KM 12(1983)
pag 20-31

MEETFOUT GEMETEN

Aldi J.M. Hagenaars en Bernard M.S. van Praag*

1. INLEIDING

Bij het verzamelen en analyseren van inkomensgegevens moet in het algemeen rekening worden gehouden met de aanwezigheid van meetfouten. Deze meetfouten kunnen verschillende oorzaken hebben.

In de eerste plaats kunnen meetfouten ontstaan door een foutieve opgave van het inkomen door de respondenten, die hun inkomen zelden exact kennen.

In de tweede plaats kunnen door fouten bij de technische verwerking van de data meetfouten ontstaan, die niet altijd met behulp van standaardcontroles kunnen worden opgespoord.

Een derde bron van meetfouten is classificatie. Respondenten wordt vaak gevraagd aan te geven tot welke inkomensklasse hun inkomen behoort; door deze groepering van inkomens ontstaan eveneens meetfouten.

In dit artikel zal een methode worden beschreven om deze meetfouten te schatten, met een toepassing op een Nederlandse steekproef uit 1975.

De onderzoeksresultaten blijken zowel relevant voor de methode van dataverzameling als voor de analyse van de verkregen gegevens.

De meeste in dit artikel beschreven resultaten zijn eerder gepubliceerd in Econometrica (Van Praag, Hagenaars en Van Eck (1983)). Voor de technische details verwijzen we naar deze eerdere publikatie.

* Centrum voor Onderzoek van de Economie van de Publieke Sector, Rijksuniversiteit Leiden, Hugo de Grootstraat 32, 2311 XK LEIDEN. Tel. 071 - 149641.

2. MEETFOUT DOOR FOUTIEVE INKOMENSOPGAVE

Laat het echte inkomen van een respondent weergegeven worden door ξ, en de observatie van dit inkomen door x.
Veronderstel tevens dat de respondenten hun inkomen niet systematisch onder- of overschatten, maar dat een eventuele meetfout wel afhankelijk zal zijn van het inkomensniveau.
Dit is weer te geven in een multiplicatief meetfoutenmodel

$$x = \xi u \tag{1}$$

Wanneer aan beide kanten van deze vergelijking de natuurlijke logaritme wordt genomen levert dit op

$$y = \eta + \varepsilon \tag{2}$$

waarbij $y = \ln x$
$\eta = \ln \xi$
$\varepsilon = \ln u$

We nemen aan dat η en ε onafhankelijk verdeeld zijn, en dat de individuele ε's onderling onafhankelijk en normaal verdeeld zijn met verwachting 0 en variantie σ^2_ε.
Definieer vervolgens

$$\begin{aligned} \mu_y &= \int_{-\infty}^{\infty} y\,dF(y) \\ \mu_\eta &= \int_{-\infty}^{\infty} \eta\,dG(\eta) \\ \sigma^2_y &= \int_{-\infty}^{\infty} (y-\mu_y)^2\,dF(y) \\ \sigma^2_\eta &= \int_{-\infty}^{\infty} (\eta-\mu_\eta)^2\,dG(\eta) \end{aligned} \tag{3}$$

waarbij F(.) en G(.) de verdelingsfunctie van respectievelijk het geobserveerde en het echte log-inkomen weergeven.
Onder de bovengenoemde veronderstellingen met betrekking tot de storingsterm ε geldt nu

(4) $\mu_y = \mu_\eta$

(5) $\sigma^2_y = \sigma^2_\eta + \sigma^2_\varepsilon$

Indien de parameters van de verdeling van het echte inkomen, μ_η en σ^2_η, worden geschat door de geobserveerde waarden μ_y en σ^2_y, wordt het gemiddeld log-inkomen wel zuiver geschat, maar de variantie van de echte verdeling systematisch overschat. Deze variantie van de log-inkomens kan worden gezien als een maatstaf voor inkomensongelijkheid; de inkomensongelijkheid wordt dus overschat met een term die de variantie van de logaritme van de meetfout weergeeft.
De constatering dat de aanwezigheid van meetfouten leidt tot overschatting van de inkomensongelijkheid kan worden gegeneraliseerd naar een groot aantal andere inkomensongelijkheidmaatstaven; zie hiervoor Van Praag, Hagenaars en Van Eck (1983).
Indien de variantie van de logaritme van de meetfout kan worden geschat, kunnen de variantie van de log-inkomens en andere ongelijkheidsmaatstaven worden gecorrigeerd. Een schatting van de meetfoutvariantie is mogelijk indien twee observaties van het log-inkomen bekend zijn, aangeduid met y_{1n} en y_{2n} voor elke respondent n, n = 1,.....,N.

Indien elke observatie een meetfout bevat krijgen we

(6) $y_{1n} = \eta_n + \varepsilon_{1n}$ $\qquad n = 1,.....,N$

$y_{2n} = \eta_n + \varepsilon_{2n}$

We veronderstellen weer onafhankelijkheid van η en ε_i, en

$$\begin{bmatrix} \varepsilon_{1n} \\ \varepsilon_{2n} \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma^2_{\varepsilon_1} & r\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} \\ r\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \sigma^2_{\varepsilon_2} \end{bmatrix})$$

voor alle observaties, waarbij r de correlatiecoefficient tussen de beide meetfouten is.
Gelijkstelling van populatie- met steekproefmomenten van de eerste en tweede orde van y_1 en y_2 levert de volgende vergelijkingen op

(7) $\hat{\mu}_\eta = (\bar{y}_1 + \bar{y}_2)/2$

$$\begin{bmatrix} \hat{\sigma}^2_\eta + \hat{\sigma}^2_{\varepsilon_1} & \hat{\sigma}^2_\eta + \hat{r}\hat{\sigma}_{\varepsilon_1}\hat{\sigma}_{\varepsilon_2} \\ \hat{\sigma}^2_\eta + \hat{r}\hat{\sigma}_{\varepsilon_1}\hat{\sigma}_{\varepsilon_2} & \hat{\sigma}^2_\eta + \hat{\sigma}^2_{\varepsilon_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{(N-1)} \Sigma (y_{1n}-\bar{y}_1)^2 & \frac{1}{(N-1)} \Sigma (y_{1n}-\bar{y}_1)(y_{2n}-\bar{y}_2) \\ \frac{1}{(N-1)} \Sigma (y_{1n}-\bar{y}_1)(y_{2n}-\bar{y}_2) & \frac{1}{(N-1)} \Sigma (y_{2n}-\bar{y}_2)^2 \end{bmatrix}$$

Dit stelsel levert vier onafhankelijke vergelijkingen op, die niet voldoende zijn om de vijf onbekende parameters μ_η, σ^2_η, $\sigma^2_{\varepsilon_1}$, $\sigma^2_{\varepsilon_2}$ en r te schatten.

Wel is het mogelijk om bij een gegeven waarde van r de overige parameters te berekenen; in het kader van de door ons veronderstelde normaliteit kan de waarde van r niet uit hogere orde momenten worden geschat.

Het model kan eenvoudig worden uitgebreid indien de veronderstelling $E(\varepsilon_i) = 0$ wordt gewijzigd in $E(\varepsilon_i) = \alpha$ voor één van beide observaties; bijvoorbeeld omdat verwacht wordt dat één van beide observaties een systematische onderschatting van het echte inkomen zal zijn (zodat $\alpha<0$).

3. MEETFOUT DOOR CLASSIFICATIE

Naast de hierboven beschreven meetfout als gevolg van vergissingen of onwetendheid bij de ondervraagde ontstaan ook dikwijls meetfouten als gevolg van classificatie.

Inkomensgegevens kunnen in de vorm van een frequentie verdeling per inkomensklasse bij de onderzoeker terecht komen omdat de dataverzamelende instantie continue individuele gegevens heeft gegroepeerd, of omdat de vraagstelling bij het interview zelf op inkomensklassen was gericht.

In het eerste geval zijn meestal wel gegevens over gemiddelden en spreiding van inkomens in een klasse bekend; in het laatste geval beschikken we niet over deze informatie.

Laat het aantal respondenten n_k per inkomensklasse k $(k = 1,\ldots,K)$ bekend zijn. De natuurlijke logarithme van de ondergrens van klasse k wordt aangegeven met c_k, $k = 1,\ldots,K$.

De ondergrens van de laagste klasse, c_1, wordt per definitie op nul gesteld, en de bovengrens van de hoogste klasse op ∞.

Indien nu ook de verwachte waarde van de log-inkomens per klasse bekend is, aan te geven met $E(c_k)$, kan de verwachte waarde van alle log-inkomens worden geschreven als

$$\mu_y = \int_0^{\infty} y\,dF(y)$$

$$(8) \qquad = \sum_{k=1}^{K} \int_{c_k}^{c_{k+1}} y\,dF(y)$$

$$= \sum_{k=1}^{K} \int_{c_k}^{c_{k+1}} dF(y) \cdot \frac{\int_{c_k}^{c_{k+1}} y\,dF(y)}{\int_{c_k}^{c_{k+1}} dF(y)}$$

Een consistente schatter is

$$(9) \qquad \hat{\mu}_y = \sum_{k=1}^{K} \frac{n_k}{N} E(c_k)$$

waarbij $N = \sum_{k=1}^{K} n_k$.

De variantie van de log-inkomens kan worden geschreven als

$$\sigma_y^2 = \int_0^{\infty} (y - \mu_y)^2 \, dF(y)$$

$$(10) \qquad = \sum_{k=1}^{K} (E(c_k) - \mu_y)^2 \int_{c_k}^{c_{k+1}} dF(y) + \sum_{k=1}^{K} \int_{c_k}^{c_{k+1}} (y - E(c_k))^2 dF(y)$$

$$= \sigma^2_{c(y)} + \sigma^2_{\delta}$$

De eerste term aan de rechterkant van deze vergelijking is de tussenvariantie, die met $\sigma^2_{c(y)}$ zal worden aangeduid. Voor dit deel van de totale variantie is een consistente schatter

$$(11) \qquad \hat{\sigma}^2_{c(y)} = \frac{1}{N} \sum_{k=1}^{K} (E(c_k) - \hat{\mu}_y)^2 \, n_k$$

De tweede term aan de rechterzijde van vergelijking (10), σ^2_{δ}, is echter niet bekend. Wanneer $\hat{\sigma}^2_{c(y)}$ als een schatter voor de totale variantie wordt gebruikt, leidt dit tot onderschatting van σ^2_y. (Deze conclusie geldt niet in z'n algemeenheid wanneer als benadering van de klassemiddens waarden worden gebruikt die sterk afwijken van $E(c_k)$).
Er zijn twee methoden om een schatter voor σ^2_{δ} te bepalen. De eerste is een verdelingsvrije methode: met zo min mogelijk veronderstellingen over de verdeling F(y) wordt een beneden- en een bovengrens van σ^2_{δ} afgeleid (zie bijvoorbeeld Gastwirth (1972a,1972b).
Een tweede methode leidt een schatter af uit eigenschappen van een specifieke verdeling van de inkomens: wanneer de parameters van zo'n verdeling geschat zijn kan σ^2_{δ} analytisch of numeriek berekend worden.

4. EEN COMBINATIE VAN MEETFOUTEN

De twee verschillende soorten meetfouten die hierboven beschreven zijn kunnen ook tegelijkertijd voorkomen.
Wanneer in een mondelinge of schriftelijke ondervraging de respondent gevraagd wordt aan te geven tot welke klasse zijn inkomen behoort, kan eerst een foutieve inkomensschatting worden gedaan, die vervolgens wordt geclassificeerd. Bij een dergelijke combinatie van meetfouten is niet a priori te voorspellen of de echte variantie σ^2_η zal worden onderschat of overschat door de geclassificeerde variantie $\sigma^2_{c(y)}$.
Behalve de relatieve grootte van de overschatting en onderschatting, aangegeven door resp. σ^2_ε en σ^2_δ, zijn uiteraard ook de absolute groottes van beide meetfout-varianties van belang.
Beide meetfout-varianties kunnen worden geschat indien van elke respondent twee observaties van het inkomen bekend zijn, één observatie op een continue schaal, en één observatie in inkomensklassen.
Over een dergelijk databestand beschikken wij in de vorm van een CBS-enquête uit 1975, uitgevoerd ten behoeve van het Consumenten Conjunctuur Onderzoek (CCO) en het onderzoek Inkomenswaardering (IW) dat door het CBS in samenwerking met het Centrum voor Onderzoek van de Economie van de Publieke Sector wordt gehouden.
Deze enquêté bestaat uit een mondeling interview waarin de CCO-vragen worden gesteld, en een schriftelijke IW-vragenlijst die na afloop wordt achtergelaten en door de respondent wordt teruggestuurd.
In het mondelinge CCO-deel wordt gevraagd aan te geven tot welke klasse het netto gezinsinkomen van de respondent behoort. In het schriftelijke deel wordt gevraagd zo exact mogelijk het netto gezinsinkomen op een continue schaal op te geven.
Wanneer we aannemen dat bij beide inkomensopgaven meetfouten mogelijk zijn, kunnen we dit weergeven met behulp van het volgende model:

$$(12) \qquad c(y_{1n}) = c(\eta_n + \varepsilon_{1n}) \qquad n = 1,\ldots,N$$

$$y_{2n} = \eta_n + \varepsilon_{2n}$$

De hierboven beschreven modellen kunnen nu worden toegepast om schattingen te vinden van de relevante parameters σ^2_η, $\sigma^2_{\varepsilon_1}$ en $\sigma^2_{\varepsilon_2}$.
We laten de mogelijkheid van verschillende meetfout-varianties toe omdat bij de schriftelijke enquête meer tijd is om na te denken en even-

tueel inkomensgegevens op te zoeken; bovendien worden respondenten extra attent gemaakt op alle onderdelen die tot het inkomen dienen te worden gerekend (vakantiegeld, kinderbijslag). We vermoeden op grond hiervan dat $\sigma^2_{y_1} > \sigma^2_{y_2}$ en $E(y_1) < E(y_2)$.
Dit laatste kan worden opgenomen door te veronderstellen

$$E(\varepsilon_1) = \alpha \qquad (\alpha < 0)$$
$$E(\varepsilon_2) = 0$$

We nemen aan dat η niet gecorreleerd is met ε_1 en ε_2. Ook over de verdeling van η dient een veronderstelling te worden gemaakt. We hebben geëxperimenteerd met zowel een gamma-verdeling als een log-normale verdeling van de echte inkomens ξ.
De gamma-verdeling bleek een aanzienlijk slechtere beschrijving van de data, en was bovendien niet consistent met het geponeerde meetfoutenmodel; we vervolgen daarom met de veronderstelling dat ξ lognormaal verdeeld is, en η dus normaal $N(\mu_\eta, \sigma^2_\eta)$. (Voor de experimenten met de gamma-verdeling verwijzen we naar Van Praag, Hagenaars en Van Eck (1983)).
Deze verdelingsassumptie heeft ook consequenties voor de veronderstelling met betrekking tot de correlatie tussen ε_1 en ε_2: zoals hiervoor is aangeduid is de tweede-orde momentenmatrix niet voldoende voor het berekenen van σ^2_η, $\sigma^2_{\varepsilon_1}$, $\sigma^2_{\varepsilon_2}$ èn r. Aangezien de hogere orde momenten van de bivariaat normaal verdeelde y_1 en y_2 weer functies zijn van de tweede orde-momenten kan r niet worden geschat. In plaats van een schatting worden verschillende waarden van r vastgeprikt; de kleinste waarde stellen we op nul (geen correlatie tussen de meetfouten) en de grootste waarde stellen we op

$$\frac{\sigma_{y_1 y_2}}{\sigma_{y_1}\,\sigma_{y_2}} = \rho$$

de enige op dit moment nog onbekende parameter (alle correlatie tussen y_1 en y_2 is het gevolg van correlatie in de meetfouten).
De voorwaardelijke verdeling van y_{1n} gegeven y_{2n} is

$$N\left[\mu_\eta + \alpha + \rho \frac{\sigma_{y_1}}{\sigma_{y_2}} (y_{2n} - \mu_\eta); \sigma_{y_1}\sqrt{1-\rho^2}\right]$$

De log-aannemelijkheidsfunctie van de steekproef, die gemaximaliseerd

moet worden met betrekking tot ρ, kan worden vereenvoudigd tot

$$(13)\quad \ln(L) = \sum_{n=1}^{N} \ln \left[N \left\{ \frac{c_{k+1}(y_{1n}) - \hat{\mu}_\eta - \hat{\alpha} - \frac{\hat{\sigma}_{y_1}}{\hat{\sigma}_{y_2}} \rho \, (y_{2n} - \hat{\mu}_\eta)}{\sqrt{\sigma^2_{y_1}(1-\rho^2)}} ; 0,1 \right\} + \right.$$

$$\left. - N \left\{ \frac{c_k(y_{1n}) - \hat{\mu}_\eta - \hat{\alpha} - \frac{\hat{\sigma}_{y_1}}{\hat{\sigma}_{y_2}} \rho \, (y_{2n} - \hat{\mu}_\eta)}{\sqrt{\sigma^2_{y_1}(1-\rho^2)}} ; 0,1 \right\} \right]$$

Hierin is $\hat{\mu}_\eta = \frac{1}{N} \sum_{n=1}^{N} y_{2n}$

$$\hat{\alpha} = \hat{\mu}_{y_1} - \hat{\mu}_\eta$$

$$\hat{\sigma}^2_{y_2} = \frac{1}{N} \sum_{n=1}^{N} (y_{2n} - \hat{\mu}_\eta)^2$$

en zijn $\hat{\mu}_{y_1}$ en $\hat{\sigma}^2_{y_1}$ de schatters die gevonden worden door het aanpassen van een lognormale verdeling aan de geclassificeerde data.
Vergelijking (13) is de aannemelijkheidsfunctie van de voorwaardelijke verdeling van y_1, gegeven een waarneming y_2; deze aannemelijkheidsfunctie kan met behulp van een eenvoudig zoekprogramma worden gemaximaliseerd naar ρ.

5. SCHATTINGSRESULTATEN

In tabel 1 geven we de resultaten aan van de schattingen, waarbij we voor de correlatiecoëfficiënt r tussen ε_1 en ε_2 vijf verschillende waarden hebben gekozen van nul tot 0.838, de waarde van de correlatiecoëfficiënt ρ tussen y_1 en y_2, die bij het maximaliseren van vergelijking (13) is gevonden.

	$\rho = \frac{\sigma^2_\eta + r\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}}{\sigma_{y_1}\sigma_{y_2}}$	r				
		0	0.250	0.500	0.750	0.838
σ^2_η		0.149	0.139	0.120	0.061	0
$\sigma^2_{\varepsilon_1}$		0.033	0.043	0.062	0.121	0.182
$\sigma^2_{\varepsilon_2}$		0.025	0.035	0.054	0.113	0.174

$\hat{\mu}_\eta$	=	9.855
$\hat{\alpha}$	=	-0.037
$\hat{\sigma}^2_{y_1}$	=	0.182
$\hat{\sigma}^2_{y_2}$	=	0.174
$\hat{\sigma}_{y_1 y_2}$	=	0.149
N	=	1706
$\sigma^2_{c(y)}$	=	0.155

TABEL 1

We zien dat $\sigma^2_{y_1} > \sigma^2_{y_2}$ en $E(y_1)\,(= \hat{\mu}_\eta + \hat{\alpha}) < E(y_2)\,(= \hat{\mu}_\eta)$, zoals we op grond van de enquête-opzet reeds vermoedden.

Uit de tabel blijkt dat de meetfout als gevolg van foutieve inkomensopgave zeker niet te verwaarlozen is. In het gunstigste geval, wanneer de meetfouten onderling niet gecorreleerd zijn, is het aandeel van de meet-

foutvariantie in de totale inkomensvariantie 0.18 voor de eerste en 0.14 voor de tweede inkomensopgave.
Dit aandeel loopt sterk op wanneer de meetfouten positief gecorreleerd zijn; als de correlatie tussen de meetfouten gelijk is aan 0.50 zijn deze aandelen al respectievelijk 0.34 en 0.31.
Wanneer de meetfouten onderling in dezelfde mate gecorreleerd zijn als de twee inkomensmetingen dan bestaat de logvariantie alleen nog maar uit meetfoutvariantie.
De conclusie is dat foutieve opgaven inderdaad leiden tot een overschatting van de echte inkomensongelijkheid; in het gunstigste geval is dit een overschatting met $\frac{\sigma^2_{y_2}}{\sigma^2_{\eta}}$, dus met ca. 17%.

Wanneer we $\sigma^2_{y_2}$ vergelijken met $\sigma^2_{c(y_2)}$ zien we dat classificatie inderdaad leidt tot onderschatting; de geclassificeerde variantie is 89% van de totale variantie. Als we echter $\sigma^2_{c(y)}$ vergelijken met σ^2_{η}, de "echte" logvariantie, dan blijkt het effect van de foutieve inkomensopgave te domineren: $\sigma^2_{\eta} < \sigma^2_{c(y)}$. De geclassificeerde variantie is een overschatting van de echte variantie; de overschatting is echter teruggebracht tot $\frac{\sigma^2_{c(y)}}{\sigma^2_{\eta}}$, dat wil zeggen tot 4%.

Deze conclusie geldt ook wanneer de berekeningen voor verschillende subgroepen worden uitgevoerd. (Zie Van Praag, Hagenaars en Van Eck (1983)).
De grootte van de meetfout varieert echter sterk met het aantal kostwinners (meer kostwinners, grotere meetfout) en het beroep (zelfstandigen hebben een aanzienlijk grotere meetfout dan loontrekkers).

6. CONCLUSIE

In het voorgaande is een methode besproken die het mogelijk maakt meetfouten in inkomensopgaven te schatten. Het procedé is eenvoudig: in enquêtes wordt getracht tweemaal een inkomensopgave te krijgen, waarvan ten minste één op een continue schaal wordt gemeten.
Deze opgaven worden met elkaar vergeleken, waarbij zonodig een assumptie over de verdeling van de inkomens wordt gemaakt (die later getoetst kan worden). Op deze wijze kan een schatting verkregen worden van de grootte van de meetfout bij de continue inkomensschatting; deze meetfoutvariantie geeft allereerst informatie over de betrouwbaarheid van de data en kan tevens gebruikt worden om bijvoorbeeld inkomensongelijkheidsmaatstaven te corrigeren.
Tenslotte kan deze variantie worden gebruikt in errors-in-variables modellen bij regressievergelijkingen waarbij het inkomen een verklarende variabele is. De eenvoud van deze methode en het grote aantal toepassingsmogelijkheden zijn sterke aanbevelingen om haar veelvuldig in de statistische praktijk te gebruiken.

REFERENTIES

Gastwirth, J.L. (1972a), "The estimation of the Lorenz-curve and Gini Index", Review of Economics and Statistics, vol. 54, pp. 306-316.

Gastwirth, J.L. (1972b), "A New Goodness of Fit Test", *Business and Economics Statistics Proceedings of the American Statistical Association.*

Van Praag, B.M.S., A.J.M. Hagenaars en W. Van Eck (1983), "The Influence of Classification and Observation Errors on the Measurement of Income Inequality", Econometrica, Vol. 51, no. 4 (July, 1983)

Ontvangen: 24-2-1983