CENTRAAL PLANBUREAU

Econometristenbijeenkomsten

VERSLAG van de 21ste econometristenbijeenkomst, gehouden op 28 Juni 1949 in het gebouw van het C.P.B.

Voorzitter: dr P.J. Verdoorn
Voordracht van de heer H. Theil

EEN METHODE VAN LINEAIRE REGRESSIE-ANALYSE

1. Inleiding

In de - gebruikelijke - regressie-analyse volgens de methode der kleinste kwadraten gaat men van de veronderstelling uit, dat de afwijkingen der waarnemingen van de regressielijn, welke veroorzaakt worden door meetfouten en door de buiten beschouwing gelaten verklarende variabelen, aan een normale verdeling voldoen. In de door spreker uiteengezette methode wordt deze veronderstelling losgelaten. In de plaats daarvan wordt de eis gesteld, dat in de relatie

$$y = \beta x + u,$$

waarin y de verklaarde, x de verklarende variabele en β de regressie-coëfficient voorstelt[1], u een stochastische variabele is, die aan tenminste één van de volgende voorwaarden voldoet:

A: de waarnemingen u zijn onderling onafhankelijk verdeeld; zij hebben dezelfde cumulatieve verdelingsfunctie F(u), die onafhankelijk is van x; de waarschijnlijkheidsdichtheid F'(u) is continu;

B: de waarnemingen u zijn onderling onafhankelijk verdeeld; de waarschijnlijkheidsdichtheid $F'_x(u)$ van u mag afhankelijk zijn van x, maar is continu en symmetrisch met gemiddelde α.

Stel, dat van u twee waarnemingen zijn verricht, u_1 en u_2. Dan is, indien één van bovenstaande voorwaarden vervuld is, de stochastische variabele $(u_2 - u_1)$ symmetrisch om nul verdeeld.

---

1) In het hierna volgende wordt uitsluitend het geval van 2 variabelen behandeld; de methode is echter ook toepasselijk voor 3 of meer variabelen; voor deze generalisatie wordt verwezen naar een binnenkort in "Econometrica" te verschijnen artikel van de heer Theil.

Indien x en y niet nauwkeurig gemeten zijn, nemen we aan, dat x' en y'zijn gemeten, zodanig, dat

$$x' = x + \xi$$
$$y' = y + \eta$$

en dat ξ en η, hetzij aan voorwaarde A, hetzij aan voorwaarde B voldoen.

2. De regressie-coëfficient in het geval van twee variabelen

Uit de collectie, gekarakteriseerd door

$$y = \beta x + u$$

is een steekproef ter grootte van n getrokken. We rangschikken de punten (x,y) naar toenemende x. Noem $m = \left[\frac{1}{2}n\right]$; indien n oneven is, wordt de (m + 1)ste waarneming niet gebruikt. We bepalen nu de volgende statistische grootheden

$$b_{(i)} = \frac{y_{n-m+i} - y_i}{x_{n-m+i} - x_i}$$
$$= \beta + \frac{u_{n-m+i} - u_i}{x_{n-m+i} - x_i} \qquad (i = 1,\ldots,m)$$

Deze grootheden stellen dus voor de hellingen in het x,y-diagram tussen het eerste en het (n - m + 1)ste, tussen het tweede en het (n - m + 2)de punt, enz.. De kans, dat $b > \beta$ is gelijk aan de kans, dat $b < \beta$, nl. $\frac{1}{2}$.

De waarden b_i (i = 1, 2, ... m) worden nu opnieuw gerangschikt en wel naar opklimmende waarden van b. Spreker toont aan, dat hieruit een betrouwbaarheidsinterval te berekenen is:

$$P\left[b_r \leq \beta \leq b_{m-r+1}\right] = 2^{1-m}\left\{\binom{m}{0} + \binom{m}{1} + \ldots\binom{m}{r-1}\right\}$$
$$= 1\ 2\ I_{0,5}(m - r + 1, r) \quad \ldots(1)$$

De betrouwbaarheidscoëfficient kan afgelezen worden uit K. Pearson's "Tables of the incomplete Beta-function". Voor hogere waarden van n is een approximatie door een normale verdeling met gemiddelde $\frac{1}{2}m$ en spreiding $\frac{1}{2}\sqrt{m}$ voldoende nauwkeurig.

Op eenvoudige wijze kan worden aangetoond, dat meetfouten niet tot wijziging van bovenstaande methode aanleiding geven, indien hun verdelingen aan voorwaarde A of aan voorwaarde B voldoen.

Indien de steekproef waarnemingen bevat met gelijke x-waarden, dan is de rangschikking van de steekproefpunten niet eenvoudig meer. In een dergelijk geval is het echter voldoende, dat de rangschikking van punten met gelijke x-waarden willekeurig plaats heeft.

3. Gebieden van lineaire regressie

Spreker veronderstelt vervolgens, dat de relatie tussen de variabelen x en y wordt weergegeven door $y = \beta x + u$, waarbij u voldoet aan voorwaarde A. Indien een waarde x gegeven is, moet een betrouwbaarheidsinterval voor y worden geconstateerd.

Hierbij wordt gebruik gemaakt van een stelling van W.R. Thompson [2], die als volgt luidt:

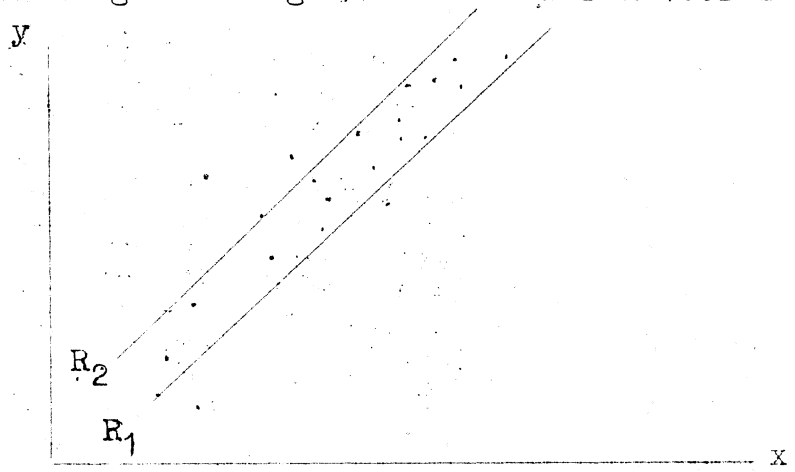
indien een steekproef ter grootte van n wordt getrokken uit een collectie, gekarakteriseerd door een cumulatieve verdelingsfunctie F(u) met continue verdelingsdichtheid F'(u), de steekproefwaarden vervolgens worden gerangschikt naar opklimmende waarde van u ($u_1 < u_2 < \ldots\ldots < u_n$) en tenslotte een (n + 1)ste trekking u wordt verricht, dan is de waarschijnlijkheid, dat u ligt in het interval (u_s, u_{n-s+1}) gelijk aan $1 - \frac{2s}{n+1}$ ..... (2).

Indien β gegeven is, vindt men n steekproefwaarden $u_i = y_i - \beta x_i$. Deze grootheden worden gerangschikt naar opklimmende waarden van u. De waarschijnlijkheid, dat een (n + 1)ste trekking een u oplevert, die ligt in het interval (u_s, u_{n-s+1}) is gelijk aan $1 - \frac{2s}{n+1}$. Hieruit volgt, dat met deze waarschijnlijkheid de volgende relatie geldt:

$$\beta x + u_s \leqq y \leqq \beta x + u_{n-s+1}$$

Onderstaand diagram 1 moge dit illustreren voor s = 2

Diagram 1



2) "Biological applications of the normal range and associated significance tests in ignorance of original distribution forms" (Annals of Mathematical Statistics, IX, 1938, blz. 281-187); aldaar is ook het bewijs van de stelling opgenomen.

De richting van de rechten R_1 en R_2 wordt door β bepaald; dienovereenkomstig geeft spreker het gebied tussen R_1 en R_2 weer door $G(\beta)$. Indien het steekproefpunt van de $(n + 1)$ste trekking wordt weergegeven door D en de steekproefpunten van de n trekkingen door D_i $(i = 1, .., n)$, dan is de waarschijnlijkheid, dat D in het gebied $G(\beta)$ is bevat, gelijk $1 - \frac{2s}{n+1}$. In het algemeen is β niet bekend, zodat men gebruik moet maken van het hierboven weergegeven betrouwbaarheidsinterval

$$P\left[b_r \leq \beta \leq b_{m-r+1} \right] = 1 - 2\, I_{0,5}\,(m - r + 1,\ r).$$

Dit benuttende, kan men stellen, dat de waarschijnlijkheid, dat
1) het interval $(b_r,\ b_{m-r+1})$ de ware β bevat, en
2) de vereniging van alle $G(\beta)$'s voor $b_r \leq \beta \leq b_{m-r+1}$ het punt D bevat,
gelijk is aan of groter is dan

$$\left(1 - \frac{2s}{n+1}\right)\left\{1 - 2\, I_{0,5}\,(m - r + 1,\ r)\right\} \quad \ldots\ldots\ldots \quad (3)$$

Diagram 2 geeft een grafische voorstelling van deze vereniging. In het algemeen zal deze begrensd worden door grenzen van $G(b_r)$ en van $G(b_{m-r+1})$. Men dient te bedenken, dat $G(\beta)$ voor iedere β steeds begrensd wordt door twee rechten, die ieder door een steekproefpunt gaan. Variatie van β geeft een omwenteling van deze rechten om het betreffende punt, totdat een nieuw steekproefpunt op de rechte komt te liggen: dan heeft de omwenteling plaats om dit nieuwe punt. Men zal op geometrische gronden inzien, dat - behoudens gevallen van uitzonderlijke ligging van bepaalde steekproefpunten - het voldoende is, dat uitsluitend de uitersten $G(b_r)$ en $G(b_{m-r+1})$ worden beschouwd.

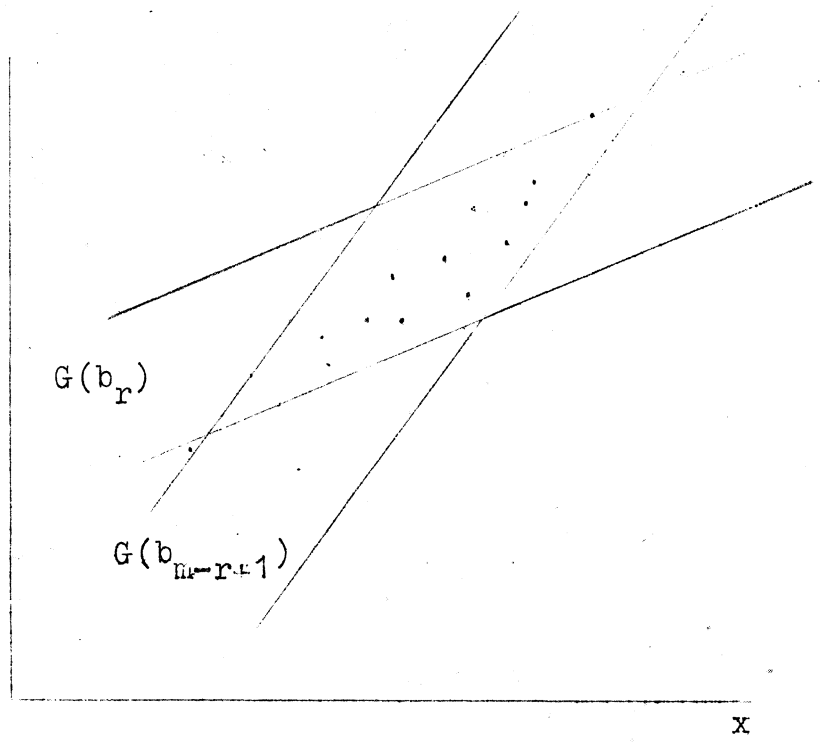


Diagram 2
(s = 2)

De weergegeven vereniging kan genoemd worden: een gebied van lineaire regressie. Met de waarschijnlijkheid (3) of met een grotere waarschijnlijkheid zal dit gebied een (n + 1)ste steekproefpunt bevatten, voor welke x ook.

4. Een toepassing: huishoudrekeningen 1935

In samenwerking met de heer H.S. Houthakker heeft spreker uit de Statistiek der Huishoudrekeningen (C.B.S. 1935) gegevens gelicht voor het totale inkomen (x) en het totale verbruik (y). Getracht is voor de regressiecoëfficient β in de relatie $y = \beta x + u$ een betrouwbaarheidsinterval te berekenen; daarbij zijn de gezinnen ingedeeld in 4 klassen

		Overschrijdingskans (ε)	Betrouwbaarheids-interval
Landarbeiders	r = 5	0,012	0,64 - 1,045
m = 20	6	0,041	0,65 - 1,038
Boeren	r = 10	0,004	0,32 - 0,995
m = 36	11	0,011	0,33 - 0,99
	12	0,029	0,38 - 0,99
	13	0,065	0,40 - 0,97

		Overschrijdingskans (ε)	Betrouwbaarheids-interval
Hoofdarbeiders	r = 39	0,007	0,84 - 1,01
m = 103	40	0,012	0,85 - 1,00
	41	0,020	0,86 - 1,00
	42	0,033	0,87 - 1,00
	43	0,050	0,87 - 1,00
	44	0,073	0,88 - 0,99
Handarbeiders	r = 55	0,007	0,87 - 0,97
m = 139	56	0,011	0,87 - 0,96
	57	0,017	0,87 - 0,96
	58	0,026	0,88 - 0,96
	59	0,036	0,89 - 0,95
	60	0,055	0,89 - 0,94
	61	0,075	0,89 - 0,94

In bijlage wordt een tabel gegeven voor de waarschijnlijkheid $P = 1 - 2\, I_{0,5}\,(m - r + r,\ r) = 1 - \varepsilon$ voor $m \leq 50$.

Bijlage

Hieronder volgt een tabel voor de waarschijnlijkheid

$$P = 1 - 2\, I_{0,5}\,(m - r + 1,\ r) = 1 -$$

voor $m \leqq 50$ (dus voor steekproeven tot de uitgeverijgrootte 100) en voor r zodanig, dat $P \geqq 0{,}700$ is. In de rijen is m afgezet, in de kolommen r. De getallen geven P weer in duizendste delen.

m ↓ \ r →	1	2	3	4	5	6	7	8	9	10	11	12
3	750											
4	875											
5	937											
6	969	781										
7	984	875										
8	992	930	711									
9	996	961	820									
10	998	979	891									
11	999	988	935	773								
12	1000	994	962	854								
13		997	978	908	733							
14		998	987	943	820							
15		999	993	965	882							
16	999	996	979	932	790							
17	1000	998	987	951	857							
18		999	994	970	905	763						
19		999	996	982	937	834						
20		1000	997	988	959	885	737					
21	999	993	973	922	811							
22	999	996	983	950	866	714						
23	1000	997	989	965	907	790						
24		998	993	977	936	848						
25		999	996	985	957	892	770					
26	999	998	991	971	924	831						
27	1000	998	994	981	948	878	752					
28		999	996	987	964	913	815					
29		999	998	992	976	939	864	735				
30		1000	999	995	984	957	901	800				

	7	8	9	10	11	12	13	14		
31	999	997	989	971	929	850	719			
32	999	998	993	980	949	890	785			
33	1000	999	995	986	965	920	837	704		
34		999	997	991	976	942	879	771		
35		999	998	994	983	959	910	825		

	8	9	10	11	12	13	14	15	16	17
36	1000	999	996	989	971	935	868	757		
37		999	997	992	980	953	901	812		
38		1000	998	995	986	966	927	857	744	
39			999	996	991	976	947	889	798	
40			999	998	994	983	962	919	846	732

	10	11	12	13	14	15	16	17	18	19
41	1000	999	996	988	972	940	883	789		
42		999	997	992	980	956	912	836	720	
43		999	998	995	986	968	934	874	778	
44		1000	999	996	990	977	951	904	826	709
45			999	998	993	984	964	928	865	767

	12	13	14	15	16	17	18	19	20	21
46	999	998	995	989	974	946	896	816		
47	1000	999	997	992	981	960	921	856	757	
48		999	998	994	987	971	941	889	807	
49		1000	999	996	991	979	956	915	848	747
50			999	997	993	985	967	935	881	797