld te meer omdat men bij de trainingssset geselecteerd heeft op leeftijd c.q. op gevallen waarin de leeftijd met zekerheid kon worden vastgesteld. (Overigens worden in het geciteerde artikel curvilineaire relaties verondersteld en worden om pragmatische redenen inverse regressies berekend.)

Voor het bepalen van $\bar{x}_{GLS}$ zal men eerst regressies van elke y op x vaststellen, alsmede de covariantiematrix van de residuen Σ en vervolgens rechtstreeks $\bar{x}_{GLS}$ berekenen.

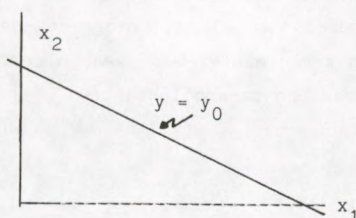
Het is hierbij niet noodzakelijk dat elke y-variabele in belangrijke mate samenhang met x laat zien. Multivariate calibratie is ook van nut als

zeg y_2 alleen gecorreleerd is met de residuen van de regressie van y_1 op x . Als deze correlatie gelijk is aan r zal de variantie van $\tilde{x}_{GLS}$ door meenemen van y_2 reduceren van σ_1^2/β^2 tot $(1-r^2)\sigma_1^2/\beta^2$. Voor het corrigeren van de effecten van verontreinigingen en dergelijke kan dit interessante perspectieven bieden bij het verbeteren van de resultaten van calibratie van enkelvoudige metingen via covariabelen. Het voorbeeld maakt tevens duidelijk dat de keuze van covariabelen niet alleen dient af te hangen van hun lineaire samenhang met x .

De schatter $\tilde{x}_{BLP}$ kan in voorkomend geval gemakkelijk worden vastgesteld door berekening van de regressie van x op y_1 en y_2 . Onafhankelijk van de vraag of men $\tilde{x}_{GLS}$ dan wel $\tilde{x}_{BLP}$ zal gebruiken kan men het selecteren van relevante y -variabelen wellicht altijd het handigst realiseren via regressies van x op y . Men zal er overigens voor dienen te waken dat het aantal y -variabelen groter is dan of gelijk is aan het aantal x -variabelen dat men wenst te kunnen identificeren, tenzij men gebruik kan maken van voorinformatie over de verdeling van x en dan de inverse regressie kan toepassen.

Voorbeeld 3 (één y -variabele en twee x -variabelen)

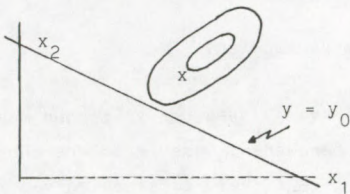
Hoewel in deze situatie de regressie van y op x_1 en x_2 zonder meer kan worden vastgesteld kan men $\tilde{x}_{GLS}$ niet berekenen. Deze is immers onbepaald, omdat aan de vergelijking $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ vele combinaties van x_1 en x_2 voldoen:



Een uitweg kan zijn dat men afzonderlijke regressies van y op x_1 en van y op x_2 berekent, waarmee het wel mogelijk is $\tilde{x}_{GLS}$ te berekenen voor zowel x_1 als x_2 . Deze uitweg geeft echter in het algemeen misleidende resultaten, omdat zowel het niveau als de helling van beide regressies afhangen van de waarden van de andere x -variabelen in de oefenset. Bovendien krijgt men op deze wijze, ook als x_1 en x_2 in de trainingssset orthogonaal

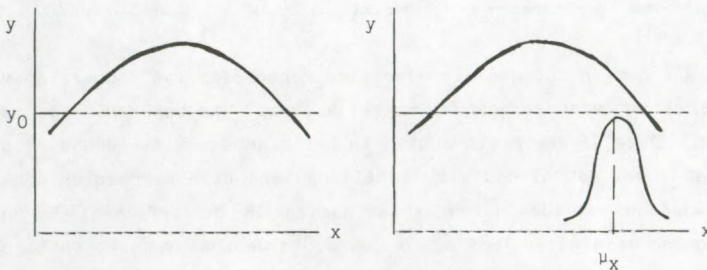
zouden zijn, een overschatting van de grootte van de meetfouten.

De schatter $\bar{x}_{BLP}$ is wel te berekenen, immers hiervoor zijn alleen de regressies van x_1 op y en van x_2 op y nodig en deze leveren $\bar{x}_{BLP}$ als $(\bar{x}_{1BLP}, \bar{x}_{2BLP})'$. Dat dit juist is kan men zien door in de uitdrukking voor $\bar{x}_{BLP}$ de term $(B\Sigma^{-1}B')\bar{x}_{GLS}$ te vervangen door $B\Sigma^{-1}(y-\alpha)$, zoals in de appendix is gedaan. Een enigszinsanschouwelijke voorstelling verkrijgt men door in bovenstaande plaatje de tweedimensionale dichtheid van (x_1, x_2) te symboliseren via een stel concentrische ellipsen. De (x_1, x_2) combinaties op elke ellips zijn even aannemelijk op basis van de bekende informatie over x ; de aannemelijkheid neemt af naarmate de combinaties verder van μ_x liggen. Anderzijds zijn met betrekking tot de informatie op basis van y alle combinaties (x_1, x_2) op de niveau-krommen van $\beta_1 x_1 + \beta_2 x_2$ even aannemelijk; de aannemelijkheid neemt af naarmate de combinaties verder afdrijven van $\beta_1 x_1 + \beta_2 x_2 = y - \alpha$. De beide soorten informatie samen laten blijkbaar een keuze van een beste (x_1, x_2) toe, die wordt gegeven door de regressie van x_1 en x_2 op y .



In dit voorbeeld helpt de informatie over de verdeling van x bij het kiezen van de oplossing. Een bijzonder voorbeeld hiervan vindt men in de situatie met één y -variabele en één x -variabele, maar waarin de functionele relatie tussen y en x de vorm van een parabool heeft:

$$y = \alpha + \beta x + \gamma x^2.$$



Gegeven een nieuwe waarde van y zijn er 2 kandidaten voor $\tilde{x}_{\text{GLS}}$ (tenzij y groter is dan of gelijk is aan de top van de parabool). Als men echter weet in welke regio men x moet zoeken kan dit helpen om een keuze te maken. Overigens is de afleiding van $\tilde{x}_{\text{BLP}}$ in dit geval niet zo voor de hand liggend als boven.

Het laatste voorbeeld suggereert dat extra informatie over x het mogelijk maakt ook niet-monotone relaties te gebruiken voor calibratiedoeleinden.

Omdat het daarbij in wezen niet veel ter zake doet of die extra informatie rechtstreeks gegeven is dan wel in de vorm van informatie die opgeslagen is in andere y -variabelen biedt dit wellicht perspectief voor verder onderzoek.

5 Proeftechnische aspecten

5.1 Schattingsonnauwkeurigheid van de parameters

Bij de voorgaande beschouwingen is er van uitgegaan dat de parameters van het model bekend zijn of als zodanig kunnen worden beschouwd. Onder die conditie zijn de eigenschappen van de meetmethoden beschreven. In de praktijk moeten de parameters eerst worden geschat. De verkregen schattingen kunnen worden ingevuld op de plaatsen van de parameters in de gegeven uitdrukkingen. Dit levert bijvoorbeeld $\tilde{\tilde{x}}_{\text{GLS}}$ en $\tilde{\tilde{x}}_{\text{BLP}}$. Als de schattingen consistent zijn, dat wil zeggen als deze in kans convergeren naar de werkelijke waarden van de parameters bij toenemende steekproefgrootte, zullen ook de verkregen uitdrukkingen in kans convergeren naar de uitdrukkingen met de werkelijke parameterwaarden ingevuld (Naes, 1985a):

$$\tilde{\tilde{x}}_{\text{GLS}} \xrightarrow{P} \tilde{x}_{\text{GLS}}$$

$$\tilde{\tilde{x}}_{\text{BLP}} \xrightarrow{P} \tilde{x}_{\text{BLP}}$$

Uiteraard heeft een beperkte steekproefomvang tot gevolg dat $\tilde{\tilde{x}}_{\text{GLS}}$ en $\tilde{\tilde{x}}_{\text{BLP}}$ veel kunnen afwijken van $\tilde{x}_{\text{GLS}}$ en $\tilde{x}_{\text{BLP}}$. Deze afwijkingen fungeren bij herhaald gebruik van de eenmaal vastgestelde voorspeller als een

systematische fout. In tegenstelling met de in de voorgaande paragraaf gegeven voorspellingsfouten kunnen deze schattingsfouten naar believen worden verkleind door een grotere steekproefomvang en door een optimale allocatie van de instelwaarden van x (of van y). Als de voorspellers bedoeld zijn voor herhaald gebruik dient er naar gestreefd te worden deze schattingson nauwkeurigheid zo klein te maken dat de invloed daarvan op de precisie van de voorspeller verwaarloosbaar klein is in het praktisch interessante traject van meetwaarden.

Men kan ruwweg stellen dat de bijdrage van de schattingsfouten aan de MSE van de voorspeller een fractie van de orde van grootte $(1+p)/n$ is. Een indicatie van het vereist aantal punten levert derhalve $(1+p)/n < \gamma$, waarin γ een te kiezen constante is waarmee de orde van grootte van de gemiddelde systematische fout in de voorspellers kan worden beperkt.

Onderstaande tabel geeft een numerieke indruk:

p	$\gamma = .10$	$\gamma = .20$
1	20	10
2	30	15
5	60	30
10	110	55

De onnauwkeurigheid in de schatter voor Σ heeft geen systematische afwijking in $\bar{x}_{GLS}$ ten gevolge. Immers

$$E\bar{x}_{GLS} | x = x$$

ongeacht de schatting voor Σ , die wordt ingevuld. Het effect op de precisie van $\bar{x}_{GLS}$ is waarschijnlijk gering. Wel werkt schattingson nauwkeurigheid van Σ door in de systematische afwijking van $\bar{x}_{BLP}$. Deze kan echter op dezelfde wijze als boven beperkt worden gehouden omdat $\bar{x}_{BLP}$ van de inverse regressie is afgeleid. De eis zou nu kunnen zijn: $n > (1+q)/\gamma$.

Het is duidelijk dat om de schattingson nauwkeurigheid van de parameters te kunnen verwaarlozen zelfs bij optimale allocatie grote aantallen waarnemingen in de oefenset nodig zullen zijn en dat die aantallen toenemen

met het aantal variabelen in het probleem. Daarom zal het in de praktijk veelal nodig zijn met deze onnauwkeurigheid rekening te houden bij het opgeven van de precisie van de uitkomsten. Zie daarvoor de literatuur (Brown, 1982). Dit neemt niet weg dat het altijd zinvol is de grootte van beide bijdragen ook afzonderlijk te beschouwen, en te trachten de schattingsonnauwkeurigheid door een voldoende aantal goed gekozen instelpunten zo klein mogelijk te maken.

5.2 Keuze van de instelwaarden van x (of van y)

Uit de theorie van optimale proefopzetten is bekend dat men bij lineaire regressie het best punten kan kiezen op een minimaal aantal lokaties regelmatig gespreid op de rand van het gebied waarin het model geldig wordt geacht. In de praktijk zal men dit advies meestal niet strict volgen omdat niet echt bekend is waar het model geldt dan wel omdat men wil kunnen nagaan of het model past of alternatieve modellen wil kunnen schatten (vgl. Buonaccorsi, 1986). Wel zal het vaak lonend zijn bij de keuze van instelpunten af te wijken van de eventuele frequentieverdeling in de doelpopulatie (als die bestaat). Als men $\tilde{x}_{\text{GLS}}$ zal gaan gebruiken zal men selectie plegen op de instelwaarden van x. Men kan met de oefenset in dat geval niet meer rechtstreeks $\tilde{x}_{\text{BLP}}$ berekenen. Daarvoor zou men op andere wijze verkregen informatie over μ_x en D moeten combineren met $\tilde{x}_{\text{GLS}}$, zoals boven aangegeven. Als men rechtstreeks $\tilde{x}_{\text{BLP}}$ wil bepalen zal selectie op instelwaarden van y, indien dit mogelijk is, de voorkeur hebben boven aselechte keuze van punten. Een goed advies zou kunnen zijn uit de doelpopulatie aselekt een steekproef te trekken, vervolgens voor alle elementen in de steekproef de "eenvoudige" metingen te verrichten (waarmee informatie over de doelpopulatie wordt verkregen) en tenslotte de "dure" metingen uit te voeren aan strategisch gekozen objecten.

5.3 Aantal y-waarden groter dan het aantal punten in de oefenset

Als men te doen heeft met problemen waarin veel y-variabelen worden gemeten, bijvoorbeeld bij gegevens afkomstig van apparatuur die een pieken-spectrum als output levert, zal het gemakkelijk voorkomen dat het aantal y-variabelen groter is dan het aantal gemeten punten. In dat geval kan, zolang het aantal punten groter is dan het aantal x-variabelen nog wel de

multivariate regressie van y op x worden berekend maar levert de schatting voor Σ een singuliere matrix op. Dan kan $\tilde{x}_{\text{GLS}}$ niet meer berekend worden met de gegeneraliseerde kleinste kwadratenmethode. Overeenkomstige problemen krijgt men als men rechtstreeks $\tilde{x}_{\text{BLP}}$ zou willen berekenen omdat dan de singulariteit van de matrix van kwadraten en producten van y dit verhindert. In zulke gevallen zal men een uitweg moeten zoeken in een beperking van het aantal parameters dat moet worden geschat. Daarbij kan worden gedacht aan beperking van het aantal onafhankelijke y -variabelen:

- door selectie van variabelen. Eerder is aangegeven dat dit zou kunnen via regressies van x op y en toepassing van daarvoor geschikte methoden,
- door clustering van variabelen en het samenstellen van een nieuwe variabele voor elk relevant geacht cluster (bijvoorbeeld door optellen of andere ad hoc procedures),
- door het uitvoeren van principale componentenanalyse op de y -variabelen en de grootste principale componenten als nieuwe y -variabelen in te voeren,
- door het uitvoeren van principale componentenanalyse op de residuen van de y -variabelen, na regressie op x , om daarmee de parameters van een gestructureerde covariantiematrix Σ te schatten, die vervolgens kan worden gebruikt voor berekening van $\tilde{x}_{\text{GLS}}$ op basis van alle y -variabelen (Naes, 1985b). Een gestructureerde Σ zal in de praktijk vaak zeer redelijk zijn omdat het aantal variatiebronnen (niet gemeten storingscomponenten) kleiner kan blijken te zijn dan het aantal y -variabelen.

Van de hier genoemde mogelijkheden lijken de tweede en de vierde methode, of eventueel een combinatie daarvan, de aantrekkelijkste. Voor voorbeelden zie Naes (1985b), die toepassingen geeft bij NIR-spectrophotometergegevens voor de bepaling van het watergehalte van vleesmonsters en voor de bepaling van het vetgehalte van vismonsters. De derde methode is niet bevredigend omdat de selectie van de nieuwe y -variabelen niet mede gebaseerd is op de relatie met x zoals in canonische correlatie-analyse het geval zou zijn. Deze laatste methode kan echter niet worden toegepast vanwege dezelfde singulariteiten als boven beschreven.

5.4 Fouten in de x-variabelen

Als men x-variabelen instelt is het mogelijk dat zogenaamde applicatiefouten optreden, die niet worden geregistreerd (het zogenaamde Berkson model). Deze applicatiefouten vergroten de varianties van de y-variabelen, maar behoren niet tot de meetfouten in y. Men zou in feite Σ hiervoor dienen te corrigeren alvorens $\bar{x}_{\text{GLS}}$ of $\bar{x}_{\text{BLP}}$ te berekenen, zeker bij $\bar{x}_{\text{BLP}}$ omdat deze hiervoor het gevoeligst is. In de praktijk zal dit vaak om verwaarloosbare bijdragen gaan. In het voorbeeld van de fetale leeftijd van Oman & Wax (1984) kan dit wel degelijk een factor van belang zijn omdat fetale leeftijd niet exact kan worden vastgesteld, maar wordt gerelateerd aan het tijdstip van de laatste menstruatie.

De problemen worden groter wanneer de x-variabelen door meting worden vastgesteld en daarbij een niet verwaarloosbare meetfout wordt gemaakt. Dit kan zich in de praktijk voordoen bij "comparative calibration" situaties, dat wil zeggen bij het onderling calibreren van niet-foutvrije methoden. Het definiëren van de functionele relaties en het schatten daarvan wordt dan aanzienlijk gecompliceerder. Indien $\bar{x}_{\text{BLP}}$ relevant is voor het probleem heeft men hiervan geen last als deze rechtstreeks wordt vastgesteld. Ik zal hierop niet verder ingaan; zie enkele TNO-rapporten over deze materie (Engel, 1984; Jansen, 1983).

5.5 Herhaalde meting van y

Wanneer men bij het gebruik van het calibratie model in plaats van enkelvoudige y-waarnemingen bij iedere x het gemiddelde van m onafhankelijke waarnemingen gaat gebruiken ter verbetering van de precisie dan kan men $\bar{x}_{\text{GLS}}$ gebruiken zoals deze op de oefenset is vastgesteld. De voorspeller $\bar{x}_{\text{BLP}}$ moet echter aan de nieuwe situatie worden aangepast omdat in $\Sigma + B'DB$ in uitdrukking (8) Σ vervangen moet worden door Σ/m . Dit kan alleen wanneer $\bar{x}_{\text{BLP}}$ indirect wordt berekend. Als men deze werkwijze toepast dient wel zeer zorgvuldig te worden nagegaan of de herhalingswaarnemingen werkelijk onderling onafhankelijk zijn met betrekking tot de variatiebronnen waarop Σ betrekking heeft. In de praktijk is dit meestal niet het geval bij herhalingsmetingen aan hetzelfde object omdat Σ ook voor een deel door object-specifieke modelafwijkingen wordt veroorzaakt.

Model

$$y = \alpha + B'x + e$$

$$E y = \alpha + B'x$$

$$\text{Var } y = \Sigma$$

Indien x een verdeling heeft is de covariantiematrix van x en y gelijk aan:

	x	y
x	D	DB
y	$B'D$	$B'DB + \Sigma$

Klassieke schatter

Uit de normaal-vergelijkingen $B\Sigma^{-1}B'\tilde{x}_{GLS} = B\Sigma^{-1}(y-\alpha)$ volgt:

$$\tilde{x}_{GLS} = (B\Sigma^{-1}B')^{-1} B\Sigma^{-1}(y-\alpha)$$

Invullen van $E y$ en $\text{var } y$ overeenkomstig bovenstaand model levert op:

$$E\tilde{x}_{GLS} \Big| x = x$$

$$\text{var } \tilde{x}_{GLS} \Big| x = (B\Sigma^{-1}B')^{-1}$$

Deze variantie is derhalve onafhankelijk van x .

Inverse schatter

Afgeleid als gewogen gemiddelde van μ_x en $\tilde{x}_{GLS}$:

$$\begin{aligned} \tilde{x}_{BLP} &= (D^{-1} + B\Sigma^{-1}B')^{-1} [D^{-1}\mu_x + B\Sigma^{-1}B' \tilde{x}_{GLS}] \\ &= [I - DB(B'DB + \Sigma)^{-1}B']\mu_x + DB(B'DB + \Sigma)^{-1}B' \tilde{x}_{GLS} \\ &= (I - A)\mu_x + A \tilde{x}_{GLS} \\ &= \tilde{x}_{GLS} - (I-A)(\tilde{x}_{GLS} - \mu_x) \end{aligned}$$

$$\begin{aligned}
\text{want} \quad B\Sigma^{-1}(B'DB) &= (B\Sigma^{-1}B')DB \\
\text{en} \quad B\Sigma^{-1}\Sigma &= D^{-1}DB \\
\text{dus} \quad B\Sigma^{-1}(B'DB+\Sigma) &= (B\Sigma^{-1}B' + D^{-1})DB \\
\text{waaruit volgt } (B\Sigma^{-1}B' + D^{-1})^{-1}B\Sigma^{-1}B' &= DB(B'DB + \Sigma)^{-1}B' = A
\end{aligned}$$

$$\begin{aligned}
\text{en} \quad (D^{-1} + B\Sigma^{-1}B')^{-1}(D^{-1} + B\Sigma^{-1}B') &= I \\
\text{dus} \quad (D^{-1} + B\Sigma^{-1}B')^{-1}D^{-1} &= I - (D^{-1} + B\Sigma^{-1}B')^{-1}B\Sigma^{-1}B' \\
&= I - DB(B'DB + \Sigma)^{-1}B' = I - A
\end{aligned}$$

Afgeleid uit regressie van x op y :

$$\begin{aligned}
\tilde{x}_{BLP} &= \mu_x + DB(B'DB+\Sigma)^{-1}(y-\mu_y) \\
&= [I - DB(B'DB+\Sigma)^{-1}B']\mu_x + DB(B'DB+\Sigma)^{-1}(y-\alpha) \\
&= (I - A)\mu_x + (B\Sigma^{-1}B' + D^{-1})^{-1}B\Sigma^{-1}B'(B\Sigma^{-1}B')^{-1}B\Sigma^{-1}(y-\alpha) \\
&= (I - A)\mu_x + (B\Sigma^{-1}B' + D^{-1})^{-1}B\Sigma^{-1}B' \tilde{x}_{GLS} \\
&= (I - A)\mu_x + A\tilde{x}_{GLS}
\end{aligned}$$

Uiteraard gaat dit alleen op als $B\Sigma^{-1}B'$ inverteerbaar is, dus als $q \geq p$. De regressiegedaante van $\tilde{x}_{BLP}$ kan ook rechtstreeks worden verkregen uit de voorstelling van $\tilde{x}_{BLP}$ als gewogen gemiddelde van μ_x en $\tilde{x}_{GLS}$, door daarin $(B\Sigma^{-1}B')\tilde{x}_{GLS}$ te vervangen door $B\Sigma^{-1}(y-\alpha)$. Dit levert ook voor $p > q$ een zinvolle uitdrukking op.

Voorts kan men afleiden:

$$\begin{aligned}
\tilde{x}_{BLP} - x &= -(I - A)(x - \mu_x) + DB(B'DB+\Sigma)^{-1}e \\
E\tilde{x}_{BLP} \Big| x &= (I - A)\mu_x + Ax \\
&= x - (I - A)(x - \mu_x) \\
E \text{ MSE } \tilde{x}_{BLP} &= E(\tilde{x}_{BLP} - x)(\tilde{x}_{BLP} - x)' \\
&= (I - A)D(I - A)' + DB(B'DB+\Sigma)^{-1}B'D \\
&= (I - A)D(I - A)' + (I - A)DA' \\
&= (I - A)D \\
&= (D^{-1} + B\Sigma^{-1}B')^{-1} \\
&= A(B\Sigma^{-1}B')^{-1}
\end{aligned}$$

Men vindt hiervoor derhalve de variantie-covariantiematrix van de rest-termen in het model voor regressie van x op y .

REFERENTIES

- Brown, P.J. - Multivariate calibration (with discussion).
J.R. Statist. Soc. B44, 287-321 (1982).
- Buonaccorsi, J.P. - Design considerations for calibration.
Technometrics 28, 148-155 (1986).
- Engel, B. - The linear functional relationship model for two variables.
IWIS-rapport A 84 ST 85 (1984).
Kwantitatieve methoden 19, 3-29 (1985).
- Jansen, A.A.M. - Some practical problems in establishing functional relationships. IWIS-rapport A 83 ST 90 51 (1983).
- Naes, T. - Comparison of approaches to multivariate linear calibration.
Biom.J. 27, 265-275 (1985a).
- Naes, T. - Multivariate calibration when the error covariance matrix is structured. Technometrics 27, 301-311 (1985b).
- Oman, S.D. & Y. Wax - Estimating fetal age by ultrasound measurement:
an example of multivariate calibration.
Biometrics 40, 947-960 (1984)
Biometrics 41, 821-822 (1985, correction).
- Rosenblatt, J.R. & C.H. Spiegelman - Discussion on a paper by Hunter and Lamboy. Technometrics 23, 329-333 (1981).

Ontvangen: 01-10-1986
Geaccepteerd: 08-05-1987