STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

AFDELING MATHEMATISCHE STATISTIEK

Rapport S 290 M84

Toetsingen in de lineaire regressieanalyse

door

P. van der Laan

April 1962

Inleiding en model [1]

Gegeven zijn n waarnemingen (bv. metingen)
$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ van n stochastische grootheden
$(x_1,\underline{y}_1),(x_2,\underline{y}_2),\ldots,(x_n,\underline{y}_n)$.[2]

Verondersteld is dat:

a. De grootheden x_i geen stochastisch karakter bezitten. In de praktijk komt dit er vaak op neer dat de stochastische afwijkingen van de x_i veel kleiner zijn dan die van de $\underline{y}_i$ $(i=1,2,\ldots,n)$.

b. Tussen de grootheden x_i en $\underline{y}_i$ het volgende verband bestaat:

$$\underline{y}_i = \alpha + \beta(x_i - \bar{x}) + \underline{v}_i \quad \text{met} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (i=1,2,\ldots,n)$$

waarbij de stochastische afwijkingen $\underline{v}_i$ onderling onafhankelijk normaal verdeeld zijn met verwachting 0 en variantie σ^2.
Dus geldt:

$$\mathcal{E}\,\underline{y}_i = \alpha + \beta(x_i - \bar{x})$$

$$\sigma^2(\underline{y}_i) = \sigma^2 \qquad (i=1,2,\ldots,n).$$

Hieronder volgen nu methoden om enkele veronderstellingen over de parameters α en β te toetsen.

Voor toetsing van de hypothese H_o: $\beta = 0$, zie men memorandum S 73 (M 34) van het Mathematisch Centrum, getiteld: "Toetsing van de hypothese dat een regressiecoëfficiënt nul is, wanneer de afwijkingen van de regressielijn normaal verdeeld zijn met gelijke spreidingen".

Voor toetsing (voor 2 series waarnemingen) van de hypothese H_o: $\beta_1 = \beta_2$, zie men memorandum S 53 (M 27), getiteld: "Toetsing van de gelijkheid van twee regressiecoëfficiënten, indien de afwijkingen normaal verdeeld zijn met gelijke spreidingen".

---

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

2) De onderstreping geeft aan, dat de grootheid stochastisch is, d.w.z. een waarschijnlijkheidsverdeling bezit. Dezelfde letter zonder onderstreping wordt gebruikt voor een door de stochastische grootheid aangenomen waarde.

Voor toetsing van de lineariteit van de regressielijn zie men memorandum S 73 (M 32), getiteld: "Toetsing van de lineariteit van een regressielijn, indien de waarnemingsfouten onderling onafhankelijk en normaal verdeeld zijn met gelijke spreiding".

1. Te toetsen de hypothese: $\alpha = \alpha_o$.

Met behulp van de methode der kleinste kwadraten verkrijgt men voor α en β de volgende zuivere schattingen:

$$\underline{a} = \frac{1}{n}\sum_{i=1}^{n} \underline{Y}_i = \overline{\underline{Y}} \quad \text{resp.} \quad \underline{b} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(\underline{Y}_i - \overline{\underline{Y}})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\underline{S}_{xy}}{S_{xx}}$$

waarbij gedefinieerd is:

$$\underline{S}_{xy} \equiv \sum_{i=1}^{n}(x_i-\overline{x})(\underline{Y}_i-\overline{\underline{Y}})$$

$$S_{xx} \equiv \sum_{i=1}^{n}(x_i-\overline{x})^2$$

$$\underline{S}_{yy} \equiv \sum_{i=1}^{n}(\underline{Y}_i-\overline{\underline{Y}})^2.$$

Terwijl de schatting voor de variantie σ^2 is:

$$\underline{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left\{\underline{Y}_i - \underline{a} - \underline{b}\,(x_i-\overline{x})\right\}^2$$

$$= \frac{1}{n-2}\left\{\underline{S}_{yy} - \frac{(\underline{S}_{xy})^2}{S_{xx}}\right\}.$$

De verdeling van $\underline{a}$ is normaal met verwachting α en variantie $\frac{\sigma^2}{n}$, dus $\frac{\underline{a}-\alpha}{\sigma}\sqrt{n}$ is normaal verdeeld met verwachting 0 en variantie 1.

Tevens is $\frac{(n-2)\underline{s}^2}{\sigma^2}$ verdeeld als $\underline{\chi}^2$ met (n-2) vrijheidsgraden en $\underline{a}$ en $\underline{s}^2$ zijn stochastisch onafhankelijk.

Hieruit volgt dat:

$$\underline{t}_{n-2} \equiv \frac{(\underline{a} - \alpha)\sqrt{n}\,/\sigma}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left\{\frac{\underline{Y}_i - \underline{a} - \underline{b}(x_i - \bar{x})}{\sigma}\right\}^2}} = \frac{(\underline{a} - \alpha)\sqrt{n}}{\underline{s}}$$

een Student-verdeling heeft met (n-2) vrijheidsgraden.

Te toetsen de nulhypothese H_o: $\alpha = \alpha_o$. Onder H_o heeft $\underline{t}_{n-2}$, met $\alpha = \alpha_o$ gesubstitueerd, een t-verdeling met (n-2) vrijheidsgraden. Indien H_o niet vervuld is, bezit $\underline{t}_{n-2}$ een niet-centrale t-verdeling.

Uit het waarnemingsmateriaal is de toetsingsgrootheid t_{n-2} te berekenen. De bijbehorende (één-of tweezijdige) overschrijdingskans is in een tabel van de Student-verdeling met (n-2) vrijheidsgraden op te zoeken. (Voor zeer grote n kan men als benadering de normale verdeling gebruiken).

Vindt men nu een overschrijdingskans, welke kleiner is dan de, van te voren gekozen, onbetrouwbaarheidsdrempel α, dan wordt H_o verworpen, ten gunste van één van de alternatieve hypothesen $\alpha > \alpha_o$ of $\alpha < \alpha_o$.

Indien men H_o niet verwerpt, wil dit zeggen dat er, met het gegeven waarnemingsmateriaal, geen overtuigende reden is, H_o onjuist te achten.

<u>Opmerking:</u> Op analoge wijze is voor $\underline{Y}_i = \alpha + \beta x_i + \underline{v}_i$ de hypothese $\alpha = \alpha_o$ te toetsen. Als bijzonder geval heeft men $\alpha = 0$, d.w.z. de lijn gaat door de oorsprong.

2. Behoort een extra waarneming (x_{n+1}, y_{n+1}) tot de rechte?

Dezelfde veronderstellingen gelden, als in de inleiding zijn gemaakt. We doen nu nog één extra waarneming (x_{n+1}, y_{n+1}). Indien nu aan de volgende voorwaarden is voldaan, namelijk dat:

a) $\underline{y}_{n+1}$ onafhankelijk van $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_n$.

b) $\sigma^2(\underline{y}_{n+1}) = \sigma^2$

c) $\underline{y}_{n+1}$ is normaal verdeeld,

kan de nulhypothese H_o: $\mathcal{E}\, \underline{y}_{n+1} = \alpha + \beta\ (x_{n+1} - \bar{x})$ $\quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

als volgt worden getoetst.

Onder H_o geldt dat:

$$\underline{t}_{n-2} \equiv \frac{\underline{y}_{n+1} - \underline{a} - \underline{b}\ (x_{n+1} - \bar{x})}{\underline{s}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}}$$

een Student-verdeling heeft met (n-2) vrijheidsgraden. Indien H_o niet is vervuld, heeft $\underline{t}_{n-2}$ een niet-centrale t-verdeling.

Uit het waarnemingsmateriaal berekent men t_{n-2} en zoekt men de bijbehorende (één-of tweezijdige) overschrijdingskans op in een tabel van de Student-verdeling met (n-2)vrijheidsgraden. Afhankelijk van het feit, of de overschrijdingskans kleiner dan wel groter is dan de onbetrouwbaarheidsdrempel α(bv. 0.05), kan men H_o al of niet verwerpen.

Opmerking: Ten aanzien van de tweede voorwaarde is in de praktijk voorzichtigheid geboden. Uiteraard geldt dat, indien aan b) niet is voldaan, de toets zijn geldigheid verliest.

3. Zijn twee gevonden regressielijnen identiek?

Gegeven zijn twee series waarnemingsparen:
$(x_1^{(1)}, y_1^{(1)}), (x_2^{(1)}, y_2^{(1)}),.....(x_n^{(1)}, y_n^{(1)})$ en
$(x_1^{(2)}, y_1^{(2)}), (x_2^{(2)}, y_2^{(2)}),.....(x_m^{(2)}, y_m^{(2)})$.

Voor beide series onderling onafhankelijke waarnemingen gelden de voorwaarden, als in de inleiding gemaakt. Dus:

$$\mathcal{E}\, \underline{y}_i^{(1)} = \alpha^{(1)} + \beta^{(1)}(x_i^{(1)} - \overline{x}^{(1)}) \qquad \sigma^2(\underline{y}_i^{(1)}) = \sigma^2 \qquad (i=1,2,..,n)$$

$$\mathcal{E}\, \underline{y}_j^{(2)} = \alpha^{(2)} + \beta^{(2)}(x_j^{(2)} - \overline{x}^{(2)}) \qquad \sigma^2(\underline{y}_j^{(2)}) = \sigma^2 \qquad (j=1,2,..,m)$$

Te toetsen de nulhypothese $H_o: \alpha^{(1)} = \alpha^{(2)}$, $\beta^{(1)} = \beta^{(2)}$.
Onder H_o (beide series waarnemingen afkomstig van één rechte, welke uiteraard onbekend is) is de toetsingsgrootheid

$$\underline{F}_{2,n+m-4} = \frac{\{\underline{Q} - (\underline{Q}_1 + \underline{Q}_2)\}/2}{(\underline{Q}_1 + \underline{Q}_2)/(n + m - 4)}$$

verdeeld als de $\underline{F}$ van FISHER met 2 en (n + m - 4) vrijheidsgraden.
Hierin is:

$$\underline{Q}_1 = \sum_{i=1}^{n} \left\{ \underline{y}_i^{(1)} - \underline{a}^{(1)} - b^{(1)}(x_i^{(1)} - \overline{x}^{(1)}) \right\}^2$$

$$\underline{Q}_2 = \sum_{j=1}^{m} \left\{ \underline{y}_j^{(2)} - \underline{a}^{(2)} - b^{(2)}(x_j^{(2)} - \overline{x}^{(2)}) \right\}^2$$

$$\underline{Q} = \sum_{k=1}^{n+m} \left\{ \underline{y}_k - \underline{a} - \underline{b}\,(x_k - \overline{x}) \right\}^2 ,$$

waarbij de laatste uitdrukking is opgeschreven voor de beide series waarnemingsparen tezamen genomen tot één serie van (n + m) waarnemingen.

Uit het waarnemingsmateriaal is de toetsingsgrootheid $F_{2,n+m-4}$ te berekenen. De bijbehorende (één-of tweezijdige) overschrijdingskans kan men opzoeken in een tabel van de F-verdeling. Bij een, van te voren gekozen onbetrouwbaarheidsdrempel α , kan men beslissen H_o al of niet te verwerpen.

4. Uitbreiding tot het geval van meerdere regressielijnen

Gegeven zijn k series waarnemingen:

$$(x_1^{(1)}, y_1^{(1)}), (x_2^{(1)}, y_2^{(1)}), \ldots\ldots\ldots, (x_{n_1}^{(1)}, y_{n_1}^{(1)})$$

$$(x_1^{(2)}, y_1^{(2)}), (x_2^{(2)}, y_2^{(2)}), \ldots\ldots\ldots, (x_{n_2}^{(2)}, y_{n_2}^{(2)})$$

$$\vdots \qquad\qquad \ddots$$

$$(x_1^{(k)}, y_1^{(k)}), (x_2^{(k)}, y_2^{(k)}), \ldots\ldots\ldots, (x_{n_k}^{(k)}, y_{n_k}^{(k)}).$$

Wil men een gelijksoorige hypothese H_o toetsen, als onder punt 3, namelijk

$$H_o: \alpha^{(1)} = \alpha^{(2)} = \ldots\ldots\ldots = \alpha^{(k)}$$
$$\beta^{(1)} = \beta^{(2)} = \ldots\ldots\ldots = \beta^{(k)}$$

dan heeft men als samengestelde toetsingsgrootheid:

$$\underline{F}_{2k-2,\ \sum_{i=1}^{k} n_i - 2k} \equiv \frac{(\underline{Q} - \sum_{i=1}^{k} \underline{Q}_i)\ /(2k-2)}{\sum_{i=1}^{k} \underline{Q}_i\ /\left(\sum_{i=1}^{k} n_i - 2k\right)} .$$

Deze heeft onder de nulhypothese een F-verdeling met (2k-2) en $\left(\sum_{i=1}^{k} n_i - 2k\right)$ vrijheidsgraden.

Uit het waarnemingsmateriaal is de toetsingsgrootheid

$F_{2k-2,\ \sum_{i=1}^{k} n_i - 2k}$ te berekenen en uit de tabellen kan

men de (één - of tweezijdige) overschrijdingskans opzoeken. Afhankelijk van de van te voren als criterium gekozen onbetrouwbaarheidsdrempel α , kan men H_o al of niet verwerpen.

Literatuur

1. F.S. ACTON, Analysis of Straight-Line Data, New York, John Wiley & Sons, Inc. 1959.

2. H. CRAMÉR, Mathematical Methods of Statistics, Princeton University Press 1946.

3. R.A. FISHER and F. YATES, Statistical Tables for Biological agricultural and medical research, 3d Ed., Oliver & Boyd, London 1949.

4. A.M. MOOD, Introduction to the theory of Statistics Mc. Graw-Hill, New-York - Toronto-London 1950.

5. E.S. PEARSON and H.O. HARTLEY, Biometrika Tables for Statisticians, Cambridge, University Press 1954.

6. H. SCHEFFÉ, The Analysis of Variance, New York, John Wiley & Sons, Inc. 1959.