STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 73 (M 32)

Toetsing van de lineariteit van een regressielijn indien de waarnemingsfouten onderling onafhankelijk en normaal verdeeld zijn met gelijke spreidingen.

[1952]

MATHEMATISCH CENTRUM,
2de Boerhaavestr. 49,
Amsterdam (O)
Statistische Afdeling
S 73 (M 32).

SM 32

Toetsing van de lineariteit van een regressielijn indien de waarnemingsfouten onderling onafhankelijk en normaal verdeeld zijn met gelijke spreidingen.[1]

Gegeven zij dat van twee grootheden ξ en η de grootheid ξ foutloos kan worden waargenomen, terwijl bij de waarnemingen van η onderling onafhankelijke fouten [2] worden gemaakt, die alle een normale [3] verdeling bezitten met gemiddelde 0 en onbekende, maar steeds dezelfde, spreiding σ.

Zij verder gegeven dat bij iedere waarneming $x_i'(i=1,\dots,k)$ van ξ een aantal waarnemingen $y_{ij}'(j=1,\dots,n_i)$ van η beschikbaar is.

Gevraagd wordt om, op grond van dit waarnemingsmateriaal, dus:

$$(x_1',y_{11}'),(x_1',y_{12}')\dots(x_1',y_{1n_1}'),$$
$$\vdots$$
$$(x_h',y_{h_1}'),(x_h',y_{h_2}')\dots(x_h',y_{hn_1}'),$$

de hypothese H_o te toetsen dat er een lineair verband tussen η en ξ bestaat:

$$H_o:\dots\dots \qquad \eta=\beta\xi+\alpha\,.$$

Om H_o te toetsen worden uit de waarnemingen schattingen b en a en voor β en α berekend door toepassing van de methode der kleinste kwadraten, die in dit geval overeenkomt met de methode der meest aannemelijke schattingen (Eng. method of maximum likelihood). Hierdoor wordt gevonden:

---

[1] Dit memorandum is slechts bedoeld ter orientatie en streeft niet naar volledigheid of volledige exactheid.

[2] Onder fouten vallen in dit geval ook toevallige afwijkingen van andere aard, bv. physiologische afwijkingen.

[3] De stochastische grootheid $\underline{x}$ bezit een normale verdeling met gemiddelde μ en spreiding σ, indien voor iedere a geldt:

$$P\left[\underline{x}\leq a\right]=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}}\,du.$$

$$b = \frac{\sum_{i=1}^{h} n_i(x_i'-\overline{x'})y_i'}{\sum_{i=1}^{h} n_i(x_i'-\overline{x'})^2},$$

$$a = \overline{y'} - b\overline{x'},$$

waarin, als $N = \sum_{i=1}^{h} n_i$,

$$y_i' = \frac{\sum_{j=1}^{n_i} y_{ij}'}{n_i}\ ;\ \overline{y'} = \frac{\sum_{i=1}^{h} n_i y_i'}{N}\ ;\ \overline{x'} = \frac{\sum_{i=1}^{h} n_i x_i'}{N},$$

Een schatting voor σ^2 wordt gegeven door:

$$s^2 = \frac{\sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij}'-y_i')^2}{N-h} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{n_i} y_{ij}'^2 - \sum_{i=1}^{h} n_i y_i'^2}{N-h}.$$

Uit de afwijkingen van de gemiddelden y_i' van de aangepaste regressielijn kan, indien de hypothese H_0 juist is, een tweede schatting voor σ^2 worden afgeleid:

$$s'^2 = \frac{\sum_{i=1}^{h} n_i(y_i'-bx_i'-a)^2}{h-2} = \frac{\sum_{i=1}^{h} n_i y_i'^2 - b^2\sum_{i=1}^{h} n_i(x_i'-\overline{x'})^2 - N\overline{y'}^2}{h-2}.$$

Beschouwen we niet alleen de gevonden waarden $y_{ij}'(j=1,\dots,n_i;i=1,\dots,h)$ maar de verzameling van alle mogelijke waarden die aangenomen kunnen worden, dan bezitten de genoemde schattingen op deze verzameling een waarschijnlijkheidsverdeling, die afhankelijk is van de (exact waargenomen) waarden $x_i'(i=1,\dots,h)$. Dit wordt door onderstreping van de desbetreffende symbolen aangegeven.

De grootheden $\underline{s}^2$ en $\underline{s}'^2$ zijn onderling onafhankelijk verdeeld, zoals bv. door Mood [1] bewezen wordt (hoofdstuk 13) en

$\frac{(N-h)\underline{s}^2}{\sigma^2}$ en $\frac{(h-2)\underline{s}'^2}{\sigma^2}$ bezitten χ^2-verdelingen met respectievelijk (N-h) en (h-2) vrijheidsgraden.

Hieruit volgt, dat de grootheid

$$\underline{F} = \frac{\underline{s}'^2}{\underline{s}^2}$$

verdeeld is volgens de verdeling van de grootheid $\underline{F}$ van Snedecor met (h-2) en (N-h) vrijheidsgraden.

Ook wanneer er geen lineair verband tussen η en ξ bestaat zal $\underline{s}^2$ een goede schatting van σ^2 blijven, voor $\underline{s}'^2$ daarentegen en dus ook voor $\underline{F}$ zullen de grotere waarden een grotere waarschijnlijkheid bezitten. Voor toetsingen van de hypothese H_0 gebruikt men dan ook een éénzijdige kritieke zône van de vorm $\underline{F} \geq F_0$ waarbij F_0, behorende bij een bepaalde onbetrouwbaarheidsdrempel α en bepaalde aantallen vrijheidsgraden, opgezocht kan worden in tabellen vermeld in onderstaande litteratuurlijst

Litteratuur.

[1] A.M. Mood. Introduction to the theory of statistics, Mc. Graw-Hill, New York-Toronto-London, 1950, hoofdstuk 13(regressietheorie), hoofdstuk 14, p.322 (lineariteitstoets), pp.426-427 (tabel).

[2] H. Cramer. Mathematical methods of statistics, Princeton Un. Press., Princeton 1946.

[3] C.W. Emmens. Principles of Biologi al assay, Chapman and Hall Ltd, London, 1948, pp. 30-33(tabellen).