S 168 (M 70)

<u>Een exacte toets tegen kettingcorrelatie bij lineaire regressieproblemen van E.J. Hannan (1955)</u>

Stichting Mathematisch Centrum

Statistische Afdeling

1955

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
Amsterdam-O.

S 168 (M 70)

Een exacte toets tegen kettingcorrelatie bij lineaire regressieproblemen van E.J. HANNAN (1955) 1)

1. Model

We gaan uit van een lineair regressiesysteem:

(1) $$\underline{y}_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt} + \underline{v}_t \qquad (t = 1, \dots, 2n+1),$$

waarin $\beta_0, \beta_1, \dots, \beta_k$ onbekende regressie-coëfficiënten zijn en $(x_{j1}, \dots, x_{j,2n+1})$ $(j = 1, \dots, k)$ k bekende regressie-vectoren. Tussen de stochastische grootheden $\underline{v}_t$, die onafhankelijk van alle x_{jt} normaal verdeeld zijn met $\mathcal{E}\underline{v}_t = 0$ en gelijke spreidingen, kan een eerste orde kettingcorrelatie bestaan, d.w.z.:

(2) $$\underline{v}_t = \rho\, \underline{v}_{t-1} + \underline{u}_t \qquad (t = 1, \dots, 2n+1),$$

waarin $\underline{u}_1, \dots, \underline{u}_{2n+1}$ onderling onafhankelijk $N(0, \sigma_{\underline{u}}^2)$ 2) verdeeld zijn en $|\rho| < 1$ is.

Nu is het bij de toepassing van de in de regressieanalyse gebruikelijke toetsen (t-toets van STUDENT, F-toets van FISHER) voor het toetsen van hypothesen over de β's noodzakelijk dat de grootheden $\underline{v}_t$ onderling onafhankelijk verdeeld zijn. We zullen dus willen onderzoeken of de kettingcorrelatie-coëfficiënt $\rho = 0$ is. Hiervoor geeft E.J. HANNAN (1955), voortbouwend op resultaten van M. OGAWARA (1951) voor het geval k = 0 (Zie memorandum S 168 (M 69)) de volgende exacte toetsingsmethode.

2. Toetsingsmethode

De voorgestelde toetsingsmethode berust erop dat we de voorwaardelijke simultane verdeling van $\underline{y}_2, \underline{y}_4, \dots, \underline{y}_{2n}$ beschouwen onder voorwaarde van de gevonden waarden van $y_1, y_3, \dots, y_{2n+1}$. Onder deze voorwaarde zijn namelijk de grootheden $\underline{y}_{2s}$ $(s = 1, \dots, n)$ onderling onafhankelijk normaal verdeeld. Voor het gemiddelde en de variantie van $\underline{y}_{2s}$ kan men vrij eenvoudig afleiden:

(3) $$\mathcal{E}\{\underline{y}_{2s} \mid y_1, y_3, \dots, y_{2n+1}\} = \beta_0 \left(1 - 2\rho(1+\rho^2)\right) + \sum_{j=1}^{k} \beta_j x_{j,2s} + \rho(1+\rho^2)^{-1}(y_{2s-1} + y_{2s+1}) - \sum_{j=1}^{k} \rho(1+\rho^2)^{-1} \beta_j (x_{j,2s-1} + x_{j,2s+1}),$$

1) Dit memorandum is slechts bedoeld als oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) $N(\mu, \sigma^2)$ is een symbool voor de normale verdeling met gemiddelde μ en variantie σ^2.

(4) $$\sigma^2\{\underline{y}_{2s} \mid y_1, y_3, \ldots, y_{2n+1}\} = \sigma_u^2 (1+\rho^2)^{-1}$$

In plaats van (3) kunnen we ook schrijven

(5) $$\underline{y}_{2s} = \gamma_0 + \gamma_1 x_{1s} + \ldots + \gamma_{2k+1} x_{2k+1,s} + \underline{v}'_{2s} \qquad (s = 1, \ldots, n),$$

waarin dus:

$$\gamma_0 = \beta_0 \{1 - 2\rho(1+\rho^2)^{-1}\}$$

(6) $$\gamma_j = \begin{cases} \beta_j \\ 2\rho(1+\rho^2)^{-1} \\ -\gamma_{k+1}\beta_{j-k-1} \end{cases} \qquad x_{j,s} = \begin{cases} x_{j,2s} & (j = 1, \ldots, k) \\ 2^{-1}(y_{2s-1} + y_{2s+1}) & (j = k+1) \\ 2^{-1}(x_{j-k-1,2s-1} + x_{j-k-1,2s+1}) & (j = k+2, \ldots, 2k+1). \end{cases}$$

In het regressiesysteem (5) zijn nu de grootheden $\underline{v}'_{2s}$ wel onderling onafhankelijk normaal verdeeld $(N(0, \sigma_u^2 (1+\rho^2)^{-1})$. Maken we geen gebruik van de wetenschap dat $\gamma_{j+k+1} = -\gamma_{k+1}\gamma_j$ $(j = 1, \ldots, k)$ (zie (6)) dan kunnen voor het schatten van de γ's en voor het toetsen van hypothesen over de γ's de in de regressie-analyse (variantie-analyse) gebruikelijke methoden worden gebruikt (kleinste kwadraten schattingen, F-toetsen). In het bijzonder voor het toetsen van de hypothesen $\rho = 0$, dus $\gamma_{k+1} = 0$, vinden we als toetsingsgrootheid (zie H.B. MANN(1949), stelling 4,3):

$$\underline{F}_{1, n-2k-2} = (n-2k-2)\, \hat{\gamma}_{k+1}^2\, |L| \, \{|L_{k+1,k+1}|\, \underline{Q}\}^{-1},$$

waarin $\hat{\gamma}_{k+1}$ de kleinste kwadratenschatting voor γ_{k+1} voorstelt; dus het element op de $(k+1)^e$ rij van de $1 \times (2k+1)$ -matrix

$$\hat{\gamma} = (Z'Z)^{-1} Z' y$$

als

$$Z = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \ldots & x_{2k+1,1} - \bar{x}_{2k+1} \\ \vdots & \vdots & & \vdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \ldots & x_{2k+1,n} - \bar{x}_{2k+1} \end{pmatrix}, \quad y = \begin{pmatrix} y_2 - \bar{y} \\ y_4 - \bar{y} \\ \vdots \\ y_{2n} - \bar{y} \end{pmatrix}$$

met $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$ en $\bar{y} = \sum_{s=1}^{n} y_{2s}/n$ (zie memorandum S 168 (M71)); L de $(2k+1) \times (2k+1)$ -matrix $\{l_{ij}\}$ opgebouwd uit elementen

$$l_{ij} = \sum_{s=1}^{n} (x_{is} - \bar{x}_i)(x_{js} - \bar{x}_j)$$

(dus L = Z'Z): $L_{k+1,k+1}$ de minor behorende bij het element $l_{k+1,k+1}$ en $\underline{Q}$ tenslotte de som van de kwadraten van de residuen:

$$\begin{aligned} \underline{Q} &= \sum_{s=1}^{n} (y_{2s} - \hat{\gamma}_0 - \hat{\gamma}_1 x_{1s} \ldots - \hat{\gamma}_{2k+1} x_{2k+1,s})^2 \\ &= \sum_{s=1}^{n} (y_{2s} - \bar{y})^2 - \sum_{j=1}^{2k+1} \hat{\gamma}_j \sum_{s=1}^{n} (x_{js} - \bar{x}_j)(y_{2s} - \bar{y}). \end{aligned}$$

3. Opmerkingen

1. De schattingen $\hat{\gamma}_j$ voor γ_j $(j=1,\ldots,k)$ zijn tegelijkertijd schattingen voor β_j uit het oorspronkelijke model (1) (immers $\beta_j = \gamma_j$ voor $j=1,\ldots,k$, zie (6)). Deze schattingen bezitten echter wanneer ρ dicht bij 0 ligt asymptotisch voor grote steekproeven een grotere variantie dan de rechtstreekse kleinste kwadraten schattingen $\hat{\beta}_j$ uit (1), zijn dan dus minder doeltreffend (zie HANNAN (1955)). Ligt ρ dichtbij 1 (of ook, mits er geen kettingcorrelatie bestaat tussen $x_{j1},\ldots,x_{j,2n+1}$, dichtbij -1) dan is de schatting $\hat{\gamma}_j$ voor β_j doeltreffender dan $\hat{\beta}_j$ (d.w.z. bezit asymptotisch een kleinere variantie). In dit geval, ρ dicht bij 1 of -1, geeft bovendien de regressieanalyse toegepast op model (5) exacte toetsen voor hypothesen over de γ_j $(j=1,\ldots,k)$, dus de β_j, in tegenstelling tot regressieanalyse toegepast op model (1).

2. Als bijzonder geval vinden we bij k = 0 een toets tegen kettingcorrelatie voor een steekproef met model

$$\underline{y}_t = \beta_0 + \underline{v}_t \quad \text{met} \quad \mathcal{E}\,\underline{v}_t = 0 \qquad (t=1,\ldots,2n+1).$$

Dit wordt dan herleid tot een regressiemodel met één regressievector:

$$\underline{y}_{2s} = \beta_0\left(1-2\rho(1+\rho^2)^{-1}\right) + \rho(1+\rho^2)^{-1}(\underline{y}_{2s-1}+\underline{y}_{2s+1}) + \underline{v}'_{2s}$$

$(s=1,\ldots,n)$. Zie voor een iets uitvoeriger behandeling van dit geval: memorandum S 168 (M 69).

3. Aan de uitvoering van de toets van HANNAN zijn enkele BEZWAREN verbonden. In de eerste plaats moet, om de methode te kunnen toepassen, in model (5) het aantal regressie-coëfficiënten γ_j kleiner zijn dan het aantal waarnemingen y_{2s}; dit betekent dus dat $2k+2<n$. Komen dus in het oorspronkelijke model $k+1$ onbekende coëfficiënten voor, $\beta_0, \beta_1, \ldots, \beta_k$, dan moet het aantal waarnemingen y_t minstens $2(2k+3)+1 = 4k+7$ zijn.

In de tweede plaats moet bij de berekeningen van $\hat{\gamma}_j$ een matrix met $2k+1$ rijen en kolommen (de matrix L) geïnverteerd worden, wat wel een ernstig nadeel is t.o.v. bv. de toets van DURBIN & WATSON (1950 en 1951), of memorandum S 168 (M 66)), waarbij (voor het bepalen van de kleinste kwadratenschattingen $\hat{\beta}_1,\ldots,\hat{\beta}_k$) een matrix met slechts k rijen en kolommen geïnverteerd moet worden. Bovendien zal men indien de hypothese $\rho=0$ niet verworpen wordt in sommige gevallen toch alsnog deze schattingen $\hat{\beta}_1,\ldots,\hat{\beta}_k$ uit model (1) willen berekenen in verband met de grotere doeltreffendheid (zie opmerking 1).

Een derde nadeel is, dat de toets niet ondubbelzinnig is. Men kan nl. ook de waarnemingen $\underline{y}_3, \underline{y}_5, \ldots, \underline{y}_{2n-1}$ beschouwen onder voorwaarde van de gevonden waarden voor $y_2, y_4, \ldots, y_{2n}$ en dan kan men een andere uitkomst verkrijgen.

Daar staat tegenover dat de toets minder het karakter van een benadering draagt dan die van DURBIN en WATSON.

Literatuur

J. DURBIN and G.S. WATSON (1950 en 1951), Testing for serial correlation in least squares regression.
Part I, Biometrika, 37 (1950), pp. 409-428.
Part II, Biometrika, 38 (1951), pp. 159-178.

E.J. HANNAN (1955), Exact tests for serial correlation. Biometrika, 42, pp. 133-143.

H.B. MANN (1949), Analysis and design of experiments. New York, Dover Publications, Inc.

M. OGAWARA (1951), A note on the test of serial correlation coefficients. Annals, 22, pp 115-118.

Memoranda van de Statistische Afdeling van het Mathematisch Centrum te Amsterdam (1955)

S 168 (M 66) Een toets tegen kettingcorrelatie bij lineaire regressie van J. DURBIN en G.S. WATSON.

S 168 (M 69) Een exacte toets tegen kettingcorrelatie van M.OGAWARA (1951).

S 168 (M 71) De kleinste kwadratenschattingen voor de regressie coëfficiënten in lineaire regressiesystemen.