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

S 168[A] (M 66)

Een toets tegen kettingcorrelatie bij lineaire regressie van J. DURBIN en G.S. WATSON (1950 en 1951).[1)]

1. Inleiding

Indien er in een reeks waarnemingen correlatie bestaat tussen opeenvolgende waarnemingen en geen correlatie of een veel zwakkere tussen verder uit elkaar gelegen waarnemingen, dan spreken wij van kettingcorrelatie (Engels: serial correlation). Een onderzoek naar deze correlatie is vooral van belang bij tijdreeksen waarop variantie- of regressieanalyse toegepast zal worden, zoals b.v. bij economische problemen vaak voorkomt.

Wij gaan uit van het regressiemodel:

$$(1)\quad \underline{y}_i = \beta_0 + \beta_1 x_{1i} + \ldots\ldots + \beta_k x_{ki} + \underline{v}_i \quad ; \quad i = 1, \ldots\ldots, n \quad ; \quad k < n,$$

waarin $\beta_1, \ldots\ldots, \beta_k$ de onbekende regressiecoëfficiënten zijn, $x_{1i}, \ldots\ldots, x_{ki}$ de waargenomen waarden van de onafhankelijke variabelen $x_1, \ldots\ldots, x_k$ die wij niet-stochastisch onderstellen en $\underline{y}_i$ de stochastische afhankelijke variabelen. De stochastische grootheden $\underline{v}_i$ zijn onbekend.

Opdat de kleinste-kwadraten-schattingen voor $\beta_1, \ldots\ldots, \beta_k$ de beste lineaire zuivere schattingen zijn ("beste" wil zeggen: de kleinste variantie bezitten) en voor het toetsen van hypothesen over de β's de t-toets (STUDENT) en de F-toets (FISHER) gebruikt kunnen worden, moeten de grootheden $\underline{v}_i$ onderling onafhankelijk $N(0,\sigma)$-verdeeld zijn. Met de toets van DURBIN en WATSON wordt nu de onderlinge onafhankelijkheid getoetst tegen het alternatief: kettingcorrelatie zoals deze wordt weergegeven in het model:

$$(2)\quad \underline{v}_i = \rho\, \underline{v}_{i-1} + \underline{u}_i \quad ; \quad i = 2, \ldots\ldots, n,$$

waarin $|\rho| < 1$ is en $\underline{u}_i$ onderling onafhankelijke $N(0,\sigma)$-verdeelde grootheden zijn. De nulhypothese luidt dus: $\rho = 0$; $\rho > 0$ betekent

---

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

2) $N(\mu,\sigma)$: Symbool voor normale verdeling met gemiddelde μ en spreiding σ.

positieve correlatie en $\rho < 0$ negatieve correlatie.

2. Toetsingsgrootheid

Om de toetsingsgrootheid bij deze toets te berekenen moeten eerst de kleinste-kwadraten-schattingen $\underline{b}_0, \ldots\ldots, \underline{b}_k$ voor $\beta_0, \ldots, \beta_k$ bepaald worden. Deze schattingen zijn het eenvoudigst uit te drukken met behulp van de matrix notatie. Hiervoor definiëren wij de gemiddelden:

$$\bar{x}_1 \overset{def}{=} n^{-1} \sum_{i=1}^{n} x_{1i}, \ldots\ldots\ldots, \bar{x}_k \overset{def}{=} n^{-1} \sum_{i=1}^{n} x_{ki} \quad \text{en} \quad \bar{\underline{y}} \overset{def}{=} n^{-1} \sum_{i=1}^{n} \underline{y}_i$$

en de matrices:

$$\underline{y} \overset{def}{=} \begin{pmatrix} \underline{y}_1 \\ \vdots \\ \underline{y}_n \end{pmatrix}, \quad X \overset{def}{=} \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \ldots\ldots & x_{k1} - \bar{x}_k \\ \vdots & \vdots & & \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \ldots\ldots & x_{kn} - \bar{x}_k \end{pmatrix} \quad \text{en} \quad \underline{b} \overset{def}{=} \begin{pmatrix} \underline{b}_1 \\ \vdots \\ \underline{b}_k \end{pmatrix}.$$

Dan vindt men met de methode der kleinste kwadraten:

$$\underline{b}_0 = \bar{\underline{y}} - (\underline{b}_1 \bar{x}_1 + \ldots\ldots + \underline{b}_k \bar{x}_k) \quad \text{en} \quad \underline{b} = (X'X)^{-1} X' \underline{y}$$

waarbij X' uit X ontstaat door verwisseling van rijen en kolommen en $(X'X)^{-1}$ de inverse van $(X'X)$ is.

Noemen wij de residuen $\underline{z}_i$:

$$\underline{z}_i \overset{def}{=} \underline{y}_i - (\underline{b}_0 + \underline{b}_1 x_{1i} + \ldots\ldots + \underline{b}_k x_{ki}) =$$

$$= \underline{y}_i - \bar{\underline{y}} - b_1 (x_{1i} - \bar{x}_1) - \ldots\ldots - \underline{b}_k (x_{ki} - x_k) \tag{3}$$

dan is de toetsingsgrootheid:

$$\underline{d} \overset{def}{=} \left[\sum_{i=1}^{n} \underline{z}_i^2 \right]^{-1} \cdot \sum_{i=2}^{n} (\underline{z}_i - \underline{z}_{i-1})^2. \tag{4}$$

Bestaat er een positieve kettingcorrelatie tussen de $\underline{y}_i$ $(\rho > 0)$ dan zullen de kleinere waarden van $\underline{d}$ waarschijnlijker worden dan onder de nulhypothese; het kritieke gebied bij de eenzijdige toets tegen positieve kettingcorrelatie zal dus uit kleine waarden van $\underline{d}$ bestaan. Bij een negatieve correlatie $(\rho < 0)$ worden de grotere waarden van $\underline{d}$ waarschijnlijker.

Het blijkt dat de verdeling van $\underline{d}$ beperkt is tot het interval $(0 ; 4)$

3. Grenzen voor de kritieke waarden van de toetsingsgrootheid.

DURBIN en WATSON geven in hun artikel voor de éénzijdige toets tegen positieve kettingcorrelatie (dus $\rho > 0$), bij drie waarden van de onbetrouwbaarheid: 0,05; 0,025 en 0,01, onder- en bovengrenzen, resp. d_L en d_u voor de kritieke waarden van $\underline{d}$ voor $k = 1, \ldots, 5$ (zij noemen dit aantal onafhankelijke variabelen k') en voor $n = 15(1)\,40(5)\,100$. Er zijn dus bij een bepaalde onbetrouwbaarheid en afhankelijkheid van de waarde die voor $\underline{d}$ wordt gevonden drie mogelijke resultaten van de toets:

1) $d \leqq d_L$: de nulhypothese $\rho = 0$ wordt verworpen ten gunste van positieve correlatie $\rho > 0$;
2) $d \geqq d_u$: de nulhypothese wordt niet verworpen en
3) $d_L < d < d_u$: de toets eindigt onbeslist.

De eenzijdige toets tegen negatieve correlatie is niets anders dan de toets tegen positieve correlatie toegepast op de waarde $4 - \underline{d}$.

De tweezijdige toets is een combinatie van de twee éénzijdige toetsen. De onbetrouwbaarheid is dan natuurlijk de som van de onbetrouwbaarheden van de eenzijdige toetsen. De verschillende mogelijke resultaten van de eenzijdige toetsen kunnen wij voor de tweezijdige samenvatten tot:

1) d buiten het interval $(d_L, 4 - d_L)$:
de nulhypothese wordt verworpen,
2) d in het interval $(d_u, 4 - d_u)$:
de nulhypothese wordt niet verworpen en
3) alle andere gevallen:
de toets eindigt onbeslist.

4. Nader onderzoek indien de toets onbeslist geëindigd is.

Om in die gevallen waarin de boven beschreven toets onbeslist is geëindigd toch tot een uitspraak over het al dan niet verwerpen van de nulhypothese te komen wordt aan de waarschijnlijkheidsverdeling van $\underline{d}$ een bekende verdeling aangepast. Omdat $\underline{d}$ beprkt is tot het interval $(0;4)$ dus $\underline{d}/4$ tot het interval $(0;1)$ wordt hiervoor een B -verdeling gekozen. Voor deze aanpassing worden de eerste twee momenten van $\underline{d}$ gebruikt die met behulp van sporen van enkele matrices bepaald kunnen worden.

Het spoor van een vierkante matrix is de som van de elementen op zijn hoofddiagonaal.

Met de eerste verschillen

$$\Delta x_{ij} \overset{def}{=} x_{ij} - x_{i,j-1}$$

en de tweede verschillen

$$\Delta^2 x_{ij} \overset{def}{=} \Delta x_{ij} - \Delta x_{i,j-1} = (x_{ij} - x_{i,j-1}) - (x_{i,j-1} - x_{i,j-2})$$

definiëren wij de matrices

$$\Delta X \overset{def}{=} \begin{pmatrix} \Delta x_{12} & \cdots & \Delta x_{k2} \\ \vdots & & \vdots \\ \Delta x_{1n} & \cdots & \Delta x_{kn} \end{pmatrix} \quad \text{en} \quad \Delta^2 X = \begin{pmatrix} \Delta^2 x_{13} & \cdots & \Delta^2 x_{k3} \\ \vdots & & \vdots \\ \Delta^2 x_{1n} & \cdots & \Delta^2 x_{kn} \end{pmatrix}$$

De matrices waarvan wij de sporen nodig hebben zijn dan

$$S_1 = (\Delta X)'(\Delta X)(X'X)^{-1} \quad \text{en}$$
$$S_2 = (\Delta^2 X)'(\Delta^2 X)(X'X)^{-1}.$$

De eerste twee momenten van $\underline{d}$ zijn dan:

(5) $$\mathcal{E}(\underline{d}) = (n-k-1)^{-1}\left[2(n-1) + sp\, S_1\right],$$

(6) $$\sigma^2(\underline{d}) = 2\left[(n-k-1)(n-k+1)\right]^{-1}\left[2(3n-4) - (n-k-1)\left\{\mathcal{E}(\underline{d})\right\}^2 - 2\, sp\, S_2 + sp\, S_1^2\right]$$

Hierin is $sp\, S_1^2$ gelijk aan de som van de kwadraten van alle elementen van S_1.

Onderstellen wij nu dat $\underline{d}/4$ een B-verdeling bezit, dus dat de waarschijnlijkheidsdichtheid gegeven wordt door:

$$\left[B(p,q)\right]^{-1}\left(\tfrac{d}{4}\right)^{p-1}\left(1-\tfrac{d}{4}\right)^{q-1}$$

en

$$\mathcal{E}(\underline{d}) = \frac{4p}{p+q} \quad \text{en} \quad \sigma^2(\underline{d}) = 16\,pq\left[(p+q)^2(p+q+1)\right]^{-1}$$

dan geldt voor p en q :

(7) $$p+q+1 = \mathcal{E}(\underline{d})\left[4 - \mathcal{E}(\underline{d})\right]\left[\sigma^2(\underline{d})\right]^{-1},$$

(8) $$p = \mathcal{E}(\underline{d}) \cdot (p+q)/4.$$

Voor het bepalen van de kritieke waarden van deze verdeling kunnen de tabellen van CATHERINE THOMPSON (1941) gebruikt worden òf tabellen van de F-verdeling (b.v. FISHER and YATES (1949)) voor de grootheid

(9) $$\underline{F} = p(4-\underline{d})/(q\underline{d})$$

met $n_1 = 2q$ en $n_2 = 2p$ vrijheidsgraden (bij de toets tegen positieve correlatie zijn voor F juist de grote waarden kritiek!) òf, indien $2p$ en $2q$ geen gehele getallen zijn de benadering van CARTER (1947) voor de kritieke waarden van $z = \frac{1}{2}\ln F$ (natuurlijke logarthme). Deze laatste benadering is, bij de onbetrouwbaarheden $\alpha = 0{,}05$ en $\alpha = 0{,}01$ (voor de toets tegen positieve correlatie):

$$\xi\sqrt{h+\lambda}/h - \left[\tfrac{1}{2q} - \tfrac{1}{2p}\right]\left[\lambda + \tfrac{5}{6} - \tfrac{5}{3}\right]$$

waarin

$$S = \tfrac{1}{2p} + \tfrac{1}{2q}\ , \quad h = \tfrac{2}{S} \quad \text{en} \quad \lambda = (\xi^2 - 3)/6$$

is en ξ en λ de volgende waarden bezitten:

α	0,05	0,01
ξ	1,6449	2,3263
λ	0,0491	0,4020

Bij de toets tegen negatieve correlatie verloopt de benadering volkomen analoog, mits overal $4-\underline{d}$ in plaats van $\underline{d}$ gebruikt wordt; voor de parameters p en q zullen dan andere waarden gevonden worden (immers $\mathcal{E}(4-\underline{d}) = 4 - \mathcal{E}(\underline{d})$ en $\sigma^2(4-\underline{d}) = \sigma^2(\underline{d})$).

5. Enkele opmerkingen

1. Zijn de onafhankelijke variabelen directe functies van de tijd (vooral monotone functies) dan kan in vele gevallen ook een slechte aanpassing (een slecht gekozen regressiemodel) leiden tot kleinere waarden voor $\underline{d}$ dan men bij een betere aanpassing zou vinden. In deze gevallen verdient het dus aanbeveling eerst een aanpassingstoets (b.v. χ^2) toe te passen.

2. De toets tegen kettingcorrelatie kan ook toegepast worden in variantie-analyse problemen, voor het onderzoeken van een eventuele kettingcorrelatie in de tijdsvolgorde der waarnemingen en bij de aanpassing van orthogonale polynomen. Bij variantie analyse schema's zijn de onafhankelijke variabelen x_j gewoonlijk bekende eenvoudige waarden die direct met de proefopzet

samenhangen. Hier kan men dan $\mathcal{E}(\underline{d})$ en $\sigma^2(\underline{d})$ uitdrukken in de parameters van deze proefopzet (zoals de aantallen groepen per classificatie bij enkel- of tweevoudige variantie analyse).

Bij de aanpassing met orthogonale polynomen kunnen $\mathcal{E}(\underline{d})$ en $\sigma^2(\underline{d})$ worden uitgedrukt in grootheden die met behulp van tabellen van deze orthogonale polynomen te bepalen zijn. DURBIN en WATSON geven zowel voor enkel- en tweevoudige variantie analyse als voor de aanpassing met orthogonale polynomen voorbeelden.

3. De toets van DURBIN en WATSON is slechts voor zeer speciale regressiesystemen (speciale gedaanten van X), die wij hier niet nader zullen bespreken, uniform meest onderscheidend ten opzichte van alle mogelijke zuivere toetsen. Een toets is zuiver indien zijn onderscheidingsvermogen, althans onder alternatieve hypothesen die weinig van de nulhypothese verschillen (hier dus voor waarden van ρ in de buurt van 0), niet kleiner is dan zijn onbetrouwbaarheid onder de nulhypothese.

Literatuur

J. DURBIN and G.S. WATSON (1950, 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.

CATHERINE M. THOMPSON (1941), Tables of percentage points of the incomplete bêta-function.
Biometrika, 32, pp 151-181.

R.A. FISHER and F YATES (1949), Statistical tables for biological agricultural and medical research.
3rd Ed., Oliver & Boyd, London, Table V.

A.H. CARTER (1947), Approximation to percentage points of the -distribution.
Biometrika, 34, pp 352-358.

GERDA KLERK-GROBBEN (1955), Een toets tegen kettingcorrelatie bij toepassing van regressiemethoden bij tijdreeksen. Verslag van een colloquium voordracht over een artikel van J. DURBIN en G.S. WATSON.
Rapport S 174(M 62) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam, 1955.
Verschijnt vermoedelijk in Statistica, 1956.