

Polynome Regressie-analyse met meetfouten in de variabelen

Marco Vriens^{*}, Kees van Montfort^{**}, en Ab Mooijart^{**}

Abstract

As is known for a long time, serious identification problems exist in the linear regression model with "errors-in-the-variables" (EIV). As might be expected these same problems exist in the polynomial regression model with EIV. In this study we examine a bivariate polynomial EIV regression model of order two. By analyzing only first and second order moments (i.e. means and covariances) one cannot identify the model without making very unrealistic assumptions. However, it is possible to use additional information in the data. This can be done in several ways, but we shall only apply one of them. We show that with the help of moments up to order three it is possible to identify the model, and to derive consistent moment estimators. Furthermore we over-identify the model by making one additional assumption about the distribution of the errors. For this over-identified model we show how an optimal solution can be obtained with the help of the generalized least squares method.

Key words: polynomial regression analysis with "errors-in-the-variables", third order moments, BAN-estimates, BGLS-estimates.

* K.U.B. Faculteit Sociale Wetenschappen, Vakgroep Sociale Zekerheidswetenschap, Hogeschoollaan 225, 5037 gc Tilburg. Tel: 013-663001.

** R.U.L. Faculteit Sociale Wetenschappen, Vakgroep Methoden en Technieken, Hooigracht 15, 2312 km Leiden. Tel: 071-273768

Inleiding

In deze studie onderzoeken we het bivariate tweedegraads polynome regressiemodel met meetfouten in de variabelen. Het bivariate tweedegraads polynome regressiemodel met "error-in-the-variables" (EIV) (Model 1) ziet er als volgt uit :

$$y = \alpha + \beta_1 \xi + \beta_2 \xi^2 + \epsilon$$

$$x = \xi + \delta$$

Hierbij zijn x en y twee geobserveerde "random" variabelen die uitgedrukt worden in termen van ongeobserveerde of latente variabelen ξ , ϵ , en δ . Het intercept wordt gerepresenteerd door α . De regressiecoëfficiënten worden gerepresenteerd door β_1 en β_2 .

Aan het lineaire model met EIV (zie Van Montfort e.a., 1987) wordt een kwadratische term toegevoegd. We laten zien dat ook dit model geïdentificeerd kan worden met behulp van eerste, tweede, en derde orde momenten van de geobserveerde variabelen (sectie 2). Verder laten we in sectie 3 zien dat, in de situatie van symmetrisch verdeelde meetfouten, op numerieke wijze een optimale schatter geconstrueerd kan worden, als blijkt dat dit niet op algebraïsche wijze geschieden kan zoals in Van Montfort e.a. (1987).

1 Polynome regressieanalyse met EIV

Er is weinig verschenen over polynome regressieanalyse met EIV. Het onderwerp komt onder andere ter sprake in Griliches & Ringstad (1970), Kenny & Judd (1984), Aigner e.a. (1984), en Mooijaart en Bentler (1987).

De eerste publicatie over non-lineaire regressie-analyse met EIV is van Griliches en Ringstad (1970). Deze research noot is een bijproduct van een meer omvattende studie en de auteurs laten zien dat de OLS-schatters van de regressie coëfficiënten in een polynoom EIV regressiemodel "biased" zijn naar 0 toe. De ols-schatter van de regressiecoëfficiënt die bij de kwadratische term hoort is "biased". "(...) the nonlinear term (...) is also biased towards zero but as the square of the bias factor of the linear term" (Griliches & Ringstad, 1970). Bovendien gaan zij ervan uit dat de geobserveerde variabelen en de latente variabelen normaal verdeeld zijn. In

deze studie gebruiken we voor de identificatie van model 1 geen aanname over de verdeling van de geobserveerde variabelen.

In de tweede referentie van Kenny & Judd (1984) wordt onderzocht hoe men niet-lineaire en interactieve effecten van latente variabelen kan schatten in het structurele model. Hun benadering wordt echter gekenmerkt door het gebruik van producten van de geobserveerde variabelen. Deze producten worden gebruikt als indicatoren voor latente product variabelen. De als derde genoemde publicatie van Aigner e.a. (1984) geeft een aantal literatuurverwijzingen waarbij uitgegaan wordt van een specifiek polynoom model (de auteurs zelf veronderstellen een algemene regressiefunctie $f(\xi, \epsilon)$). Deze literatuurverwijzingen zijn voor ons echter niet interessant omdat wij niet, zoals bij deze referenties, er van uitgaan dat de variantie-covariantie matrix van de errors bekend is. De vierde publicatie waarin het onderwerp onder de noemer "random polynomial factor analysis" behandeld wordt is het artikel van Mooijaart & Bentler (1987). In dit artikel wordt gedemonstreerd hoe men via het gebruik van hogere orde momenten de oneindige verzameling oplossingen, die men bij factoranalyse meestal heeft en die normaliter gereduceerd worden tot een enkele door middel van rotatietechnieken, terug kan brengen tot een eindig aantal. In dit artikel bracht men voor een specifiek factormodel het aantal oplossingen terug tot 4, waarin er twee de spiegeloplossing van de andere twee waren. Men liet zien dat een "polynome factor" in een voorbeeld een zelfde orde van variantie verklaarde als twee "lineaire factoren". Bovendien liet men zien dat een "polynome factor" beter interpreteerbaar was.

2 Identificatie van het bivariate polynome model

Voor de identificatie van het polynome regressie-model met EIV (Model 1) gaan we de populatiemomenten van x en y uitdrukken in modelparameters. Hierbij gaan we uit van volgende aannamen:

A1 : De latente variabelen ϵ en δ zijn onafhankelijk van elkaar en van ξ .

A2 : $E(\epsilon) = E(\delta) = 0$

Verder nemen we aan, zonder verlies van algemeenheid, dat $E(\xi) = 0$. Als $E(\xi)$ niet gelijk aan nul is dan kunnen we de $E(\xi)$ rechtstreeks schatten met behulp van het steekproefgemiddelde van x .

De notatie voor de verschillende momenten is als volgt: $\mu_n = E(x^n y^n)$, $\lambda_n = E(\xi^n)$, $\omega_n = E(\delta^n)$ en $\zeta^n = E(\epsilon^n)$. Opgemerkt zij dat als gevolg van A2 alle centrale momenten gelijk zijn aan de corresponderende niet-centrale momenten.

We nemen verder aan dat alle momenten die in onze vergelijkingen voorkomen bestaan.

Wanneer we de eerste, tweede, en derde orde populatie-momenten van x en y uitdrukken in model-parameters krijgen we een onder-geïdentificeerd model (zie Vriens e.a., 1987). We hebben dan acht vergelijkingen met twaalf onbekenden. Om het model te identificeren zijn we aangewezen op twee methoden te weten: 1) het gebruik van extra informatie in de data, en 2) het gebruik van extra model aannamen (zie Van Montfort, 1987).

De methode om gebruik te maken van extra informatie die in de data zit is een mogelijkheid. Het fitten van nog hogere orde momenten zou echter rekenkundig erg uit de hand lopen en zou bovendien leiden tot zeer instabiele schatters. Het gebruik van instrumentele variabelen (zie Sargan, 1958; en Pal, 1981) als oplossing voor het identificatieprobleem ligt niet zo voor de hand, omdat instrumentele variabelen niet altijd aanwezig zijn. Het gebruik van karakteristieke functies bovenop het gebruik van derde orde momenten is in principe ook een mogelijkheid die echter in deze studie niet onderzocht zal worden. Van Montfort (1987) heeft deze methode toegepast om een lineair regressie-model met EIV te identificeren.

We kunnen echter ook het model identificeren door gebruik te maken van extra aannamen. Het model kan in ons geval geïdentificeerd worden door toevoeging van een extra aanname.

Als we het model exact-geïdentificeerd (just-identified) willen hebben moeten we vier onbekenden parameters elimineren. Wanneer we aannemen dat ξ normaal verdeeld is dan kunnen we alle hogere momenten (> 2) van ξ herleiden tot functies van λ_2 (Johnson, N.L & Kotz, S., 1970 p. 47). Het stelsel van eerste, tweede, en derde orde populatie-momenten bevat nu acht vergelijkingen met acht onbekenden, en we kunnen laten zien dat de model-parameters in dit model geïdentificeerd zijn (zie Vriens e.a., 1987).

3 Het geval van symmetrisch verdeelde meetfouten

Een model dat evenveel parameters bevat als de data die het moet representeren is minder interessant dan een model dat minder parameters bevat dan de data. Wanneer we te maken hebben met symmetrisch verdeelde meetfouten hebben we een model dat over-geïdentificeerd is. De symmetrisch verdeelde meetfouten impliceren dat de ω_1 en S_1 gelijk aan 0 zijn. We houden dan zeven vergelijkingen over met zes onbekenden. Onder de aannamen dat ξ normaal verdeeld is en dat de meetfouten symmetrisch verdeeld zijn komt de moment-structuur er als volgt uit te zien:

- (1) $\mu_{01} = \alpha + \beta_2 \lambda_2$
- (2) $\mu_{20} = \lambda_2 + \omega_2$
- (3) $\mu_{11} = \beta_1 \lambda_2$
- (4) $\mu_{02} = \alpha^2 + (\beta_1^2 + 2\alpha\beta_2)\lambda_2 + 3\beta_2^2 \lambda_2^2 + S_2$
- (5) $\mu_{21} = (\alpha + \beta_2 \omega_2)\lambda_2 + 3\beta_2 \lambda_2^2 + \alpha \omega_2$
- (6) $\mu_{12} = 2\alpha\beta_1 \lambda_2 + 6\beta_1 \beta_2 \lambda_2^2$
- (7) $\mu_{03} = \alpha^3 + (\alpha^2 \beta_2 + \alpha \beta_1^2 + \beta_2 S_2)3\lambda_2 + (\alpha \beta_2^2 + \beta_1^2 \beta_2)9\lambda_2^2 + 15\beta_1^3 \lambda_2^3 + 3\alpha S_2$

Het ontbreken van een unieke oplossing impliceert, dat we met behulp van derde orde momenten meerdere consistente schatters voor β_1 en β_2 kunnen afleiden. Het construeren van een optimale schatter in de set van consistente derde orde schatters van β_1 en β_2 kan op verschillende manieren gebeuren. Van Montfort e.a. (1988) construeren, in het lineaire bivariate regressie model, op algebraïsche wijze een optimale derde orde schatter. Voor ons model zijn echter niet alle consistente derde orde basis schatters op algebraïsche wijze te bepalen, en daarom gebruiken we de GLS-methode van Browne (1984).

Door middel van de methode van Browne (1984) worden de zogeheten "Generalized Least Squares"-schattingen (GLS) verkregen. Browne bespreekt een klasse van BGLS-schatters die asymptotisch minimale variantie hebben (Mooijjaart, 1985). Browne's klasse van BGLS-schattingen is verschillend van BAN-schattingen vanwege het feit dat de GLS-methode alleen BAN-schatters oplevert als $\mathbf{m}$ (zie (8)) een vector is bestaande uit "sufficient statistics". Een voordeel van de methode van Browne is dat standaardfouten van de schattingen berekend kunnen worden en dat de fit van het model getoetst kan worden. In deze methode moet onderstaande functie geminimaliseerd worden:

$$(8) \quad X^2 = (m - \mu(\eta))' W^{-1} (m - \mu(\eta))$$

Waarbij m een kolomvector is die de steekproefschattingen bevat van de momenten die we gebruikt hebben. In de kolomvector $\mu(\eta)$ zitten de populatiemomenten als functies van de latente parameters die verzameld zijn in de vector η . De matrix W is een gewichtenmatrix. Wanneer we X-kwadraat naar de waarden van de latente parameters minimaliseren krijgen we 'generalized least squares'-schattingen. Indien voor W de identiteitsmatrix genomen wordt levert minimalisatie van X-kwadraat naar η OLS-schattingen op. Wanneer we voor W de inverse van een consistente schatter van de variantie-covariantie matrix van m nemen dan vinden we BGLS-schattingen voor de latente parameters. In dit geval is X-kwadraat asymptotisch chi-kwadraat verdeeld met 1 vrijheidsgraad. We kunnen dan een specifiek model toetsen. Zo is het mogelijk om te toetsen of de kwadratische term in het model of de meetfout op de x-variabele in het model significant van 0 afwijkt.

4 Minimalisatie van de GLS-verliesfunctie

Voor het minimaliseren van de GLS-verliesfunctie bestaan een aantal gradient methoden die gebruikt kunnen worden om algemene non-lineaire functies te minimaliseren, zie Walsh (1979). We hebben in ons onderzoek de Davidon-Fletcher-Powell methode (DFP-methode) geprogrammeerd (Walsh, 1979), waarmee het lukte om een lokaal minimum van de GLS-verliesfunctie te vinden.

In Vriens e.a. (1987) staan de resultaten van een Monte Carlo studie vermeld. Hierbij is nagegaan hoe goed de GLS-methode in staat is om de ware modelparameters terug te vinden uit data die uit een bekend model gegenereerd zijn. Uit deze studie blijkt dat alle GLS-schaters van de modelparameters, behalve de schatter van de variantie van de meetfout ϵ , zich "goed" gedragen bij steekproeven groter dan 100. De GLS-schatter voor de variantie van ϵ gedraagt zich pas bevredigend bij steekproeven groter dan 1000.

Vervolgonderzoek zal zich enerzijds richten op de multi-variate uitbreiding van model 1, en tevens zal onderzocht worden in hoeverre de variantie van de meetfout op de y-variabele stabiel geschat kan worden.

5 Geraadpleegde Literatuur

- Aigner, D.J., C. Hsiao, A. Kapteyn, & T. Wansbeek (1984). Latent variable models in Econometrics. In: Griliches, Z. & M.D. Intriligator (Eds.), Handbook of Econometrics. North Holland P.C., Amsterdam.
- Browne, M.W. (1984). Asymptotically distribution free methods for the analysis of covariance structures, British Journal of Mathematical and Statistical Psychology, vol. 37, pp. 62-82.
- Griliches, Z. & Ringstad, V. (1970). Errors-in-the-variables bias in non-linear contexts, Econometrica, vol. 38, no. 2.
- Johnson, N.L. & Kotz, S. (1970). Continuous Univariate Distributions-I. Boston: Houghton Mifflin.
- Kenny, D.A. & Judd, C.M. (1984). Estimating the Non-Linear and Interactive Effects of Latent Variables, Psychological Bulletin, vol. 96, No. 1, pp. 201-210.
- Montfort, K. van, Mooijaart, A. & De Leeuw, J. (1987). Regression with errors in variables: estimators based on third order moments, Statistica Neerlandica, vol. 4, pp. 223-237.
- Montfort, K. van (1987). Estimation of regression coefficients with the help of characteristics functions. Department of Psychology, Leiden University.
- Mooijaart, A. (1985). Factor analysis for non-normal variables, Psychometrica, vol. 50, pp. 323-342.
- Mooijaart, A. & Bentler, P. (1987). Random Polynomial Factor Analysis. In: E. Diday (Ed.), Data analysis and Informatics IV. Amsterdam: North Holland.
- Pal, M. (1981). Estimation in Error-in-Variables Models, Unpublished dissertation.
- Sargan, J.D. (1958). The estimation of Economic relationships using instrumental variables, Econometrica, vol. 26, pp. 393-415.
- Vriens, M. Montfort, K. van, Mooijaart, A. (1987). Bivariate polynome regressie-analyse met errors in de variabelen. Department of Psychology, Leiden University.
- Walsh, G.R. (1979). Methods of Optimization. New York: John Wiley & Sons.

Ontvangen:

Geaccepteerd: 19-05-1988