STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 168

SM 58a

(M 58a, M 59, M 61, M 63, M 64, M 66, M 68, M 69, M 70, M 71)

1955

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

S 168

S 168 (M 58) Enige in de covariantieanalyze gebruikte toetsen

S 168 (M 59) Een toets voor de kleinste van een aantal geschatte varianties van normale verdelingen

R. Doornbos

S 168 (M 61) Een parametervrije toets tegen verloop voor groepen waarnemingen

S 168 (M 63) Toetsen voor één of twee uitbijters bij een normale verdeling

S 168 (M 64) Het aanpassen van een exponentiële regressielijn

H. Kesten
J.Th. Runnenburg

S 168 (M 66) Een toets tegen kettingcorrelatie bij lineaire regressie van J. DURBIN en G;S. WATSON (1950/1951)

S 168 (M 68) Sequente toets voor het gemiddelde van een normale verdeling met gegeven spreiding, of voor het verschil der gemiddelden van twee normale verdelingen, waarvan de spreidingen gelijk zijn

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

S 168 (M 70) Een exacte toets tegen kettingcorrelatie bij lineaire regressieproblemen van E.J. HANNAN (1955)

S 168 (M 71) Kleinste kwadraten-schattingen voor de regressie-coëfficïenten van een meervoudige lineaire regressie-vergelijking

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.
Statistische Afdeling
S 168 (M 58)

Enige in de covariantieanalyse gebruikte toetsen. [1] [2]

Wij gaan uit van k groepen stochastische grootheden

$$\begin{matrix} y_{11}, \ldots, y_{1n_1}, \\ y_{21}, \ldots, y_{2n_2}, \\ \vdots \\ y_{k1}, \ldots, y_{kn_k}. \end{matrix}$$

De verwachtingen van deze grootheden voldoen aan de volgende relaties:

$$(1) \qquad \begin{cases} \mathcal{E}\, y_{1j} = \alpha_1 + \beta_1 x_{1j} & (j = 1, \ldots, n_1), \\ \mathcal{E}\, y_{kj} = \alpha_k + \beta_k x_{kj} & (j = 1, \ldots, n_k). \end{cases}$$

Hierin stellen α_i en β_i $(i = 1, \ldots, k)$ onbekende parameters voor, terwijl de waarden x_{ij} $(i = 1, \ldots, k\,;\ j = 1, \ldots n_i)$ gegeven zijn. Verder wordt ondersteld, dat alle y's onderling onafhankelijk en normaal verdeeld zijn met dezelfde spreiding σ. Hierin is dus begrepen het geval waarin x_{ij} een waarneming is van de stochastische grootheid $\underline{x}_{ij}$, waarbij y_{ij} en $\underline{x}_{ij}$ een tweedimensionale normale verdeling hebben. Onder de voorwaarde dat $\underline{x}_{ij}$ de waarde x_{ij} heeft aangenomen, is y_{ij} dan immers normaal verdeeld met als gemiddelde een lineaire functie van x_{ij}. Wij stellen nu toetsen op voor de volgende hypothesen:

(A) $\qquad \beta_1 = \ldots = \beta_k$

met als toegelaten hypothesen willekeurige α's en β's.

(B) $\qquad \alpha_1 = \ldots = \alpha_k$ en $\beta_1 = \ldots = \beta_k$

met als toegelaten hypothesen willekeurige α's maar gelijke β's en

(C) $\qquad \alpha_1 = \ldots = \alpha_k$ en $\beta_1 = \ldots = \beta_k$

met als toegelaten hypothesen willekeurige α's en β's.

---

1) Dit memorandum dient slechts ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) De eerste van de in dit memorandum besproken toetsen wordt ook in memorandum S 73 (M 37) behandeld met gebruik van een enigszins andere notatie.

Met (A) toetsen wij dus de evenwijdigheid van de k regressielijnen uitgaande van het model (1). Met (B) toetsen wij of de k lijnen samenvallen, aangenomen dat ze evenwijdig zijn en met (C) toetsen wij direct of wij slechts met één regressielijn te maken hebben, uitgaande van (1).

Wij voeren de volgende afkortingen in:

$$C_{xx_i} = \sum_{j=1}^{n_i} \left(x_{ij} - \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}\right)^2 = \sum_{j=1}^{n_i} x_{ij}^2 - \frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i},$$

$$\underline{C}_{yy_i} = \sum_{j=1}^{n_i} \left(y_{ij} - \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}\right)^2 = \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^2}{n_i},$$

$$\underline{C}_{xy_i} = \sum_{j=1}^{n_i} \left(x_{ij} - \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}\right)\left(y_{ij} - \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}\right) =$$

$$= \sum_{j=1}^{n_i} x_{ij}\, y_{ij} - \frac{\sum_{j=1}^{n_i} x_{ij} \sum_{j=1}^{n_i} y_{ij}}{n_i},$$

$$C_{xx_W} = \sum_{i=1}^{k} C_{xx_i},$$

$$\underline{C}_{yy_W} = \sum_{i=1}^{k} \underline{C}_{yy_i},$$

$$\underline{C}_{xy_W} = \sum_{i=1}^{k} \underline{C}_{xy_W},$$

$$C_{xx_T} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}\right)^2}{\sum_{i=1}^{k} n_i},$$

$$\underline{C}_{yy_T} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}\right)^2}{\sum_{i=1}^{k} n_i},$$

$$\underline{C}_{xy_T} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}\, y_{ij} - \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{k} n_i}.$$

De som van de kwadraten van de afwijkingen van de y waarden van de geschatte regressielijnen binnen de afzonderlijke groepen, gesommeerd over de groepen is

$$S_1 = \sum_{i=1}^{k} \left(C_{yyi} - \frac{C_{xyi}^2}{C_{xxi}} \right).$$

Als het model (1) geldt, heeft S_1/σ^2 een χ^2- verdeling met $\sum_{i=1}^{k} n_i - 2k$ vrijheidsgraden.

Wij kunnen ook zo goed mogelijk evenwijdige regressielijnen in de verschillende groepen schatten. De som van de kwadraten van de afwijkingen van deze lijnen is

$$S_2 = \sum_{i=1}^{k} C_{yyi} - \frac{C_{xyw}^2}{C_{xxw}}.$$

Deze som gedeeld door σ^2 heeft een χ^2-verdeling met $\sum_{i=1}^{k} n_i - k - 1$ vrijheidsgraden als hypothese (B) of (C) waar is.

Als hypothese (A) geldt, heeft de som van kwadraten van afwijkingen van één lijn

$$S_3 = C_{yyT} - \frac{C_{xyT}^2}{C_{xxT}},$$

gedeeld door σ^2, een χ^2-verdeling met $\sum_{i=1}^{k} n_i - 2$ vrijheidsgraden.

Voor de hypothese (A) gebruiken wij de toetsingsgrootheid

$$F_A = \frac{\sum_{i=1}^{k} n_i - 2k}{k-1} \cdot \frac{S_2 - S_1}{S_1},$$

die onder de hypothese (A) een F-verdeling bezit met $k-1$ en $\sum_{i=1}^{k} n_i - 2k$ vrijheidsgraden.

Als toetsingsgrootheid voor de hypothese (B) dient

$$F_B = \frac{\sum_{i=1}^{k} n_i - k - 1}{k-1} \cdot \frac{S_3 - S_2}{S_2},$$

welke grootheid een F-verdeling met $k-1$ en $\sum_{i=1}^{k} n_i - k - 1$ vrijheidsgraden heeft, als (B) waar is.

Hypothese (C) toetsen wij tenslotte met

$$F_C = \frac{\sum_{i=1}^{k} n_i - 2k}{2(k-1)} \cdot \frac{S_3 - S_1}{S_1},$$

een F verdeelde grootheid met $2(k-1)$ en $\sum_{i=1}^{k} n_i - 2k$ vrijheidsgraden, onder (C).

Als de betreffende hypothesen niet vervuld zijn, hebben de bijbehorende F's geen F-verdeling, maar zijn grotere F waarden meer waarschijnlijk. Als kritieke zône gebruikt men daarom $F \geqq F_0$, waarin F_0 behorende bij een bepaalde onbetrouwbaarheidsdrempel ε opgezocht kan worden in tabellen van de F-verdeling met het juiste aantal vrijheidsgraden.

Literatuur over de achtergrond van bovengenoemde toetsen kan men vinden in MANN [4] en KENDALL [3]. Het rekenschema is grotendeels ontleend aan DIXON en MASSEY [1]. Al deze bocken bevatten tabellen van de F-verdeling. Deze verdeling is het uitvoerigst getabelleerd in FISHER en YATES [2] .

Literatuur

[1] W.J. DIXON en F.J. MASSEY, Introduction to Statistical Analysis, Mc Graw-Hill Book Company, Inc., New York, Toronto, London, 1951, Chapter 12.

[2] R.A. FISHER en F. YATES, Statistical Tables for Biological, Agricultural and Medical research, 3d ed., Oliver & Boyd, London 1949, Table V.

[3] M.G. KENDALL, The advanced theory of statistics, Vol II, 2nd ed., Griffin, London, 1948, pp 237-246.

[4] H.B. MANN, Analysis and Design of Experiments, Dover Publication, Inc., New York 1949.

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O
Statistische Afdeling

Rapport S 168 (M 59)

door
R. Doornbos

Maart 1955

Een toets voor de kleinste van een aantal geschatte varianties van normale verdelingen [1]

Stel wij hebben k steekproeven, ieder van dezelfde uitgebreidheid n, afkomstig uit normale verdelingen:

$$x_{11}, \ldots, x_{1n}$$
$$\vdots$$
$$x_{k1}, \ldots, x_{kn}.$$

Wij toetsen, evenals bij de toets van Bartlett het geval is de hypothese, dat de varianties van de k normale verdelingen gelijk zijn, hier evenwel speciaal tegen het alternatief, dat de steekproef met de kleinste variantie afkomstig is uit een verdeling die een kleinere variantie heeft dan de andere verdelingen.

De geschatte varianties zijn:

$$S_i^2 = \frac{\sum_{j=1}^{n} (x_{ij} - x_{i.})^2}{n-1} \quad (i = 1, \ldots, k),$$

waarin $x_{i.} = \frac{\sum_{j=1}^{n} x_{ij}}{n}$ het gemiddelde van de ie steekproef is. De kleinste onder de S_i^2 noemen wij S^2_{min}. De toetsingsgrootheid is nu:

$$A = \frac{S^2_{min}}{\sum_{i=1}^{k} S_i^2}.$$

De waarschijnlijkheidsverdeling waaruit deze waarde A een trekking is, is slechts bij benadering numeriek te berekenen. Dit heeft tot gevolg, dat in de hieronder staande tabel van kritieke waarden de bijbehorende onbetrouwbaarheidsdrempel niet exact bekend is. Wel weten wij zeker, dat deze ligt

---

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

tussen 0,05 en 0,04875. Indien de gevonden waarde van A dus kleiner is dan het in de tabel vermelde getal, kunnen wij de hypothese bij de onbetrouwbaarheidsdrempel 0,05 zeker verwerpen. De even waarden van n leveren rekentechnisch veel grotere moeilijkheden op dan de oneven waarden; vandaar dat de tabel wel berekend is voor $n=7$, maar niet voor $n=6$. Om veel nullen achter de komma te vermijden, is 0,0000046 afgekort tot $0,0^{5}46$ enz.

Tabel van kritieke waarden van A bij een onbetrouwbaarheidsdrempel $\approx 0,05$, voor verschillende steekproefuitgebreidheden n en aantallen groepen k.

k \ n	2	3	4	5	7
2	0,00154	0,02500	0,06083	0,09430	0,14663
3	$0,0^{3}278$	0,00837	0,02489	0,04262	0,07331
4	$0,0^{4}964$	0,00418	0,01401	0,02546	0,04647
5	$0,0^{4}444$	0,00251	0,00916	0,01736	0,03306
6	$0,0^{4}241$	0,00167	0,00653	0,01280	0,02518
7	$0,0^{4}145$	0,00119	0,00493	0,00992	0,02008
8	$0,0^{5}94$	$0,0^{3}895$	0,00387	0,00799	0,01654
9	$0,0^{5}65$	$0,0^{3}696$	0,00314	0,00661	0,01395
10	$0,0^{5}46$	$0,0^{3}557$	0,00261	0,00558	0,01200

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
Amsterdam - O.
Statistische Afdeling
Rapport S168 (M 61)

Een parametervrije toets tegen verloop voor groepen waarnemingen.[1]

Wij beschouwen k onderling onafhankelijke stochastische grootheden $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_k$. Van de grootheid $\underline{x}_i$ zijn n_i onderling onafhankelijke waarnemingen $x_{i,1}, \ldots, x_{i,n_i}$ gegeven $(i=1,\ldots,k)$.

De hypothese H_0 die wij willen toetsen luidt, dat $\underline{x}_1, \ldots, \underline{x}_k$ dezelfde waarschijnlijkheidsverdeling bezitten, terwijl de alternatieve hypothesen inhouden, dat de stochastische grootheden $\underline{x}_1, \ldots, \underline{x}_k$ een stijgend of dalend verloop vertonen, gedefinieerd door:

$$(1) \qquad \sum_{i<j} \sum \left\{ P[\underline{x}_i < \underline{x}_j] - P[\underline{x}_i > \underline{x}_j] \right\} \neq 0.$$

De toetsingsgrootheid $\underline{W}$ wordt als volgt gedefiniëerd. Stel voor $i<j$ is $\underline{U}_{i,j}$ het aantal paren waarnemingen $(\underline{x}_{i,\lambda}, \underline{x}_{j,\mu})$ met $\underline{x}_{i,\lambda} < \underline{x}_{j,\mu}$ vermeerderd met de helft van het aantal paren $(\underline{x}_{i,\lambda}, \underline{x}_{j,\mu})$ met $x_{i,\lambda} = x_{j,\mu}$ $(\lambda \leq n_i, \mu \leq n_j)$.[2] Stel verder

$$(2) \qquad \underline{W}_{i,j} \overset{\text{def}}{=} 2\left\{ \underline{U}_{i,j} - \mathcal{E}(\underline{U}_{i,j} \mid H_0) \right\} = 2\underline{U}_{i,j} - n_i n_j ,$$

dan is

$$(3) \qquad \underline{W} \overset{\text{def}}{=} \sum_{i<j} \sum \frac{\underline{W}_{i,j}}{n_i n_j}.$$ [3]

Als onder de $N = \sum_i n_i$ waarnemingen h groepen gelijke waarnemingen optreden en als $\underline{t}_\ell$ het aantal waarnemingen in de ℓ^{e} groep is $(\ell = 1, 2, \ldots, h)$ dan geldt

$$(4) \qquad \mathcal{E}[\underline{W} \mid t_1, t_2, \ldots, t_\ell ; H_0] = 0 ,$$

---

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

2) $\underline{U}_{i,j}$ is dus de toetsingsgrootheid van de toets van WILCOXON, toegepast op de waarnemingen van $\underline{x}_i$ en $\underline{x}_j$ en $\underline{U}_{j,i} = n_i n_j - \underline{U}_{i,j}$

3) De hier beschreven toets is een wijziging van een door T.J. TERPSTRA [4] ontwikkelde toets voor dit probleem. Hij gebruikt de toetsingsgrootheid $\underline{T} = \sum_{i<j} \sum \underline{W}_{i,j}$. Door de toetsingsgrootheid $\underline{W}$ te gebruiken bereikt men, dat de toets bruikbaar is voor een klasse van alternatieve hypothesen, die niet afhangt van de verhoudingen der steekproefgrootten (zie ook [5]).

$$(5)\quad \begin{cases} \sigma^2[\underline{W} \mid t_1, t_2, \ldots, t_h; H_0] = \\ \dfrac{\{N^3 - \Sigma t_\ell^3 - 3(N^2 - \Sigma t_\ell^2)\} \sum \frac{(k+1-2i)^2}{n_i} - \{2(N^3 - \Sigma t_\ell^3) - 3N(N^2 - \Sigma t_\ell^2)\} \sum_{i<j} \sum \frac{1}{n_i n_j}}{3N(N-1)(N-2)}. \end{cases}$$

Treden er geen gelijke waarnemingen op, dan is $t_\ell = 1$ voor iedere ℓ en:

$$(6)\quad \sigma^2[\underline{W} \mid H_0] = \tfrac{1}{3}\left\{ \sum_i \frac{(k+1-2i)^2}{n_i} + \sum_{i<j} \sum \frac{1}{n_i n_j} \right\}.$$

Als H_0 juist is, is de grootheid $\underline{W}$ voor grote waarden van N bij benadering normaal verdeeld met gemiddelde en variantie volgens (4) en (5).
Het is duidelijk, dat $\underline{W}$ in het algemeen grote positieve waarden zal aannemen als er een stijgend en grote negatieve als er een dalend verloop is; de tweezijdige kritieke zône bestaat dus uit grote waarden van $|W|$.

Opmerkingen.

1. Als $k = 2$, dus als wij twee stochastische grootheden $\underline{x}_1$ en $\underline{x}_2$ hebben met resp. n_1 en n_2 waarnemingen, dan geldt:

$$(7)\quad \underline{W} = \frac{2\underline{U} - n_1 n_2}{n_1 n_2},$$

waarin $\underline{U}$ de toetsingsgrootheid van de toets van WILCOXON is, toegepast op de steekproeven van $\underline{x}_1$ en $\underline{x}_2$. Uit (4) en (5) volgt:

$$\mathcal{E}[\underline{W} \mid t_1, t_2, \ldots, t_h; H_0] = 0,$$

$$\sigma^2[\underline{W} \mid t_1, t_2, \ldots, t_h; H_0] = \frac{N^3 - \Sigma t_\ell^3}{3 n_1 n_2 N(N-1)}.$$ (zie memorandum S 47 (M 7))

Als er in dit geval geen gelijke waarnemingen optreden, kan men de exacte overschrijdingskans opzoeken in tabellen van de exacte verdeling van $\underline{U}$ (zie b.v. [1]).

2. Als $n_i = 1$ voor iedere i dan is

$$\underline{W} = \sum_{i<j} \sum \operatorname{sgn}(\underline{x}_j - \underline{x}_i),$$

waarin

$$\operatorname{sgn} z \overset{\text{def}}{=} \begin{cases} 1 & \text{als } z > 0 \\ 0 & \text{als } z = 0 \\ -1 & \text{als } z < 0. \end{cases}$$

Uit (4) en (5) volgt:

$$\mathcal{E}[\,\underline{W} \mid t_1, t_2, \ldots, t_h; H_0] = 0,$$

$$\sigma^2[\,\underline{W} \mid t_1, t_2, \ldots, t_h; H_0] = \frac{2(N^3 - \Sigma t_\ell^3) + 3(N^2 - \Sigma t_\ell^2)}{18}$$

De toets is in dit geval identiek met de rangcorrelatie methode van KEDALL voor het geval, dat één der rijen bestaat uit de getallen $1, 2, \ldots, k$.(zie memorandum S 47 (M 13)). Treden er in dit geval geen gelijke waarnemingen op en is $k \leq 40$ dan kan men de exacte overschrijdingskans opzoeken in de tabellen van de exacte verdeling van de toetsingsgrootheid $\underline{S}$ van KENDALL (zie [2]).
Is $k \leq 10$ en $t_\ell \leq 3$ voor iedere ℓ dan kan men gebruik maken van de tabellen van de exacte verdeling van $\underline{S}$ van SILLITTO [3]

Literatuur.

[1] Van der Vaart, H.R., Gebruiksaanwijzing voor de toets van Wilcoxon, Rapport S 32 (M 4) van de Statistische Afdeling van het Mathematisch Centrum.

[2] Kaarsmaker, L. en A. van Wijngaarden, Tables for use in rankcorrelation, Rapport R 73 van de Rekenafdeling van het Mathematisch Centrum.

[3] Sillitto, G.P., The distribution of Kendall's coefficient of rankcorrelation in rankings containing ties, Biometrika 34 (1947), 36-40.

[4] Terpstra, T.J., The asymptotic normality and consistency of Kendall's test against trend when ties are present in one ranking, Proc.Kon.Ned.Ak. 55 (1952).

[5] Van Eeden, C., A test for the equality of probabilities against a class of specified alternatives, including trend, Rapport S 157(VP 3) van de Statistische Afdeling van het Mathematisch Centrum.

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

Statistische Afdeling
S 168 (M 63)

Toetsen voor één of twee uitbijters bij een normale verdeling.[1)]

Stel wij hebben een steekproef van waarnemingen, gerangschikt naar opklimmende grootte:

$$x_1 \leq x_2 \leq \cdots \leq x_n .$$

Wij willen de hypothese toetsen dat deze steekproef afkomstig is uit één normale verdeling, tegen het alternatief dat de grootste waarneming x_n uit een andere verdeling komt dan de andere waarnemingen. Wij gebruiken hiervoor de toetsingsgrootheid

$$T_n = \frac{x_n - \bar{x}}{s} ,$$

waarin $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, terwijl $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$. Deze toetsingsgrootheid is het eerst voorgesteld door E.S. PEARSON en C. CHANDRA SEKAR (1936), terwijl de exacte verdeling van T_n is afgeleid door F.E. GRUBBS (1950). Om te toetsen of de kleinste waarneming een uitbijter is, gebruikt men als toetsingsgrootheid

$$T_1 = \frac{\bar{x} - x_1}{s} .$$

In tabel 1 vinden wij kritieke waarden van T_n en T_1 getabelleerd voor verschillende onbetrouwbaarheidsdrempels en meerdere waarden van n. Deze tabel is een deel van tabel IA van GRUBBS.

---

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

Tabel 1

Kritieke waarden van T_n en T_1, bij een onbetrouwbaarheidsdrempel α en omvang van de steekproef n.

n \ α	0,01	0,025	0,05	0,10
3	1,414	1,414	1,412	1,406
4	1,723	1,710	1,689	1,645
5	1,955	1,917	1,869	1,791
6	2,130	2,067	1,996	1,894
7	2,265	2,182	2,093	1,974
8	2,374	2,273	2,172	2,041
9	2,464	2,349	2,237	2,097
10	2,540	2,414	2,294	2,146
12	2,663	2,519	2,387	2,229
14	2,759	2,602	2,461	2,297
16	2,837	2,670	2,523	2,354
18	2,903	2,728	2,577	2,404
20	2,959	2,778	2,623	2,447
25	3,071	2,880	2,717	2,537

Men kan één- of tweezijdig toetsen, al naar gelang van te voren al dan niet bekend is naar welke kant uitbijters eventueel voor kunnen komen. Bij tweezijdige toetsing moeten de onbetrouwbaarheidsdrempels van de tabel met 2 worden vermenigvuldigd.

Om te toetsen of de twee grootste waarnemingen te groot zijn, nemen wij de toetsingsgrootheid

$$\frac{S^2_{n-1,n}}{S^2} = \frac{\sum_{i=1}^{n-2} (x_i - \bar{x}_{n-1,n})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

waarin $\bar{x}_{n-1,n} = \frac{1}{n-2} \sum_{i=1}^{n-2} x_i$.

Een analoge grootheid

$$\frac{S^2_{1,2}}{S^2} = \frac{\sum_{i=3}^{n} (x_i - \bar{x}_{1,2})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

waarbij $\bar{x}_{1,2} = \frac{1}{n-2} \sum_{i=3}^{n} x_i$ kan worden gebruikt om de twee kleinste waarnemingen te toetsen Deze toetsen zijn voorgesteld door F.E. GRUBBS, die ook hiervan een tabel geeft. Enige kritieke waarden zijn in tabel 2 verzameld.

Tabel 2

Kritieke waarden van $\frac{S^2_{n-1,n}}{S^2}$ en $\frac{S^2_{1,2}}{S^2}$, bij een onbetrouwbaarheidsdrempel α en omvang van de steekproef n.

n \ α	0,01	0,025	0,05	0,10
4	0,0000	0,0002	0,0008	0,0031
5	0,0035	0,0090	0,0183	0,0376
6	0,0186	0,0349	0,0565	0,0921
7	0,0440	0,0708	0,1020	0,1479
8	0,0750	0,1101	0,1478	0,1994
9	0,1082	0,1492	0,1909	0,2454
10	0,1415	0,1865	0,2305	0,2863
12	0,2044	0,2536	0,2996	0,3552
14	0,2605	0,3112	0,3568	0,4106
16	0,3098	0,3603	0,4048	0,4562
18	0,3530	0,4025	0,4455	0,4944
20	0,3909	0,4391	0,4804	0,5269

Hier geldt weer dezelfde opmerking over één- en tweezijdige toetsing als bij tabel 1 Tabel 1 geeft rechtseenzijdige kritieke waarden, tabel 2 linkseenzijdige. Wij verwerpen de getoetste hypothese dus in het eerste geval als T_n of T_1 groter zijn dan de overeenkomstige waarde uit de tabel, terwijl in het geval wij twee uitbijters toetsen juist lage waarden van de toetsingsgrootheid worden verworpen

Literatuur:

F.E. GRUBBS (1950), Sample criteria for testing outlying observations, Ann of Math.Stat.21(1950),pp 27-58.

E.S. PEARSON and C.CHANDRA SEKAR (1936), The efficiency of statistical tools and a criterion for the rejection of outlying observations, Biometrika,28(1936),pp 308-320.

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

Rapport S 168(M 64)

door H. Kesten en
J. Th. Runnenburg.

Het aanpassen van een exponentiële regressielijn.[1)]

1. Het probleem.

Laat ξ en η twee variabelen zijn, die voldoen aan de betrekking

$$\eta = \theta + \alpha\, 10^{-\beta\xi} \qquad (1)$$

waarin θ, α en β onbekende constanten zijn. Laat verder een aantal (N) aequidistante waarden van ξ gegeven zijn en wel

$$x_i = c + i\alpha \qquad (c \text{ en } \alpha \text{ bekend};\ i = 1, \ldots, N) \qquad (2)$$

Bij iedere x_i behoort dan, volgens (1), een waarde η_i van η :

$$\eta_i = \theta + \alpha\, 10^{-\beta x_i} \qquad (i = 1, \ldots, N). \qquad (3)$$

Deze η_i worden nu waargenomen, waarbij echter toevallige afwijkingen optreden. Noemen wij de waarnemingen y_i , dan geldt dus

$$y_i = \eta_i + \underline{u}_i \qquad (i = 1, \ldots, N)\ .^{2)} \qquad (4)$$

Over de waarschijnlijkheidsverdeling der $\underline{u}_i$ is vaak wel iets bekend. Meestal wordt ondersteld, dat de $\underline{u}_i$ onderling onafhankelijk verdeeld zijn met verwachting 0. Soms ook dat zij gelijke spreidingen bezitten en normaal verdeeld zijn.

Het probleem is nu, een methode aan te geven, om de onbekende parameters θ, α en β te schatten op grond van de waarnemingen $(x_1, y_1), \ldots, (x_n, y_n)$. In dit memorandum wordt een eenvoudige methode daartoe beschreven. De eigenschappen van de verkregen schattingen, die afhangen van de over de $\underline{u}_i$ gemaakte onderstellingen, worden hier niet besproken.

---

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

2) Onderstreping van een letter duidt aan, dat de grootheid stochastisch is; dezelfde letter zonder onderstreping wordt gebruikt voor een door de stochastische grootheid aangenomen waarde.

2. <u>Schatting van β</u>.

Laat N een even getal zijn

$$N = 2n. \tag{5}$$

(Is N oneven, dan wordt het laatste punt niet gebruikt voor de schatting van β).

Beschouw nu, voor $k \leqq n$, het verschil

$$\eta_k - \eta_{n+k} = \alpha \left[10^{-\beta(c+k\alpha)} - 10^{-\beta\{c+(n+k)\alpha\}} \right] = \alpha \cdot 10^{-\beta(c+k\alpha)} \left\{ 1 - 10^{-\beta n\alpha} \right\},$$

en definieer:

$$\zeta_k \overset{\text{def}}{=\!=} {}^{10}\log(\eta_k - \eta_{n+k}) = {}^{10}\log\alpha - \beta(c+k\alpha) + {}^{10}\log\left\{1 - 10^{-\beta n\alpha}\right\} =$$

$$= C - \beta(c+k\alpha) = C - \beta x_k , \tag{6}$$

met $$C = {}^{10}\log\alpha + {}^{10}\log\left\{1 - 10^{-\beta n\alpha}\right\}.$$

Stel verder

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \left(\sum_{k=1}^{n}(x_k - \bar{x}) = 0\right) \tag{7}$$

en vorm:

$$\sum_{k=1}^{n}(x_k - \bar{x})\zeta_k = C\sum_{k=1}^{n}(x_k - \bar{x}) - \beta\sum_{k=1}^{n}(x_k - \bar{x})x_k =$$

$$= -\beta\sum_{k=1}^{n}(x_k - \bar{x})^2 . \tag{8}$$

Dan is dus

$$\beta = -\frac{\sum_{k=1}^{n}(x_k - \bar{x})\zeta_k}{\sum_{k=1}^{n}(x_k - \bar{x})^2} . \tag{9}$$

Op grond hiervan ligt het voor de hand, als schatting voor β te nemen:

$$b \overset{\text{def}}{=\!=} -\frac{\sum_{k=1}^{n}(x_k - \bar{x})z_k}{\sum_{k=1}^{n}(x_k - \bar{x})^2} = -\frac{6\sum_{k=1}^{n}\{2k - (n+1)\}z_k}{n(n^2-1)d}, \tag{10}$$

met

$$z_k \overset{\text{def}}{=\!=} \log(y_k - y_{n+k}). \tag{11}$$

3. Schattingsmethode voor θ en α.

Na bepaling van b kunnen wij θ en α met behulp van de methode der kleinste kwadraten schatten. Wij zoeken daartoe de waarden t en a van θ en α, die

$$(12) \qquad \sum_{i=1}^{N} \left(y_i - \theta - \alpha \cdot 10^{-b x_i} \right)^2$$

minimaal maken, waarbij, ook als N oneven is, alle waarnemingen gebruikt kunnen worden. Derhalve moet gelden:

$$\sum_{i=1}^{N} y_i - Nt - a \sum_{i=1}^{N} 10^{-b x_i} = 0 ,$$

en

$$\sum_{i=1}^{N} y_i \, 10^{-b x_i} - t \sum_{i=1}^{N} 10^{-b x_i} - a \sum_{i=1}^{N} 10^{-2 b x_i} = 0 .$$

Oplossing van deze vergelijkingen geeft als schattingen t en a:

$$t = \frac{\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} 10^{-2 b x_i}\right) - \left(\sum_{i=1}^{N} y_i \, 10^{-b x_i}\right)\left(\sum_{i=1}^{N} 10^{-b x_i}\right)}{\left(N \sum_{i=1}^{N} 10^{-2 b x_i}\right) - \left(\sum_{i=1}^{N} 10^{-b x_i}\right)^2}$$

of

$$(13) \qquad \boxed{t = \frac{\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} 10^{4-2 b x_i}\right) - \left(\sum_{i=1}^{N} y_i \, 10^{2-b x_i}\right)\left(\sum_{i=1}^{N} 10^{2-b x_i}\right)}{\left(N \sum_{i=1}^{N} 10^{4-2 b x_i}\right) - \left(\sum_{i=1}^{N} 10^{2-b x_i}\right)^2}}$$

en

$$a = \frac{\left(N \sum_{i=1}^{N} y_i \, 10^{-b x_i}\right) - \left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} 10^{-b x_i}\right)}{\left(N \sum_{i=1}^{N} 10^{-2 b x_i}\right) - \left(\sum_{i=1}^{N} 10^{-b x_i}\right)^2}$$

of

$$(14) \qquad \boxed{a = 10^2 \cdot \frac{\left(N \sum_{i=1}^{N} y_i \, 10^{2-b x_i}\right) - \left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} 10^{2-b x_i}\right)}{\left(N \sum_{i=1}^{N} 10^{4-2 b x_i}\right) - \left(\sum_{i=1}^{N} 10^{2-b x_i}\right)^2}} .$$

De teller en noemer zijn, om (13) en (14) te verkrijgen, met 10^4 vermenigvuldigd in de formules voor t en a om te voorkomen dat negatieve machten van 10 (i.c. $10^{-b x_i}$) opgezocht moeten worden. De exponent 4 is aangepast aan ons getallenvoorbeeld. In andere gevallen zal misschien een hogere exponent gewenst zijn.

4. Opmerkingen.

De betrekking $\eta = \theta + \alpha . 10^{-\beta \xi}$ is aequivalent met

(15) $$\eta = \theta + \alpha . e^{-\beta . {}^{e}\log 10 . \xi} = \theta + \alpha . e^{-2,3026 \beta \xi} .$$

In dit memorandum is alleen de eerste uitdrukking beschouwd omdat tabellen van ${}^{10}\log x$ beter beschikbaar zijn dan tabellen van ${}^{e}\log x$. Het is verder aan te bevelen de x-schaal zo te kiezen dat

(16) $x_1 = 0$ $(c = -d)$ en $x_i = (i-1)\, d$, eventueel met $d = 1$.

In dat geval is α precies het verschil tussen de eerste waarde van η en de asymptotische waarde ($\xi \longrightarrow \infty$) van η , en a is hiervan de schatting.

De aangepaste waarden (of: regressiewaarden) y_i worden gegeven door

(17) $$y_i = t + a . 10^{-b x_i} \qquad (i = 1, \ldots, N) .$$

5. Voorbeeld.

Waarnemingen: (met $c = -1$, $d = 1$ $N = 11$)

i	x_i	y_i
1	0	46,5
2	1	42,5
3	2	40,0
4	3	38,0
5	4	36,5
6	5	36,0
7	6	35,5
8	7	34,5
9	8	33,5
10	9	32,5
11	10	32,5

Berekening van b.

z_k | $2k-(n+1)$ | $n = 5$, $n(n^2-1) = 120$

$z_1 = {}^{10}\log\,(46{,}5 - 36{,}0) = 1{,}021$ | -4

$z_2 = {}^{10}\log\,(42{,}5 - 35{,}5) = 0{,}845$ | -2

$z_3 = {}^{10}\log\,(40{,}0 - 34{,}5) = 0{,}740$ | 0

$z_4 = {}^{10}\log\,(38{,}0 - 33{,}5) = 0{,}653$ | 2

$z_5 = {}^{10}\log\,(36{,}5 - 32{,}5) = 0{,}602$ | 4

$$\sum_{k=1}^{5} z_k\,[2k-(n+1)] = -2{,}060,$$

$$b = +\frac{6.2{,}060}{120} = 0{,}103.$$

Indien z_k tot in r decimalen nauwkeurig bepaald wordt, is de maximale fout in b tengevolge van afrondingen bij de berekening $\approx \frac{3}{2n\alpha} \cdot 10^{-3}$. De stochastische fout $(= b-\beta)$ kan echter groter zijn.

Berekening van t en a.

x_i	$2 - b x_i$	10^{2-bx_i}	y_i	$Y_i = t + a.10^{-bx_i}$
0	2,000	100,0	46,5	45,9
1	1,897	78,9	42,5	42,8
2	1,794	62,2	40,0	40,4
3	1,691	49,1	38,0	38,5
4	1,588	38,7	36,5	36,9
5	1,485	30,6	36,0	35,7
6	1,382	24,1	35,5	34,8
7	1,279	19,0	34,5	34,0
8	1,176	15,0	33,5	33,4
9	1,073	11,8	32,5	33,0
10	0,970	9,3	32,5	32,6

$$\sum_{i=1}^{N} 10^{2-bx_i} = 438.7\,,$$

$$\sum_{i=1}^{N} 10^{4-2bx_i} = 26331{,}45\,,$$

$$\sum_{i=1}^{N} y_i = 408{,}0\,,$$

$$\sum_{i=1}^{N} y_i\, e^{2-bx_i} = 17570{,}5\,,$$

$$t = \frac{408{,}0 \times 17570{,}5 - 17570{,}5 \times 438{,}7}{11 \times 26331{,}45 - 438{,}7 \times 438{,}7} = 31{,}23,$$

$$a = \frac{11 \times 17570{,}5 - 408{,}0 \times 438{,}7}{11 \times 26331{,}45 - 438{,}7 \times 438{,}7} = 14{,}70 .$$

Indien θ, α en β ongeveer de waarden hebben die hier gevonden zijn, en de berekening gemaakt wordt zoals hier is aangegeven, zal de fout in t en a, die door afronding veroorzaakt wordt, $\leqq 0{,}05$ zijn. De fout in de aanpassing is dan $\leqq 0{,}1$.

Literatuur, waarin andere methoden behandeld worden:

D.J. Cowden, Simplified methods of fitting certain types of growth curves. Journ.Am.Stat.Ass. 42 (1947) 585-590.

E.S. Keeping, A significance test for exponential regression Ann.Math.Stat. 22 (1951) 180-198.

D.S. Villars, A significance test and estimation in the case of exponential regression. Ann.Math. Stat. 18 (1947) 596-600.

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

S 168 (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 $\rho < o$ negatieve correlatie.

2. Toetsingsgrootheid

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

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

en de matrices:

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

Dan vindt men met de methode der kleinste kwadraten:

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

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 $\underline{z}_i$:

$$\begin{aligned} \underline{z}_i &\overset{def}{=} \underline{y}_i - (\underline{b}_o + \underline{b}_1 x_{1i} + \ldots\ldots + \underline{b}_k x_{ki}) = \\ &= \underline{y}_i - \bar{\underline{y}} - b_1 (x_{1i} - \bar{x}_1) - \ldots\ldots - \underline{b}_k (x_{ki} - \bar{x}_k) \end{aligned} \tag{3}$$

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 $\underline{y}_i$ $(\rho > o)$ 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 $(\rho < o)$ 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 nulhypothe $\rho = 0$ 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

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

De matrices waarvan wij de sporen nodig hebben zijn dan

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

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

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

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

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

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

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

dan geldt voor p en q :

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

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

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

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

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):

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

waarin

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

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 wa rden 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.

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

Statistische Afdeling
S 168 (M 68)

Sequente toets voor het gemiddelde van een normale verdeling met gegeven spreiding, of voor het verschil der gemiddelden van twee normale verdelingen, waarvan de spreidingen gelijk zijn.[1]

Wij beschouwen een stochastische variabele $\underline{x}$ [2], die een normale verdeling bezit met gemiddelde μ en spreiding σ. Men wil voor gegeven waarde μ_0 de hypothese $\mu \leqq \mu_0$ toetsen tegen $\mu > \mu_0$. Bij de hier te behandelen sequente methode verricht men waarnemingen $x_1, x_2, x_3, \cdots$ van $\underline{x}$. Na iedere waarneming wordt nagegaan of de som van de gevonden waarden van $\underline{x}$ tussen twee nader aan te geven grenzen ligt of niet. Zodra de som één van deze grenzen overschrijdt, is het experiment beëindigd en kan de hypothese $\mu \leqq \mu_0$ worden aanvaard of verworpen.

De hier te behandelen methode is ook van toepassing op het volgende geval: de stochastische variabelen $\underline{u}$ en $\underline{v}$ hebben normale verdelingen met gemiddelden $\mathcal{E}\underline{u} = \mu'$ en $\mathcal{E}\underline{v} = \mu''$ en dezelfde gegeven spreiding σ'; men wenst voor een gegeven waarde μ_0 de hypothese $\mu' - \mu'' \leqq \mu_0$ te toetsen tegen de hypothese $\mu' - \mu'' > \mu_0$. Men verricht nu paren waarnemingen: $(u_1, v_1), (u_2, v_2)$, enz. van $\underline{u}$ en $\underline{v}$. Indien men nu $\underline{x} = \underline{u} - \underline{v}$, $x_i = u_i - v_i$, $\mu = \mu' - \mu''$ en $\sigma = \sigma'\sqrt{2}$ stelt, is dit geval herleid tot het voorafgaande.

Het is niet mogelijk om een sequente toets te construeren, waarbij men slechts een kleine kans heeft om $\mu \leqq \mu_0$ te verwerpen voor iedere waarde van μ die iets kleiner is dan μ_0, en tevens een kleine kans om $\mu \leqq \mu_0$ te aanvaarden voor iedere waarde van μ die iets groter is dan μ_0. Wel kan men bereiken, dat behoudens een kleine kans, hoogstens α, $\mu \leqq \mu_0$ aanvaard wordt als $\mu \leqq \mu_1 < \mu_0$ is en behoudens een kleine kans, hoogstens β, $\mu \leqq \mu_0$ verworpen wordt als $\mu \geqq \mu_2 > \mu_0$ is; de waarden van μ_1, μ_2, α en β moeten door de onderzoeker vooraf vastgesteld worden.

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.
2) Stochastische grootheden worden door onderstreepte letters aangeduid.

De toets kan dan als volgt uitgevoerd worden.

Bereken:
$$a = \frac{\mu_1 + \mu_2}{2}$$
$$b_1 = \frac{\sigma^2}{\mu_2 - \mu_1} \ln \frac{\beta}{1-\alpha}$$
$$b_2 = \frac{\sigma^2}{\mu_2 - \mu_1} \ln \frac{1-\beta}{\alpha}$$

Teken in een rechthoekig coördinatenstelsel met de assen y en z de evenwijdige rechten :

$$L_1 : \quad y = az + b_1$$
$$L_2 : \quad y = az + b_2$$

Men zet nu hierin de waarneming x_1 uit in het punt: $z = 1$, $y = x_1$

Ligt dit punt onder of op L_1, dan aanvaardt men $\mu \leqq \mu_0$; ligt het boven of op L_2 , dan verwerpt men $\mu \leqq \mu_0$; ligt het tussen de beide lijnen, dan verricht men een tweede waarneming x_2 en bepaalt dan het punt: $z = 2$, $y = x_1 + x_2$

Ligt dit punt onder of op L_1 , dan aanvaardt men $\mu \leqq \mu_0$; ligt het boven of op L_2 , dan verwerpt men $\mu \leqq \mu_0$; ligt het tussen beide lijnen dan verricht men een derde waarneming enz.

De toets eindigt bij de n^{de} waarneming, als $z = n$, $y = x_1 + x_2 + \cdots + x_n$ het eerst punt is het punt, dat <u>niet</u> tussen de lijnen L_1 en L_2 ligt. Men kan bewijzen, dat dit steeds na een eindig aantal waarnemingen het geval zal zijn.

<u>Literatuur</u>:

A. Wald, Sequential Analysis (1947) chapter 7.

Statistical Research Group, Columbia University (1945), Sequential Analysis of statistical data, Section 4. (Hierin vindt men een aantal voor de toepassing nuttige tabellen).

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

S 168 (M 69)

Een exacte toets tegen kettingcorrelatie van M.OGAWARA (1951) [1]

We gaan uit van de veronderstelling dat de grootheden $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_{2n+1}$ een simultane normale verdeling bezitten met $\mathcal{E}\{\underline{y}_i\} = \mu$ en $\sigma^2\{\underline{y}_i\} = \sigma^2$ en dat er tussen de grootheden $\underline{y}_i$ een correlatie kan bestaan van de vorm

$$(1) \qquad (\underline{y}_t - \mu) = \rho(\underline{y}_{t-1} - \mu) + \underline{v}_t \ ,$$

waarin $\underline{v}_t$ onderling onafhankelijk $N(0, \sigma^2(1-\rho^2))$ verdeeld zijn en $|\rho| < 1$. Een correlatie van deze gedaante noemt men eerste orde kettingcorrelatie (Engels: first order serial correlation) Kettingcorrelatie kan men bv. verwachten bij tijdreeksen. Bovenstaand model wordt dan vaak gebruikt indien men kan veronderstellen dat de correlatie tussen twee opeenvolgende waarnemingen belangrijk sterker is dan tussen verder van elkaar gelegen waarnemingen.

M.OGAWARA (1951) leidt nu een exacte toets voor de hypothese $\rho = 0$ in bovenstaand model af, door de voorwaardelijke simultane verdeling van $\underline{y}_2, \underline{y}_4, \ldots, \underline{y}_{2n}$ te beschouwen onder voorwaarde van de gevonden waarden $y_1, y_3, \ldots, y_{2n+1}$ van de variabelen met oneven index. Onder deze voorwaarde zijn de grootheden $\underline{y}_{2s}$ $(s=1,\ldots,n)$ namelijk onderling onafhankelijk normaal verdeeld, onafhankelijk van de waarde van ρ.

Gemiddelde en variantie van $\underline{y}_{2s}$ kunnen wij als volgt afleiden:

$$(\underline{y}_{2s} - \mu) = \rho(y_{2s-1} - \mu) + \underline{v}_{2s}$$

en

$$(y_{2s+1} - \mu) = \rho(\underline{y}_{2s} - \mu) + \underline{v}_{2s+1} \ ,$$

of

$$\rho^2(\underline{y}_{2s} - \mu) = \rho(y_{2s+1} - \mu) - \rho\,\underline{v}_{2s+1} \ .$$

Tellen we de linker- en rechterleden van de eerste en derde gelijkheid op dan ontstaat:

$$(1+\rho^2)(\underline{y}_{2s} - \mu) = \rho(y_{2s-1} + y_{2s+1}) - 2\rho\mu + \underline{v}_{2s} - \rho\,\underline{v}_{2s+1} \ ,$$

dus

$$(2) \qquad \underline{y}_{2s} = \mu\{1 - 2\rho(1+\rho^2)^{-1}\} + \rho(1+\rho^2)^{-1}(y_{2s-1} + y_{2s+1}) + (1+\rho^2)^{-1}(\underline{v}_{2s} - \rho\,\underline{v}_{2s+1}) \ .$$

Hieruit volgt

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

---

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

en

$$\sigma^2\{\underline{y}_{2s} \mid y_1, y_3, \ldots, y_{2n+1}\} = (1+\rho^2)^{-2}\left[\sigma^2\{\underline{v}_{2s}\} + \rho^2\sigma^2\{\underline{v}_{2s+1}\}\right] = (1-\rho^2)(1+\rho^2)^{-1}\sigma^2.$$

Vergelijking (2) kunnen we schrijven als:

(3) $$\underline{y}_{2s} = \gamma_0 + \gamma_1 x_s + \underline{v}'_{2s} \qquad (s = 1, \ldots, n),$$

waarin dus:

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

is,

(5) $$\gamma_1 = 2\rho(1+\rho^2)^{-1}$$

en

(6) $$x_s = 2^{-1}(y_{2s-1} + y_{2s+1}),$$

terwijl de stochastische grootheden $\underline{v}'_{2s}$ (s = 1,... ,n) onderling onafhankelijk $N(0, \sigma^2(1-\rho^2)(1+\rho^2)^{-1})$ verdeeld zijn (we beschouwen y_{2s-1} en y_{2s+1} en dus ook z_s als gegeven en dus als niet-stochastische grootheden).

In de vorm (3) herkennen we het lieaire regressiemodel, waarbij z_s, het gemiddelde van de voorafgaande en volgende waarde van $\underline{y}_{2s}$ (dus $2^{-1}(y_{2s-1} + y_{2s+1})$), als onafhankelijke variabele fungeert.

Daar aan alle voorwaarden voldaan is (onderling onafhankelijk normaal verdeelde grootheden v'_{2s} met gelijke varianties) kunnen we in dit regressiesysteem met de methode der kleinste kwadraten meest aannemelijke schattingen voor de onbekenden γ_0 en γ_1 vinden:

(7) $$\hat{\gamma}_0 = \bar{\underline{y}} - \hat{\gamma}_1 \bar{x},$$

(8) $$\hat{\gamma}_1 = \sum_1^n (\underline{y}_{2s} - \bar{\underline{y}})(x_s - \bar{x}) \Big/ \sum_1^n (x_s - \bar{x})^2,$$

waarin

$$\bar{\underline{y}} = \sum_1^n \underline{y}_{2s}/n \qquad \text{en} \qquad \bar{x} = \sum_1^n x_s/n.$$

Ook geeft de regressieanalyse (variantieanalyse) exacte toetsen voor lineaire hypothesen over γ_0 en γ_1. In het bijzonder interesseert ons de hypothese $\rho = 0$; in model (3) betekent dit $\gamma_1 = 0$. Voor het toetsen van deze hypothese gebruiken we dan als toetsingsgrootheid:

(9) $$\underline{F}_{1,n-2} = (n-2)\hat{\gamma}_1^2 \Sigma(x_s - \bar{x})^2 \Big/ \Sigma(\underline{y}_{2s} - \hat{\gamma}_0 - \hat{\gamma}_1 x_s)^2 = (n-2)\hat{\gamma}_1^2 \Sigma(x_s - \bar{x})^2 \Big/ \left[\Sigma(\underline{y}_{2s} - \bar{\underline{y}})^2 - \hat{\gamma}_1^2 \Sigma(x_s - \bar{x})^2\right],$$

die onder de hypothese $\gamma_1 = 0$ een F-verdeling (Fisher) met 1 en n - 2 vrijheidsgraden bezit, of

(10) $$\underline{t}_{n-2} = \sqrt{\underline{F}_{1,n-2}} = \hat{\gamma}_1 \Big/ \sqrt{\left[\Sigma(\underline{y}_{2s} - \bar{\underline{y}})^2 - \hat{\gamma}_1^2 \Sigma(x_s - \bar{x})^2\right] \Big/ \left[(n-2)\Sigma(x_s - \bar{x})^2\right]},$$

die onder de hypothese $\gamma_1 = 0$ een t-verdeling (Student) met (n-2) vrijheidsgraden bezit (zie stelling 4.1 van H.B. Mann (1949) of memorandum S 73 (M 34) met $n_1 = n_2 = \ldots . n_h = 1$).

De toets van OGAWARA kan tot hogere orde kettingcorrelatie worden gegeneraliseerd (zie M.OGAWARA (1951) en ook E.J.HANNAN (1955)). Een andere uitbreiding gaf E.J.HANNAN voor lineaire regressieproblemen (zie E.J.HANNAN (1955) en memorandum S 168 (M 70)).

Literatuur

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

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

H.B MANN (1949) Analysis and design of experiments, hoofdstuk 4.
New York, Dover Publications, Inc.

S 73 (M 34) Toets voor de hypothese dat een regressiecoefficiënt nul is, wanneer de afwijkingen van de regressielijn normaal verdeeld zijn met gelijke spreidingen.
Memorandum van de Statistische Afdeling van het Mathematisch Centrum te Amsterdam (19..)

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

S 168 (M 70)

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

1. Model

We gaan uit van een lineair regressiesysteem:

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

waarin $\beta_0, \beta_1, \ldots, \beta_k$ onbekende regressie-coëfficiënten zijn en $(x_{j1}, \ldots, x_{j,2n+1})$ $(j = 1, \ldots, k)$ 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 <u>kettingcorrelatie</u> bestaan, d.w.z.:

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

waarin $\underline{u}_1, \ldots, \underline{u}_{2n+1}$ onderling onafhankelijk $N(0, \sigma^2_{\underline{u}})$ 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 $\underline{y}_2, \underline{y}_4, \ldots, \underline{y}_{2n}$ beschouwen onder voorwaarde van de gevonden waarden van $y_1, y_3, \ldots, y_{2n+1}$. Onder deze voorwaarde zijn namelijk de grootheden $\underline{y}_{2s}$ $(s = 1, \ldots, n)$ onderling onafhankelijk normaal verdeeld. Voor het gemiddelde en de variantie van $\underline{y}_{2s}$ kan men vrij eenvoudig afleiden:

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

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

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:

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

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

In het regressiesysteem (5) zijn nu de grootheden $\underline{v}'_{2s}$ wel onderling onafhankelijk normaal verdeeld $(N(0, \sigma^2_{\underline{u}}(1+\rho^2)^{-1})$. 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):

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

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

$$\underline{\hat{\gamma}} = (z'z)^{-1}z'\,\underline{y}$$

als

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

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 $\{\ell_{ij}\}$ opgebouwd uit elementen

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

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

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

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 $x_{j,1},\ldots,x_{j,2n+1}$, 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:

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

$(s=1,\ldots,n)$. 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 $\underline{y}_{2s}$; 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) geinverteerd 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 geinverteerd moet worden. Bovendien zal men indien de hypothese $\rho = 0$ <u>niet</u> 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.

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

S 168 (M 71)

Kleinste kwadraten-schattingen voor de regressie-coëfficiënten van een meervoudige lineaire regressie-vergelijking [1].

1. Inleiding.

We onderstellen dat n waarnemingen $y_1, \ldots, y_n$ [2] verricht wordenwelke aan de volgende betrekkingen voldoen:

$$\begin{aligned} \underline{y}_1 &= \alpha + \beta_1 x_{11} + \cdots + \beta_k x_{k1} + \underline{v}_1, \\ \underline{y}_2 &= \alpha + \beta_1 x_{12} + \cdots + \beta_k x_{k2} + \underline{v}_2, \\ &\vdots \\ \underline{y}_n &= \alpha + \beta_1 x_{1n} + \cdots + \beta_k x_{kn} + \underline{v}_n, \end{aligned} \tag{1,1}$$

waarin $\alpha, \beta_1, \ldots, \beta_k$ onbekende regressie-coëfficiënten zijn, x_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, n)$ bekende waarden van de onafhankelijke variabelen en $\underline{v}_j$ $(j = 1, \ldots, n)$ niet waarneembare onafhankelijke stochastische grootheden met verwachting $\mathcal{E}\,\underline{v}_j = 0$ en gelijke (onbekende) spreidingen.

De kleinste kwadratenschattingen voor de coëfficiënten $\alpha, \beta_1, \ldots, \beta_k$ zijn nu die waarden a, $b_1, \ldots, b_k$ welke voor $\alpha, \beta_1, \ldots, \beta_k$ gesubstitueerd de volgende kwadratische vorm minimaliseren:

$$\mathcal{S} = \sum_{j=1}^{n} (y_j - \alpha - \beta_1 x_{1j} - \cdots - \beta_k x_{kj})^2. \tag{1,2}$$

Zij voldoen dus aan de relaties:

$$\left\{\frac{\partial \mathcal{S}}{\partial \alpha}\right\}_{\substack{\alpha = a \\ \beta = b}} = 0 \tag{1,3}$$

en

$$\left\{\frac{\partial \mathcal{S}}{\partial \beta_i}\right\}_{\substack{\alpha = a \\ \beta = b}} = 0 \quad ; \quad i = 1, \ldots, k, \tag{1,4}$$

waarbij $\beta = b$ betekent $\beta_1 = b_1, \beta_2 = b_2, \ldots, \beta_k = b_k$.

---

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

2) Stochastische grootheden geven we met onderstreepte letters aan; denken we speciaal aan de door deze grootheden aangenomen waarden dan geven we deze met dezelfde letters, maar zonder onderstreping, aan.

We zullen de uitdrukking voor a en $b_1, \ldots, b_k$ eerst afleiden voor het geval k = 2 en daarna met behulp van matrixnotatie de algemene formules geven.

2. Kleinste kwadratenschattingen voor het geval k = 2.

De relaties (1,3) en (1,4) luiden nu

$$(2,1) \qquad \left\{\frac{\partial Q}{\partial \alpha}\right\}_{\substack{\alpha=a\\ \beta=b}} = -2\sum_{j=1}^{n}(y_j - a - b_1 x_{1j} - b_2 x_{2j}) = 0$$

en

$$(2,2) \qquad \begin{cases} \left\{\frac{\partial Q}{\partial \beta_1}\right\}_{\substack{\alpha=a\\ \beta=b}} = -2\sum_{j=1}^{n} x_{1j}(y_j - a - b_1 x_{1j} - b_2 x_{2j}) = 0, \\ \left\{\frac{\partial Q}{\partial \beta_2}\right\}_{\substack{\alpha=a\\ \beta=b}} = -2\sum_{j=1}^{n} x_{2j}(y_j - a - b_1 x_{1j} - b_2 x_{2j}) = 0. \end{cases}$$

Uit (2,1) volgt:

$$a = \sum_{j=1}^{n} y_j/n - b_1\sum_{j=1}^{n} x_{1j}/n - b_2\sum_{j=1}^{n} x_{2j}/n =$$
$$= \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2.$$

Substitutie in (2,2) geeft:

$$\begin{cases} \sum_{j=1}^{n} x_{1j}[y_j - \bar{y} - b_1(x_{1j} - \bar{x}_1) - b_2(x_{2j} - \bar{x}_2)] = 0, \\ \sum_{j=1}^{n} x_{2j}[y_j - \bar{y} - b_1(x_{1j} - \bar{x}_1) - b_2(x_{2j} - \bar{x}_2)] = 0, \end{cases}$$

of:

$$\begin{cases} b_1\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2 + b_2\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) = \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(y_j - \bar{y}), \\ b_1\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) + b_2\sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2 = \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)(y_j - \bar{y}). \end{cases}$$

(Immers sommen als $\sum_{j=1}^{n}\bar{x}_1(y_j - \bar{y})$ zijn nul en dus mag $\sum_{j=1}^{n} x_{1j}(y_j - \bar{y})$ vervangen worden door $\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(y_j - \bar{y})$, enz.)

De oplossing van deze lineaire vergelijkingen kunnen wij schrijven in de vorm:

$$b_1 = \begin{vmatrix} \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(y_j - \bar{y}) & \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) \\ \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)(y_j - \bar{y}) & \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2 \end{vmatrix} \Big/ \begin{vmatrix} \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2 & \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) \\ \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) & \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2 \end{vmatrix}$$

en

$$b_2 = \begin{vmatrix} \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2 & \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(y_j - \bar{y}) \\ \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) & \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)(y_j - \bar{y}) \end{vmatrix} \Big/ \begin{vmatrix} \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2 & \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) \\ \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) & \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2 \end{vmatrix}.$$

Voor het geval k = 1 zouden we op deze manier de bekende uitdrukking:

$$b_1 = \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(y_j - \bar{y}) \Big/ \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2$$

gevonden hebben.

3. Kleinste kwadratenschattingen voor het algemene geval.

Ook in het algemene geval vinden we voor a, als oplossing van

$$(3,1)\qquad \left(\frac{\partial \theta}{\partial \alpha}\right)_{\substack{\alpha=a\\ \beta=b}} = -2\sum_{j=1}^{n}(y_j - a - b_1 x_{1j} - \ldots - b_k x_{kj}) = 0,$$

de uitdrukking

$$a = \sum_{j=1}^{n} y_j/n - b_1 \sum_{j=1}^{n} x_{1j}/n - \ldots - b_k \sum_{j=1}^{n} x_{kj}/n =$$

$$= \bar{y} - b_1 \bar{x}_1 - \ldots - b_k \bar{x}_k .$$

Substitueren we dit weer in $\left(\frac{\partial \theta}{\partial \beta_i}\right)_{\substack{\alpha=a\\ \beta=b}} = 0 \ (i=1,\ldots,k)$ dan ontstaat:

$$\sum_{j=1}^{n} x_{ij}\left[y_j - \bar{y} - b_1(x_{1j} - \bar{x}_1) - \ldots - b_k(x_{kj} - \bar{x}_k)\right] = 0 \qquad (i=1,\ldots,k),$$

of

$$(3,2)\qquad b_1 \sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{ij} - \bar{x}_i) + \ldots + b_k \sum_{j=1}^{n}(x_{kj} - \bar{x}_k)(x_{ij} - \bar{x}_i) = \sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(y_j - \bar{y}) \quad (i=1,\ldots,k).$$

We zullen nu om overzichtelijker formules te verkrijgen de volgende matrix notatie invoeren:

$$\underline{y} \overset{\text{def}}{=} \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}; \quad X \overset{\text{def}}{=} \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \ldots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \ldots & x_{k2} - \bar{x}_k \\ \vdots & \vdots & & \vdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \ldots & x_{kn} - \bar{x}_k \end{pmatrix}; \quad \underline{v} \overset{\text{def}}{=} \begin{pmatrix} v_1 - \bar{v} \\ v_2 - \bar{v} \\ \vdots \\ v_n - \bar{v} \end{pmatrix};$$

$$\beta \overset{\text{def}}{=} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{en} \quad b \overset{\text{def}}{=} \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

Nu volgt uit (1,1) dat

$$\bar{y} = \alpha + \beta_1 \bar{x}_1 + \ldots + \beta_k \bar{x}_k + \bar{v},$$

zodat we voor (1,1) ook mogen schrijven:

$$(y_j - \bar{y}) = \beta_1(x_{1j} - \bar{x}_1) + \ldots + \beta_k(x_{kj} - \bar{x}_k) + v_j - \bar{v} \quad ; \quad j = 1, \ldots, n.$$

Met de bovengedefinieerde matrices is dit te schrijven als

$$(3,3)\qquad \underline{y} = X\beta + \underline{v}.$$

De kwadratische vorm welke we moeten minimaliseren heeft nu de gedaante

$$(3,4)\qquad \theta = (\underline{y} - X\beta)'(\underline{y} - X\beta)$$ [3]

3) A', de getransformeerde van matrix A, ontstaat uit A door rijen en kolommen te verwisselen, dus $\begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}' = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$.

en de lineaire vergelijkingen waaraan de kleinste kwadraten-schattingen moeten voldoen gaan over in:

$$(X'X)\,\underline{b} = X'\underline{y}.$$

Hierin is $(X'X)$ een kxk-matrix. Is deze niet singulier dan geeft

(3,5) $$\underline{b} = (X'X)^{-1}X'\underline{y}$$

de gezochte kleinste kwadratenschattingen voor $\beta_1, \ldots, \beta_k$. Om te bereiken dat $(X'X)$ een inverse bezit, moeten we eisen dat X de rang k bezit, hetgeen overeenkomt met de eis dat de regressievectoren $(x_{11}-\bar{x}_1, \ldots, x_{1n}-\bar{x}_1), (x_{21}-\bar{x}_2, \ldots, x_{2n}-\bar{x}_2), \ldots, (x_{k1}-\bar{x}_k, \ldots, x_{kn}-\bar{x}_k)$ lineair onafhankelijk zijn, dus niet via lineaire betrekkingen in elkaar uitgedrukt kunnen worden.
Dit betekent o.a. ook dat $n > k$ is.

De kleinste kwadratenschattingen zijn de beste zuivere lineaire schattingen d.w.z.: ze bezitten van alle mogelijke zuivere lineaire schattingen de kleinste variantie.
Het minimum van $\mathcal{O}$, dat voor de waarden b van β bereikt wordt, is:

(3,6)
$$\begin{aligned}\mathcal{O}_{min} &= (y - Xb)'(y - Xb) = y'y - b'X'y - y'Xb + b'X'Xb = \\ &= y'y - y'X(X'X)^{-1}X'y - y'X(X'X)^{-1}X'y + y'X(X'X)^{-1}X'X(X'X)^{-1}X'y = \\ &= y'y - y'X(X'X)^{-1}X'y = y'y - b'X'y.\end{aligned}$$

4. Verwachtingen, varianties van $\underline{b}_1, \ldots, \underline{b}_k$.

Uit de uitdrukking (3,5) voor $\underline{b}$ zien we dat de $\underline{b}$'s steeds lineaire combinaties van $\underline{y}_1, \ldots, \underline{y}_n$ zijn. Daar in (3,3) $\mathcal{E}\,\underline{v}_j = 0$ is en dus $\mathcal{E}\,\underline{y} = X\beta$, geldt voor de verwachting van b:

$$\mathcal{E}\,\underline{b} = (X'X)^{-1}X'\,\mathcal{E}\,\underline{y} = (X'X)^{-1}X'X\beta =$$

$$= I\beta = \beta.$$

De schattingen $\underline{b}_i$ zijn dus zuiver. Dit geldt ook, indien de $\underline{v}_i$ niet onderling onafhankelijk zijn en/of verschillende spreidingen bezitten, doch wel $\mathcal{E}\,\underline{v}_i = 0$ is $(i = 1, \ldots, n)$.

Zijn de grootheden $\underline{y}_1, \ldots, \underline{y}_n$ wel onderling onafhankelijk verdeeld met dezelfde variantie σ^2, dan kunnen we als volgt de covariantiematrix van $\underline{b}_1, \ldots, \underline{b}_k$ bepalen. We maken weer gebruik van het feit dat $\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(y_j-\bar{y}) = \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)\,y_j$ is, dus dat

$$X'y = \begin{pmatrix} \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(y_j-\bar{y}) \\ \vdots \\ \sum_{j=1}^{n}(x_{kj}-\bar{x}_k)(y_j-\bar{y}) \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n}(x_{1j}-\bar{x}_1)\,y_j \\ \vdots \\ \sum_{j=1}^{n}(x_{kj}-\bar{x}_k)\,y_j \end{pmatrix}.$$

We kunnen dan, volgens (3,5) $b_1, \ldots, b_k$ schrijven als:

$$\begin{cases} b_1 = A_{11}\, y_1 + \ldots + A_{1n}\, y_n , \\ \vdots \\ b_k = A_{k1}\, y_1 + \ldots + A_{kn}\, y_n , \end{cases}$$

waarin de kxk-matrix $A = (X'X)^{-1} X'$ is. Voor de <u>covariantiema-trix</u> van $\underline{b}_1, \ldots, \underline{b}_k$ geldt dan, daar $\sigma\{\underline{y}_j, \underline{y}_l\} = 0 \quad (j \neq l)$ en $\sigma^2\{\underline{y}_j\} = \sigma^2$ is,

$$\left(\sigma^2\{\underline{b}_i, \underline{b}_k\}\right) = \left(\sum_{j=1}^{n} A_{ij} A_{kj}\, \sigma^2\{\underline{y}_j\}\right) = \sigma^2 \cdot \left(\sum_{j=1}^{n} A_{ij} A_{kj}\right) = \sigma^2 \cdot AA' = \sigma^2 \cdot (X'X)^{-1} X' X (X'X)^{-1} = \sigma^2 \cdot (X'X)^{-1}. \tag{4,1}$$

Voor het geval k = 1 is dus:

$$\sigma^2\{\underline{b}_1\} = \sigma^2 / \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)^2 .$$

Voor het geval k = 2 is, wanneer we

$$Z \overset{\text{def.}}{=} \begin{vmatrix} \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)^2 & \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) \\ \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) & \sum_{j=1}^{n} (x_{2j} - \bar{x}_2)^2 \end{vmatrix} = X'X$$

stellen:

$$\sigma^2\{\underline{b}_1\} = \sigma^2 \cdot \sum_{j=1}^{n} (x_{2j} - \bar{x}_2)^2 / |Z| ,$$

$$\sigma^2\{\underline{b}_2\} = \sigma^2 \cdot \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)^2 / |Z| ,$$

$$\sigma^2\{\underline{b}_1, \underline{b}_2\} = -\sigma^2 \cdot \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) / |Z| .$$

En in 't algemeen wanneer we de minor van het element in de i^e rij en k^e kolom van Z door $|Z_{ik}|$ voorstellen:

$$\sigma^2\{\underline{b}_i, \underline{b}_k\} = \sigma^2 \cdot |Z_{ik}| / |Z| . \tag{4,2}$$

Men kan aantonen dat

$$(n - k - 1)^{-1}\, \underline{\vartheta}_{min}$$

een zuivere schatting voor σ^2 is (J. van <u>YZEREN</u> (1954)).

Zijn $\underline{v}_1, \ldots, \underline{v}_n$ onderling onafhankelijk <u>normaal</u> verdeeld met gelijke spreidingen, dan komen de kleinste kwadratenschattingen voor $\alpha, \beta_1, \ldots, \beta_k$ overeen met de meest aannemelijke schattingen (maximum likelihood estimates). De meest aannemelijke schatting voor σ^2 is dan $n^{-1}\, \underline{\vartheta}_{min}$ deze is asymptotisch aequivalent met de zuivere schatting $(n - k - 1)^{-1}\, \underline{\vartheta}_{min}$. De grootheid $\vartheta_{min} / \sigma^2$ heeft nu een -verdeling met (n - k - 1) vrijheidsgraden en is onafhankelijk verdeeld van $\underline{a}, \underline{b}_1, \ldots \underline{b}_k$.

Literatuur

..B. MANN (1949), Analysis and design of experiments, hoofdstuk 4. New York, Dover Publications, Inc.

H. CRAMÉR (1951), Mathematical methods of statistics, hoofdstuk 37. Princeton, University Press.

J. VAN IJZEREN (1954), De theoretische zijde van de methode der kleinste kwadraten. Statistica, 8, pp 21-45.