KONINKLIJKE NEDERLANDSE AKADEMIE
VAN WETENSCHAPPEN

# A rank-invariant method of linear and polynomial regression analysis

## III

BY

H. THEIL

Reprinted from Proceedings Vol. LIII, No. 9, 1950

Reprinted from Indagationes Mathematicae, Vol. XII, Fasc. 5, 1950

1950

N.V. NOORD-HOLLANDSCHE UITGEVERS MAATSCHAPPIJ

(NORTH-HOLLAND PUBLISHING COMPANY)

AMSTERDAM

# A RANK-INVARIANT METHOD OF LINEAR AND POLYNOMIAL REGRESSION ANALYSIS. III [1])

BY

H. THEIL

(Communicated by Prof. D. VAN DANTZIG at the meeting of Sept. 30, 1950)

## 3. CONFIDENCE REGIONS FOR THE PARAMETERS OF POLYNOMIAL REGRESSION EQUATIONS

*The probability set*

3. 0. The probability set $\Gamma$ underlying the probability statements of this section is the $n(\nu+2)$-dimensional Cartesian space $R_{n(\nu+2)}$ with coordinates

$$u_{11},\ldots,u_{1n},\ldots,u_{\nu 1},\ldots,u_{\nu n},\ v_1,\ldots,v_n,\ w_1,\ldots,w_n.$$

Every random variable mentioned is supposed to be defined on this probability set.

We suppose $n(\nu+2)$ random variables $\mathbf{u}_{\lambda i}$, $\mathbf{v}_i$, $\mathbf{w}_i$ $(\lambda=1,\ldots,\nu;\ i=1,\ldots,n)$ to have a simultaneous probability distribution on $\Gamma$. Furthermore we consider $n\nu$ parameters $\xi_{\lambda i}$ and $N$ parameters $\alpha_{p_1\ldots p_\nu}$ for all sets of non-negative integers $p_1,\ldots,p_\nu$ satisfying

$$0 \leqq \sum_{\lambda=1}^{\nu} p_\lambda \leqq h.$$

Now we put [2])

$$\left.\begin{aligned}
&\theta_i = \textstyle\sum \alpha_{p_1\ldots p_\nu}\xi_{1i}^{p_1}\ldots\xi_{\nu i}^{p_\nu} && (10)\\
&\eta_i = \theta_i + \mathbf{w}_i && (11)\\
&\mathbf{x}_{\lambda i} = \xi_{\lambda i} + \mathbf{u}_{\lambda i} && (12)\\
&\mathbf{y}_i = \eta_i + \mathbf{v}_i. && (13)
\end{aligned}\right\}\ \begin{cases} i=1,\ldots,n\\ \lambda=1,\ldots,\nu\end{cases}$$

So, for any set of values of the $(N+n\nu)$ parameters $\alpha_{p_1,\ldots,p_\nu}$, $\xi_{\lambda i}$, the variables $\mathbf{x}_{\lambda i}$ and $\mathbf{y}_i$ have a simultaneous distribution on $\Gamma$, and are therefore random variables.

The parameters $\xi_{\lambda i}$ $(i=1,\ldots,n)$ are interpreted as values assumed by the variable $\xi_\lambda$. The equation (10) is the polynomial regression equation. The random variables $\mathbf{w}_i$ are called "the true deviations" from the poly-

[1]) This paper is the third of a series of papers, the first of which appeared in these Proceedings, **53**, 386—392 (1950); the second appeared in these Proceedings, **53**, 521—525 (1950).

[2]) $\Sigma$ in equation (10) denotes summation over all sets $p_1,\ldots,p_\nu$.

nomial of degree $h$; the random variables $\mathbf{u}_{\lambda i}$ and $\mathbf{v}_i$ are called "the errors of observation" of the "true" values $\xi_{\lambda i}$ and $\eta_i$ respectively. [3])

*Conditions; approximation*

3. 1. In order to give confidence regions for the parameters $\alpha_{p_1,\ldots,p}$ we consider the following conditions:

*Condition I:* All $n$ $(\nu+2)$-uples $(\mathbf{u}_{\lambda i}, \mathbf{v}_i, \mathbf{w}_i)$ are stochastically independent.

*Condition IIa:* 1. Each of the errors $\mathbf{u}_{\lambda i}$ vanishes outside a finite interval $|\mathbf{u}_{\lambda i}| < g_{\lambda i}$.

2. For each $i \neq j$ we have $|\xi_{\lambda i} - \xi_{\lambda j}| > g_{\lambda i} + g_{\lambda j}$.

*Condition IIb:* 1. Each of the errors $\mathbf{u}_{\lambda i}$ vanishes outside a finite interval $|\mathbf{u}_{\lambda i}| < g_{\lambda i}$.

2. For each $i \neq j$, for each set $p_1, \ldots, p_\nu$ and for any real $h_{\lambda i}$ such that $|h_{\lambda i}| \leqq g_{\lambda i}$ we have

$$\operatorname{sgn}\left\{\prod_{\lambda=1}^{\nu}(\xi_{\lambda i}+h_{\lambda i})^{p_\lambda}-\prod_{\lambda=1}^{\nu}(\xi_{\lambda j}+h_{\lambda j})^{p_\lambda}\right\}=\operatorname{sgn}\left\{\prod_{\lambda=1}^{\nu}\xi_{\lambda i}^{p_\lambda}-\prod_{\lambda=1}^{\nu}\xi_{\lambda j}^{p_\lambda}\right\}.$$

*Condition III:* For all fixed values of the constants $\varrho_{\lambda i}$ the $n$ random variables $\sum_{\lambda=1}^{\nu} \varrho_{\lambda i}\, \mathbf{u}_{\lambda i} + \mathbf{v}_i + \mathbf{w}_i = \mathbf{z}_i$ have continuous distribution functions, which are symmetrical with the median med $(\mathbf{z})$.

Finally we mention that the solution will be given subject to the following

*Approximation*: For any positive $s$ the quantities

$$\mathbf{u}_{\lambda i}^s\, \mathbf{u}_{\lambda' i}, \quad \mathbf{u}_{\lambda i}^s\, \mathbf{v}_i, \quad \mathbf{u}_{\lambda i}^s\, \mathbf{w}_i \qquad (\lambda, \lambda' = 1, \ldots \nu;\ i = 1, \ldots, n)$$

are neglected. [4])

*Confidence regions*

3. 2. We consider the case $\nu = 1$, so that equation (10) can be written as

$$\theta_i = \sum_{p=0}^{h} \alpha_p\, \xi_i^p.$$

Let us arrange the $n$ observed points $(x_i, y_i)$ according to increasing values of $x$:

$$x_1 < \ldots < x_n.$$

---

[3]) It is clear that the random variables $\mathbf{v}_i$ and $\mathbf{w}_i$ cannot be separated in one sample of observations; if, however, the experiment is repeated for the same "true" values $\xi_{\lambda i}$, $\eta_i$ (e.g. if — when the relation between income and consumption is investigated — for the same families and the same period the amounts of their incomes and outlays are repeatedly calculated), then the errors $\mathbf{v}_i$ can be mitigated by averaging, whereas the deviations $\mathbf{w}_i$ cannot.

[4]) The approximation implies that the errors $\mathbf{u}_{\lambda i}$ are sufficiently small. This restriction is not very serious, because, unless the number of points $n$ is very large, large v lues of $\mathbf{u}_{\lambda i}$ will cause the confidence region for the parameters of the polynomial to be so large as to render the method useless.

We leave $0, 1, \ldots$ or $h$ points out of consideration until the remaining number $n'$ is such that $n'/(h+1)$ is an integer, and write $n_h = n'/(h+1)$. (It seems advisable with respect to the power of the method to omit the points with rank $n_h + 1, 2n_h + 1, \ldots$ and / or $hn_h + 1$). From now on we write $n$ for the remaining number $n'$, so that $(h+1)\, n_h = n$.

We define the following quantities:

$$\Delta(i, n_h + i) = \frac{y_i - y_{n_h+i}}{x_i - x_{n_h+i}}$$

$$\Delta^{(2)}(i, n_h + i, 2n_h + i) = \frac{\Delta(i, n_h + i) - \Delta(n_h + i, 2n_h + i)}{x_i - x_{2n_h+i}}$$

$$\vdots$$

$$\Delta^{(h)}(i, n_h + i, \ldots, hn_h + i) =$$

$$= \frac{\Delta^{(h-1)}(i, n_h + i, \ldots, \overline{h-1}\, n_h + i) - \Delta^{(h-1)}(n_h + i, 2n_h + i, \ldots, hn_h + i)}{x_i - x_{hn_h+i}}$$

We arrange the observed quantities $\Delta^{(h)}(i, \ldots, hn_h + i)$ according to increasing magnitude:

$$\Delta_1^{(h)} < \ldots < \Delta_{n_h}^{(h)},$$

in which

$$\Delta_j^{(h)} = \Delta^{(h)}(i_j, \ldots, hn_h + i_j).$$

3. 3. Then we have the following theorem:

*Theorem* 6: Under conditions I, II$a$, and III the interval $(\Delta_{r_h}^{(h)}, \Delta_{n_h - r_h + 1}^{(h)})$ is a confidence interval for $a_h$ to the approximate level of significance $2I_{\frac{1}{2}}(n_h - r_h + 1, r_h)$. [5])

In order to prove this theorem we shall use the following *Lemma.* Define for all non-negative integers $s$ and for all positive integers $c$ and $i$

$$P^s_{i, n_h+i, \ldots, (c-1)n_h+i} = \sum_{s_i} \cdots \sum_{s_{(c-1)n_h+i}} x_i^{s_i} \ldots x_{(c-1)n_h+i}^{s_{(c-1)n_h+i}}.$$

$$s_j \geqq 0,\ \Sigma s_j = s$$

Then we have

$$\frac{P^s_{i, \ldots, (c-1)n_h+i} - P^s_{n_h+i, \ldots, cn_h+i}}{x_i - x_{cn_h+i}} = P^{s-1}_{i, \ldots, cn_h+i}.$$

*Proof of the lemma*: We have

$$P^s_{i, \ldots, (c-1)n_h+i} - P^s_{n_h+i, \ldots, cn_h+i} =$$

$$= \sum_{s_i} \cdots \sum_{s_{(c-1)n_h+i}} x_{n_h+i}^{s_{n_h+i}} \ldots x_{(c-1)n_h+i}^{s_{(c-1)n_h+i}} (x_i^{s_i} - x_{cn_h+i}^{s_i}),$$

[5]) In the first and second part of this paper the arguments of the incomplete Beta-function must be reversed.

in which $\Sigma s_j = s$. It follows that

$$\frac{P^s_{i,\ldots,(c-1)n_h+i} - P^s_{n_h+i,\ldots,cn_h+i}}{x_i - x_{cn_h+i}} =$$

$$= \sum_{s_i} \cdots \sum_{s_{(c-1)n_h+i}} x^{s_{n_h+i}}_{n_h+i} \ldots x^{s_{(c-1)n_h+i}}_{(c-1)n_h+i} \left(x_i^{s_i-1} + x_i^{s_i-2} x_{cn_h+i} + \ldots + x^{s_i-1}_{cn_h+i}\right) =$$

$$= P^{s-1}_{i,\ldots,cn_h+i}.$$

*Proof of Theorem* 6: The relation between $\mathbf{x}_i$ and $\mathbf{y}_i$ is given by

$$\mathbf{y}_i = \sum_{p=0}^{h} \alpha_p (\mathbf{x}_i - \mathbf{u}_i)^p + \mathbf{v}_i + \mathbf{w}_i$$

$$\approx \sum_{p=0}^{h} \alpha_p \mathbf{x}_i^p - \mathbf{u}_i (\alpha_1 + 2\alpha_2 \xi_i + \ldots + h\alpha_h \xi_i^{h-1}) + \mathbf{v}_i + \mathbf{w}_i,$$

in which we neglected (in accordance with the Approximation) $\mathbf{u}_i^s$ for $s > 1$. Putting $\mathbf{z}_i = \varrho_i \mathbf{u}_i + \mathbf{v}_i + \mathbf{w}_i$, in which

$$-\varrho_i = \alpha_1 + 2\alpha_2 \xi_i + \ldots + h\alpha_h \xi_i^{h-1},$$

we get

$$\mathbf{y}_i \approx \sum_{p=0}^{h} \alpha_p \mathbf{x}_i^p + \mathbf{z}_i.$$

Now we have according to the lemma:

$$\Delta(i, n_h+i) \approx \alpha_1 + \alpha_2 \frac{\mathbf{x}_i^2 - \mathbf{x}^2_{n_h+i}}{\mathbf{x}_i - \mathbf{x}_{n_h+i}} + \ldots + \alpha_h \frac{\mathbf{x}_i^h - \mathbf{x}^h_{n_h+i}}{\mathbf{x}_i - \mathbf{x}_{n_h+i}} + \frac{\mathbf{z}_i - \mathbf{z}_{n_h+i}}{\mathbf{x}_i - \mathbf{x}_{n_h+i}}$$

$$= \alpha_1 + \alpha_2 \mathbf{P}^1_{i,n_h+i} + \ldots + \alpha_h \mathbf{P}^{h-1}_{i,n_h+i} + \frac{\mathbf{z}_i - \mathbf{z}_{n_h+i}}{\mathbf{x}_i - \mathbf{x}_{n_h+i}}.$$

$$\Delta^{(2)}(i, n_h+i, 2n_h+i) \approx \alpha_2 + \alpha_3 \mathbf{P}^1_{i,n_h+i,2n_h+i} + \ldots + \alpha_h \mathbf{P}^{h-2}_{i,n_h+i,2n_h+i} +$$

$$+ \frac{\dfrac{\mathbf{z}_i - \mathbf{z}_{n_h+i}}{\mathbf{x}_i - \mathbf{x}_{n_h+i}} - \dfrac{\mathbf{z}_{n_h+i} - \mathbf{z}_{2n_h+i}}{\mathbf{x}_{n_h+i} - \mathbf{x}_{2n_h+i}}}{\mathbf{x}_i - \mathbf{x}_{2n_h+i}}.$$

$$\vdots$$

$$\Delta^{(h)}(i, n_h+i, \ldots, hn_h+i) \approx \alpha_h + \mathbf{Z}_i,$$

in which $\mathbf{Z}_i$ is a random variable depending on

$$\mathbf{x}_i, \ldots, \mathbf{x}_{hn_h+i} \quad , \quad \mathbf{z}_i, \ldots, \mathbf{z}_{hn_h+i}.$$

$\mathbf{Z}_i$ can be written as a fraction, the denominator being a product of terms $(\mathbf{x}_{cn_h+i} - \mathbf{x}_{c'n_h+i})$ $(c, c' = 0, \ldots, h;\ c \neq c')$; according to condition II$a$ this denominator has a definite sign. The numerator consists of a sum of terms

$$(\mathbf{z}_{cn_h+i} - \mathbf{z}_{c'n_h+i}) \prod_{c'', c'''} (\mathbf{x}_{c''n_h+i} - \mathbf{x}_{c'''n_h+i}).$$

But to our order of approximation this is equal to

$$(z_{cn_h+i} - z_{c'n_h+i}) \prod_{c'',c'''} (\xi_{c''n_h+i} - \xi_{c'''n_h+i}),$$

so that the numerator can be written as

$$\sum_{c=0}^{h} \tau_c z_{cn_h+i} \text{ with } \sum_c \tau_c = 0.$$

It follows from condition III that this quantity has zero median. From this and from the above-mentioned property of the denominator it follows that

$$P[\Delta^{(h)}(i, \ldots, hn_h+i) < a_h \mid a_h] = P[\Delta^{(h)}(i, \ldots, hn_h+i) > a_h \mid a_h] = \tfrac{1}{2}.$$

From this and from condition I the theorem immediately follows.

*Additional remarks*

3. 5. If $a_h$ is known, a confidence interval for $a_{h-1}$ can be found. Consider the equation

$$y_i - a_h x_i^h \approx \sum_0^{h-1} a_p x_i^p + z_i,$$

which shows that the problem is reduced to the case of a polynomial of degree $(h-1)$. So, if a confidence interval for $a_h$ is given, a confidence region for $a_h$ and $a_{h-1}$ can be found. This can be generalized to an $(h+1)$-dimensional confidence region for the parameters $a_0, \ldots, a_h$ in a way analogous to the one described in 2. 2. and 2. 3.

3. 6. If $\nu > 1$, an $N$-dimensional confidence region for the parameters $a_{p_1 \ldots p_\nu}$ can be found in the following way:

1. Given the other parameters, a confidence region for the parameters $a_{p_1 0 \ldots 0}$ $(p_1 = 0, \ldots, h)$ in the $N$-dimensional parameter space can be constructed, the level of significance being $\varepsilon_1$ (cf. 3. 5.).

2. In the same way one can proceed with the parameters

$$a_{0p_2 \ldots 0}, \ldots, a_{00 \ldots p_\nu} \quad (p_\lambda = 1, \ldots, h),$$

the levels of significance being $\varepsilon_2, \ldots, \varepsilon_\nu$.

3. Finally the parameters $a_{p_1^1 \ldots p_\nu^1}$ which have at least two indices $p_\lambda^1 \neq 0$. We suppose that (apart from the conditions I and III) condition II*b* is valid, which is a more stringent condition than condition II*a*. Consider the equation

$$R_i \equiv y_i - \Sigma a_{p_1 \ldots p_\nu} x_{1i}^{p_1} \ldots x_{\nu i}^{p_\nu} \approx a_{p_1^1 \ldots p_\nu^1} x_{1i}^{p_1^1} \ldots x_{\nu i}^{p_\nu^1} + z_i,$$

in which $\Sigma$ denotes summation over all sets $p_1, \ldots, p_\nu$ except the set $p_1^1, \ldots, p_\nu^1$. We can then state regarding the quantities

$$S_{ij} = \frac{R_i - R_j}{x_{1i}^{p_1^1} \ldots x_{\nu i}^{p_\nu^1} - x_{1j}^{p_1^1} \ldots x_{\nu j}^{p_\nu^1}}$$

that

$$P[\mathfrak{S}_{ij} < \alpha_{p_1^1 \dots p_\nu^1} \mid \alpha_{p_1^1 \dots p_\nu^1}] = P[\mathfrak{S}_{ij} > \alpha_{p_1^1 \dots p_\nu^1} \mid \alpha_{p_1^1 \dots p_\nu^1}] = \tfrac{1}{2},$$

so that in a well-known way confidence regions for each of the parameters $\alpha_{p_1^1 \dots p_\nu^1}$ can be found with levels of significance

$$\varepsilon_{\nu+1}, \dots, \varepsilon_{N-(h-1)\nu-1}.$$

The common part of the $N-(h-1)\nu-1$ regions is a confidence region for the "true parameter point" in the $N$-dimensional parameter space, the level of significance being

$$\leqq \sum_{q=1}^{N-(h-1)\nu-1} \varepsilon_q.$$

## 4. CONFIDENCE REGIONS FOR THE PARAMETERS OF SYSTEMS OF REGRESSION EQUATIONS.

4. 0. In recent years considerable work has been done on the subject of systems of regression equations (see e.g. T. HAAVELMO (1943, 1944), T. KOOPMANS (1945, 1950), R. BENTZEL and H. WOLD (1946), M. A. GIRSHICK and T. HAAVELMO (1947)). In this section we shall give a brief investigation into the application of the methods considered on this subject.

*The probability set*

4. 1. Our probability set $\Gamma$ will be the $n(\nu + 2\tau)$-dimensional Cartesian space $R_{n(\nu+2\tau)}$ with coordinates

$$u_{11}, \dots, u_{1n}, \dots, u_{\nu 1}, \dots, u_{\nu n}$$
$$v_{11}, \dots, v_{1n}, \dots, v_{\tau 1}, \dots, v_{\tau n}$$
$$w_{11}, \dots, w_{1n}, \dots, w_{\tau 1}, \dots, w_{\tau n}.$$

We suppose $n(\nu + 2\tau)$ random variables $\mathfrak{u}_{\lambda i}$, $\mathfrak{v}_{\varkappa i}$, $\mathfrak{w}_{ti}$ $(i = 1, \dots, n;$ $\lambda = 1, \dots, \nu;$ $\varkappa, t = 1, \dots, \tau)$ to have a simultaneous probability distribution on $\Gamma$. Furthermore we consider $n\nu + m\tau$ parameters $\xi_{\lambda i}$, $\alpha_{tj}$ $(i = 1, \dots$ $\dots, n;$ $j = 1, \dots, m;$ $\lambda = 1, \dots, \nu;$ $t = 1, \dots, \tau)$. Finally we consider the following equations

$$f_t(\eta_{1i}, \dots, \eta_{\tau i}, \xi_{1i}, \dots, \xi_{\nu i}, \alpha_{t1}, \dots, \alpha_{tm}) = \mathfrak{w}_{ti} \tag{14}$$
$$\mathfrak{x}_{\lambda i} = \xi_{\lambda i} + \mathfrak{u}_{\lambda i} \tag{15}$$
$$\mathfrak{y}_{\varkappa i} = \eta_{\varkappa i} + \mathfrak{v}_{\varkappa i}. \tag{16}$$

$\left.\begin{matrix} i = 1, \dots, n \\ \lambda = 1, \dots, \nu \\ \varkappa, t = 1, \dots, \tau \end{matrix}\right\}$

The equations (14) are supposed to have a unique solution for $\eta_{\varkappa i}$ $(\varkappa =$ $= 1, \dots, \tau;$ $i = 1, \dots, n)$ on every element of $\Gamma$, except possibly on a set of elements with zero probability.

The equations (14) are called the "stochastic regression equations". The parameters $\xi_{\lambda i}$ $(i = 1, \dots, n)$ are interpreted as the values which the variable $\xi_\lambda$ assumes $(\lambda = 1, \dots, \nu)$. The random variables $\mathfrak{w}_{ti}$ are called

the "true deviations" in the stochastic regression equations. Finally, the random variables $\mathbf{u}_{\lambda i}$ and $\mathbf{v}_{\varkappa i}$ are called the "errors of observation" of the "true" values $\xi_{\lambda i}$ and $\eta_{\varkappa i}$ respectively.

The problem is again, to determine confidence regions for the parameters $\alpha_{ij}$.

*Confidence regions*

4. 2. We reduce the equations (14), (15) and (16) to the forms

$$\mathbf{y}_{\varkappa i} = g_\varkappa(\mathbf{x}_{1i},\ldots,\mathbf{x}_{\nu i},\mathbf{u}_{1i},\ldots,\mathbf{u}_{\nu i},\mathbf{v}_{1i},\ldots,\mathbf{v}_{\tau i},\mathbf{w}_{1i},\ldots,\mathbf{w}_{\tau i},\ \alpha_{11},\ldots,\alpha_{1m},\ldots,\alpha_{\tau 1},\ldots,\alpha_{\tau m}). \qquad (\varkappa = 1,\ldots,\tau)$$

Consider e.g. the case

$$f_t \equiv H_t(\xi_{1i},\ldots,\xi_{\nu i}) + \sum_{\varkappa=1}^{\tau} \beta_{t\varkappa}\,\eta_{\varkappa i}, \qquad (t = 1,\ldots,\tau)$$

in which $\beta_{t\varkappa}$ are real numbers and $H_t$ are polynomials of degree $h$ in the $\xi$'s. Suppose that the errors $\mathbf{u}_{\lambda i}$ are sufficiently small in order that terms containing $\mathbf{u}_{\lambda i}\,\mathbf{u}_{\lambda' i}$ $(\lambda, \lambda' = 1,\ldots,\nu)$ can be neglected (cf. the Approximation of section 3. 1.); then we have

$$H_t(\mathbf{x}_{1i},\ldots,\mathbf{x}_{\nu i}) + \sum_{\varkappa=1}^{\tau} \beta_{t\varkappa}\,\mathbf{y}_{\varkappa i} \approx \mathbf{z}_{ti}, \qquad (t = 1,\ldots,\tau)$$

in which $\mathbf{z}_{ti}$ are linear functions of $\mathbf{u}_{\lambda i}, \mathbf{v}_{\varkappa i}, \mathbf{w}_{ti}$ $(\lambda = 1,\ldots,\nu;\ \varkappa, t = 1,\ldots,\tau)$. The random variables $(\mathbf{z}_{1i},\ldots,\mathbf{z}_{\tau i})$ $(i = 1,\ldots,n)$ have a simultaneous probability distribution, while the $n$ $\tau$-uples $(\mathbf{z}_{1i},\ldots,\mathbf{z}_{\tau i})$ are supposed to be stochastically independent. So we have

$$\mathbf{y}_{\varkappa i} \approx \sum_{t=1}^{\tau} \frac{B_{t\varkappa}}{B}\{-H_t(\mathbf{x}_{1i},\ldots,\mathbf{x}_{\nu i}) + \mathbf{z}_{ti}\}, \qquad (17)$$

in which

$$B = \begin{vmatrix} \beta_{11} & \cdots & \beta_{1\tau} \\ \vdots & & \vdots \\ \beta_{\tau 1} & \cdots & \beta_{\tau\tau} \end{vmatrix}$$

and $B_{t\varkappa}$ is the cofactor of the element $\beta_{t\varkappa}$.

Then the problem is reduced to the case considered in section 3. Call $N$ the number of parameters of a polynomial of degree $h$. Then, in a way and under conditions which are analogous to those stated in section 3, a confidence region for $\tau N$ parameters of the equations (17) can be given. But the original equations contain $\tau(N + \tau - 1)$ parameters. This means that, if $\tau(\tau-1)$ parameters of the original equations are given, a confidence region for the remaining $\tau N$ parameters can be constructed. If the level of significance of the confidence regions for the parameters of the equations (17) are $\varepsilon_\varkappa(\varkappa = 1,\ldots,\tau)$, this level of the confidence region for the $\tau N$ parameters of the original equations is $\leqq \sum_1^{\tau} \varepsilon_\varkappa$.

4. 3. We shall now elaborate a simple example, which is due to T. HAAVELMO (1944), p. 99 seq. Suppose we have the following equations:

$$\left.\begin{aligned} \eta_{1i} - \beta\,\eta_{2i} &= \mathbf{w}_{1i} \\ \alpha\,\xi_{1i} + \eta_{1i} - \alpha\,\eta_{2i} &= \mathbf{w}_{2i} \\ x_{1i} &= \xi_{1i} \\ \mathbf{y}_{1i} &= \eta_{1i} + \mathbf{v}_{1i} \\ \mathbf{y}_{2i} &= \eta_{2i}, \end{aligned}\right\} i = 1, \ldots, n$$

in which $\alpha$ and $-\beta$ are positive.

We obtain:

$$\mathbf{y}_{1i} = \frac{\alpha\beta}{\alpha-\beta}\,x_{1i} + \frac{\alpha\,\mathbf{w}_{1i} - \beta\,\mathbf{w}_{2i}}{\alpha-\beta} + \mathbf{v}_{1i}$$

$$\mathbf{y}_{2i} = \frac{\alpha}{\alpha-\beta}\,x_{1i} + \frac{\mathbf{w}_{1i} - \mathbf{w}_{2i}}{\alpha-\beta}.$$

Suppose that the complete or the incomplete method gives two confidence intervals

$$\mathbf{a}_1 \leqq \frac{\alpha\beta}{\alpha-\beta} \leqq \mathbf{a}_2$$

$$\mathbf{b}_1 \leqq \frac{\alpha}{\alpha-\beta} \leqq \mathbf{b}_2$$

with levels of significance $\varepsilon_1$ and $\varepsilon_2$ respectively. Then we obtain two confidence regions in the $\alpha$, $\beta$-plane, bounded by hyperbolas and by straight lines respectively. The probability that the common part contains the "true" point $(\alpha, \beta)$ is $\geqq 1 - \varepsilon_1 - \varepsilon_2$. (See fig. 2).

*On multicollinearity*

4. 4. As a final application we consider the following case. The following equations are given (cf. section 2. 0.):

$$\left.\begin{aligned} \theta_i &= \alpha_0 + \alpha_1\,\xi_{1i} + \alpha_2\,\xi_{2i} \\ \eta_i &= \theta_i + \mathbf{w}_i \\ \mathbf{x}_{\lambda i} &= \xi_{\lambda i} + \mathbf{u}_{\lambda i} \\ \mathbf{y}_i &= \eta_i + \mathbf{v}_i. \end{aligned}\right\} \begin{matrix} i = 1, \ldots, n \\ \lambda = 1, 2. \end{matrix}$$

Hence

$$\mathbf{y}_i = \alpha_0 + \alpha_1\,\mathbf{x}_{1i} + \alpha_2\,\mathbf{x}_{2i} + \mathbf{z}_i$$

with

$$\mathbf{z}_i = \mathbf{v}_i + \mathbf{w}_i - \alpha_1\,\mathbf{u}_{1i} - \alpha_2\,\mathbf{u}_{2i}.$$

Suppose that the observed values $x_{1i}$, $x_{2i}$ $(i = 1, \ldots, n)$ are such that the following *condition* is satisfied:

For each pair $j$ $(i, j = 1, \ldots, n)$ the quotient

$$\frac{x_{1i} - x_{1j}}{x_{2i} - x_{2j}} \; (i \neq j)$$

has the same sign.

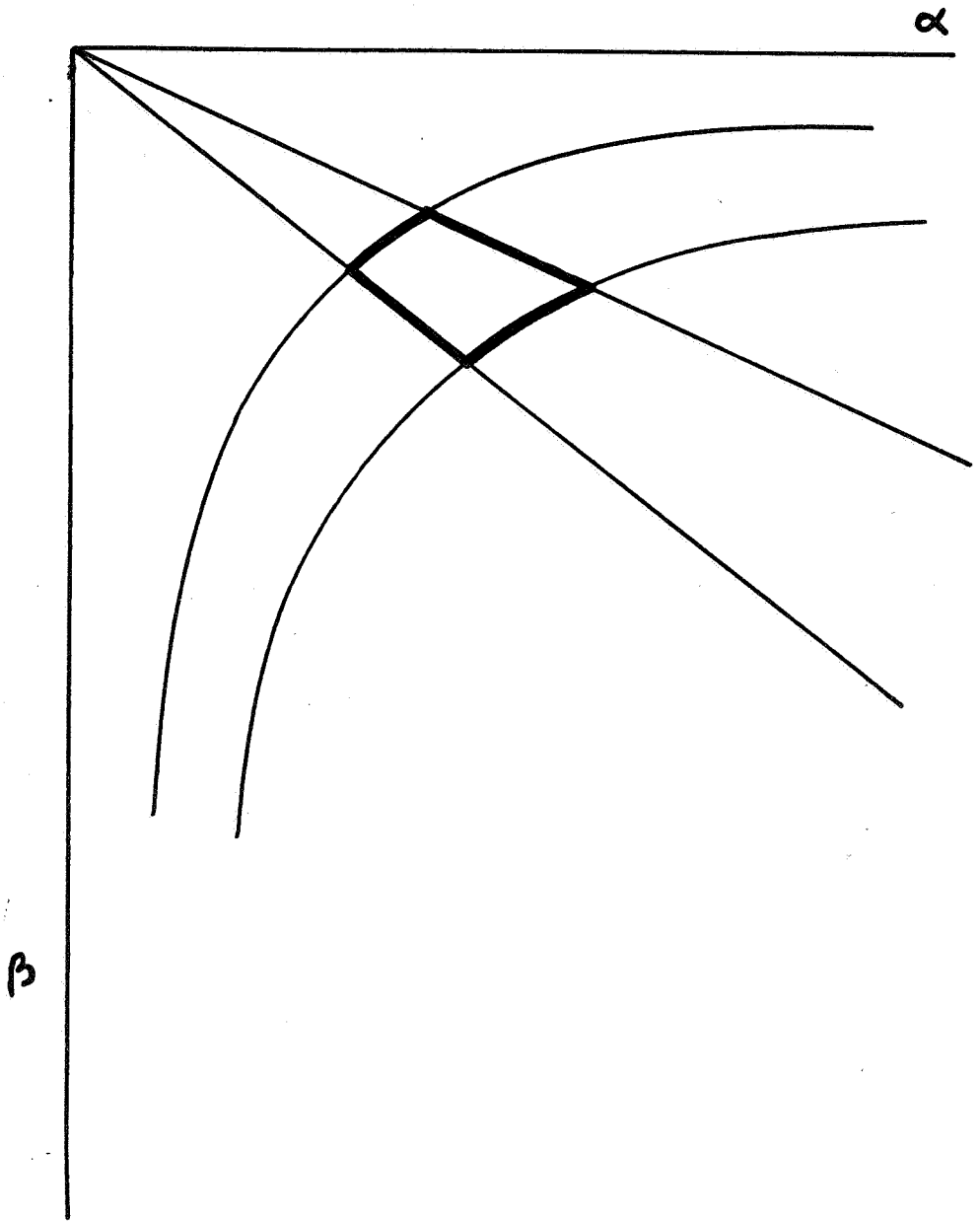


Fig. 2

This condition implies that, apart from the above-mentioned linear relation between $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{y}_i$, we have an additional monotonic relation between the observed values $x_{1i}$ and $x_{2i}$ (if this relation also is — approximately — linear, we have a case of "multicollinearity").

We now have the following

*Theorem* 7. Under the above-mentioned condition the regions $A_1$ and $A_2$ (cf. section 2. 2.) are identical, and their common part $A$ is unbounded.

Proof. If the condition is satisfied the arrangement of the observed points $(x_{1i}, x_{2i}, y_i)$ according to increasing values of $x_1$ is the same as (or just the reverse of) the arrangement according to increasing values of $x_2$. Moreover (cf. section 2. 2.) the quantities $K^{(1)}(ij)$ and $K^{(2)}(ij)$ which are estimates of $\alpha_1$, given $\alpha_2$, and of $\alpha_2$, given $\alpha_1$, respectively are represented by the same set of straight lines in the $\alpha_1$, $\alpha_2$-plane:

$$(x_{1i}-x_{1j})\,K^{(1)}(ij)+(x_{2i}-x_{2j})\,\alpha_2=y_i-y_j$$
$$(x_{1i}-x_{1j})\,\alpha\ +(x_{2i}-x_{2j})\,K^{(2)}(ij)=y_i-y_j.$$

As the slopes of these straight lines $-(x_{1i}-x_{1j})/(x_{2i}-x_{2j})$ have the same sign, the regions $A_1$ and $A_2$ are identical, from which the theorem follows.

If the incomplete method instead of the complete method is used, the same theorem holds with respect to the regions $A_1'$ and $A_2'$, whereas the condition that all quantities $(x_{1i}-x_{1j})/(x_{2i}-x_{2j})$ have the same sign is weakened to the condition that all quantities $(x_{1i}-x_{1,n_1+i})/(x_{2i}-x_{2,n_1+i})$ have the same sign $(i=1,\ldots,n_1)$.

## 5. PROBLEMS OF PREDICTION

*The probability set*

5. 0. For the probability set and the random variables defined on it we refer to section 4. 1. We assume, however, that all errors $\mathbf{u}_{\lambda i}$, $\mathbf{v}_{\varkappa i}$ are identically equal to zero $(\lambda=1,\ldots,\nu;\ \varkappa=1,\ldots,\tau;\ i=1,\ldots,n)$.

*Conditions*

5. 1. We impose the following conditions:

*Condition I*: All $n$ $\tau$-uples $(\mathbf{w}_{1i},\ldots,\mathbf{w}_{\tau i})$ are distributed independently of each other.

*Condition IIIa*: All $n$ $\tau$-uples $(\mathbf{w}_{1i},\ldots,\mathbf{w}_{\tau i})$ have the same continuous simultaneous distribution function.

Apart from these conditions we shall use the additional conditions, which are necessary for the determination of a confidence region for the parameters of the regression equations.

*The problem*

5. 2. Suppose that the following $n$ points are observed:

$$(\xi_{1i},\ldots,\xi_{\nu i},\eta_{1i},\ldots,\eta_{\tau i}).$$

Suppose further that the following $\nu$ parameters are given:

$$\xi_{1,n+1},\ldots,\xi_{\nu,n+1}.$$

These parameters are interpreted as the $\xi$-coordinates of an $(n+1)$-th point, which is not observed. The problem is to determine a confidence region for the $\eta$-coordinates of this point, i.e. for

$$\eta_{1,n+1},\ldots,\eta_{\nu,n+1}.$$

*Confidence regions*

5. 3. Consider again the case (cf. section 4. 2.):

$$f_t \equiv H_t(\xi_{1i}, \dots, \xi_{\nu i}) + \sum_{t=1}^{\tau} \beta_{t\varkappa}\, \eta_{\varkappa i}, \qquad (t = 1, \dots, \tau)$$

so that we have

$$\eta_{\varkappa i} = \sum_{t=1}^{\tau} \frac{B_{t\varkappa}}{B} \{-H_t(\xi_{1i}, \dots, \xi_{\nu i}) + \mathfrak{w}_{ti}\}. \qquad (\varkappa = 1, \dots, \tau;\ i = 1, \dots, n)$$

Putting $i = n + 1$ we can write

$$\eta_{\varkappa, n+1} = g_\varkappa(\xi_{1,n+1}, \dots, \xi_{\nu,n+1}, \beta) + h_\varkappa(\mathfrak{w}_{1,n+1}, \dots, \mathfrak{w}_{\tau,n+1}, \beta),$$

in which

$$g_\varkappa(\xi_{1,n+1}, \dots, \xi_{\nu,n+1}, \beta) = -\sum_{t=1}^{\tau} \frac{B_{t\varkappa}}{B} H_t(\xi_{1,n+1}, \dots, \xi_{\nu,n+1})$$

$$h_\varkappa(\mathfrak{w}_{1,n+1}, \dots, \mathfrak{w}_{\tau,n+1}, \beta) = \sum_{t=1}^{\tau} \frac{B_{t\varkappa}}{B} \mathfrak{w}_{ti}$$

and in which $\beta$ is the "true parameter point"; $\beta$ may be considered as a vector, the components of which are $\beta_{t\varkappa}(t, \varkappa = 1, \dots, \tau)$ and all parameters determining the polynomials $H_t\,(t = 1, \dots, \tau)$.

Suppose $\beta$ is known. Then we can arrange the $n$ quantities $h_\varkappa(w_{1i}, \dots, w_{\tau i}, \beta)$ according to increasing magnitude:

$$h_{\varkappa 1} < \dots < h_{\varkappa n}, \qquad (\varkappa = 1, \dots, \tau)$$

in which

$$h_{\varkappa j} = h_\varkappa(w_{1i_j}, \dots, w_{\tau i_j}, \beta).$$

We have the following

*Theorem* 8: Under conditions I and III*a* a confidence interval for $\eta_{\varkappa, n+1}$ is given by

$$\left(g_\varkappa(\xi_{1,n+1}, \dots, \xi_{\nu,n+1}, \beta) + \mathfrak{h}_{\varkappa s},\ g_\varkappa(\xi_{1,n+1}, \dots, \xi_{\nu,n+1}, \beta) + \mathfrak{h}_{\varkappa, n-s+1}\right)$$

if $\beta$ is the known "true parameter point"; the level of significance is $2s\,(n + 1)^{-1}$.

In order to prove this theorem, we shall use the following lemma (see W. R. THOMPSON (1936)):

*Lemma*: If a random sample of size $n$ is drawn from a universe with continuous distribution function; if the sample values are arranged in ascending order; if an $(n + 1)$-th draw from the same universe is to be effected; then the probability that the stochastic interval bounded by the $s$-th and the $(n - s + 1)$-th of these values will contain the $(n + 1)$-th is equal to $1 - 2s/(n + 1)$.

*Proof of Theorem* 8: As $g_\varkappa(\xi_{1,n+1}, \dots, \xi_{\nu,n+1}, \beta)$ is ex hypothesi a known quantity, the problem is to determine a confidence interval for

$h_\varkappa(\mathbf{w}_{1,n+1}, \ldots, \mathbf{w}_{\tau,n+1}, \beta)$. But $n$ sample values $h_\varkappa(\mathbf{w}_{1i}, \ldots, \mathbf{w}_{\tau i}, \beta)$ from the same universe (cf. condition III$a$) are obtained; hence the lemma is sufficient in order to show the validity of the theorem.

5. 4. Generally, however, $\beta$ is unknown, and we can only calculate a confidence region $\mathbf{R}$ for $\beta$. Let now $\beta$ vary through $\mathbf{R}$, and denote by $\mathbf{J}_\varkappa$ the interval bounded by the lowest of all lower limits of the interval considered in Theorem 8 and by the highest of all upper limits ($\varkappa = 1, \ldots, \tau$). If the level of significance of $\mathbf{R}$ is $\varepsilon$, the following theorem immediately follows:

*Theorem* 9:

$$P[\eta_{1,n+1} \in \mathbf{J}_1, \ldots, \eta_{\tau,n+1} \in \mathbf{J}_\tau, \beta \in \mathbf{R}] \geqq (1-\varepsilon)\left(1-\frac{2\tau s}{n+1}\right).$$

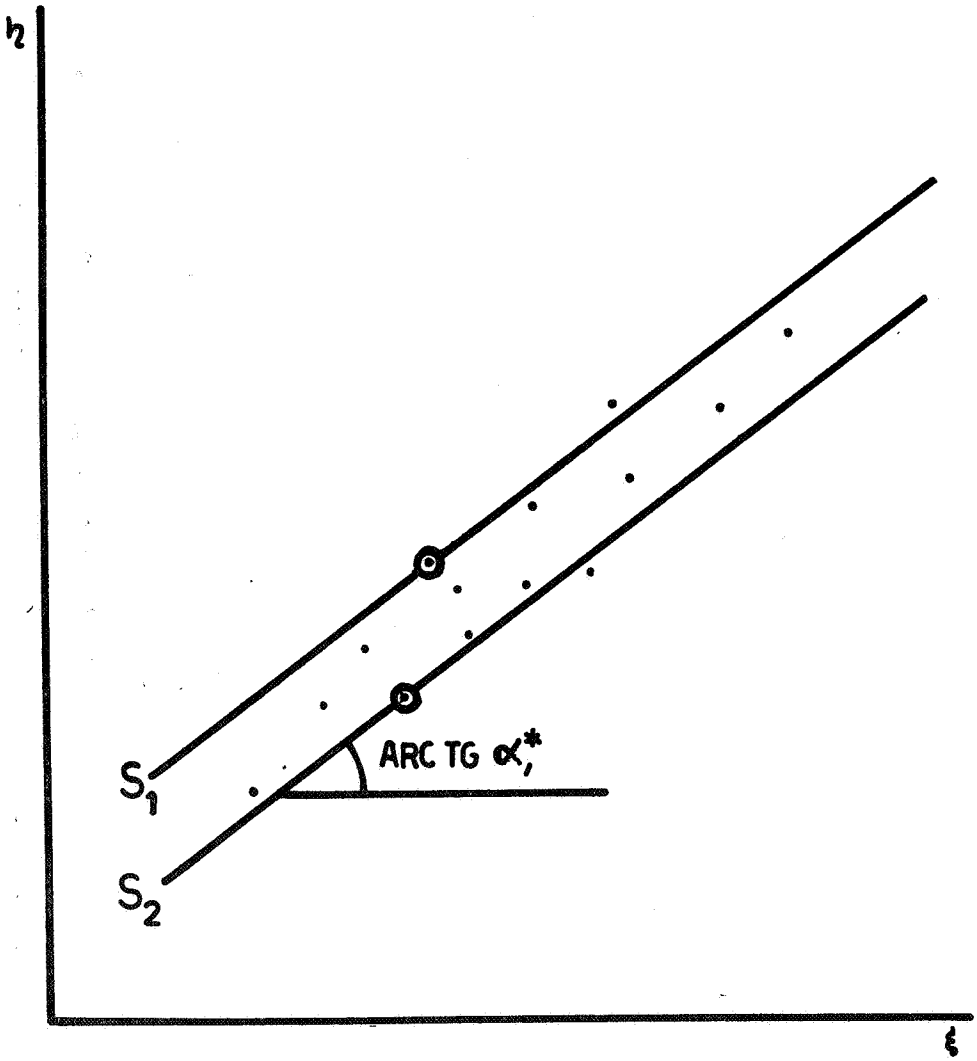


Fig. 3. $n = 14 \quad s = 2$

*The linear case in two variables*

5. 5. For the linear case of two variables

$$\eta_i = \alpha_0 + \alpha_1 \xi_i + w_i$$

a simple graphical representation can be given. Suppose that $\alpha_1^*$ is the "true" $\alpha_1$; after arranging the sample values

$$w_i(\alpha_1^*) = \eta_i - \alpha_0 - \alpha_1^* \xi_i$$

in order we find two straight lines:

$S_1$: $$\eta = \alpha_0 + \alpha_1^* \xi + w_s(\alpha^*)$$

and $S_2$: $$\eta = \alpha_0 + \alpha_1^* \xi + w_{n-s+1}(\alpha^*).$$

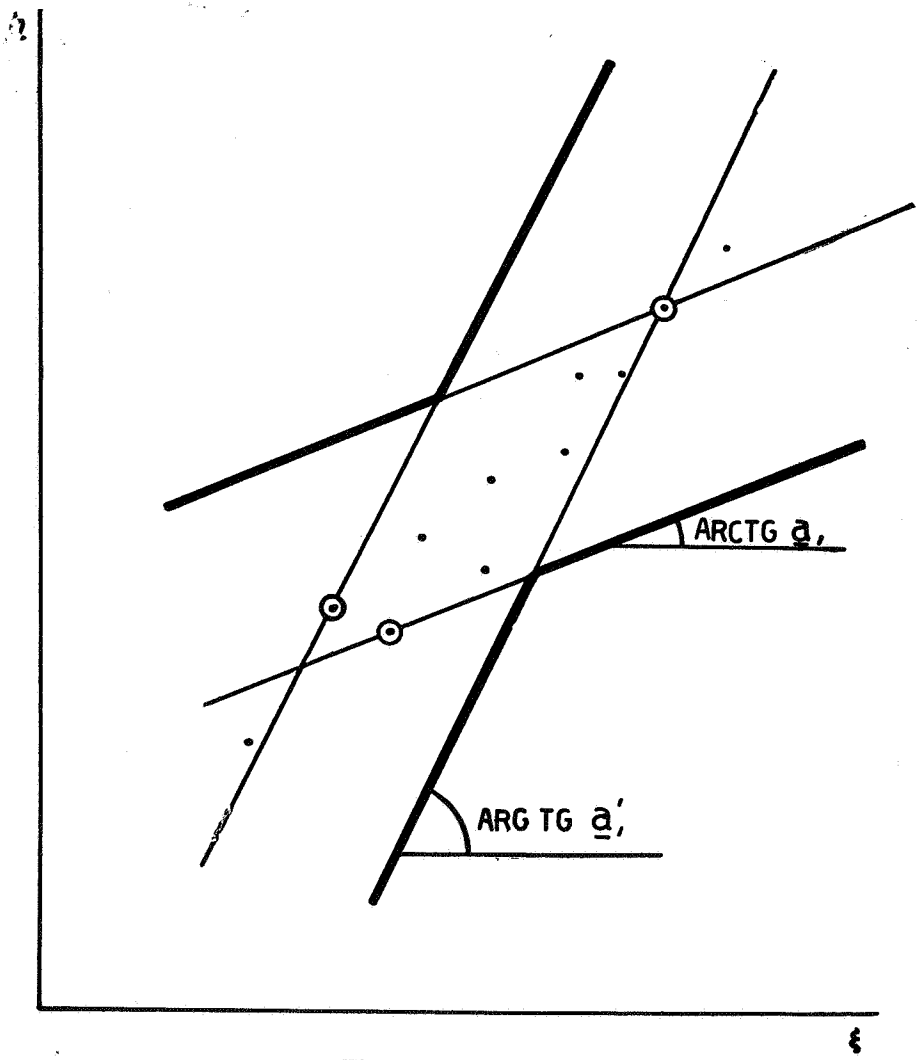


Fig. 4. $n = 11 \quad s = 2$

The probability that the region bounded above by $S_1$ and below by $S_2$ will contain an $(n+1)$-th sample point $D$ is, under the condition that $\alpha_1^* = \alpha_1$, equal to $1 - 2s/(n+1)$. (See fig. 3).

Suppose that the confidence interval for $\alpha_1$ is $(a_1, a_1')$. When $\alpha_1^*$ varies through this interval the lines $S_1$ and $S_2$ revolve around the observed points $(\xi_i, \eta_i)$ with ranks $s$ and $(n - s + 1)$ respectively with respect to increasing values of $w$. As long as the observed points having these properties remain the same, $S_1$ and $S_2$ revolve around one point; but as soon as variation of $\alpha_1^*$ causes another point to have this property, the revolution takes place around this point. The figures 4 and 5 elucidate the fact that sometimes the region is bounded by the straight lines $S_1$ and $S_2$ for $\alpha^* = a_1$ and $a_1'$ only, whereas it is sometimes necessary to consider values between $a_1$ and $a_1'$ as well.

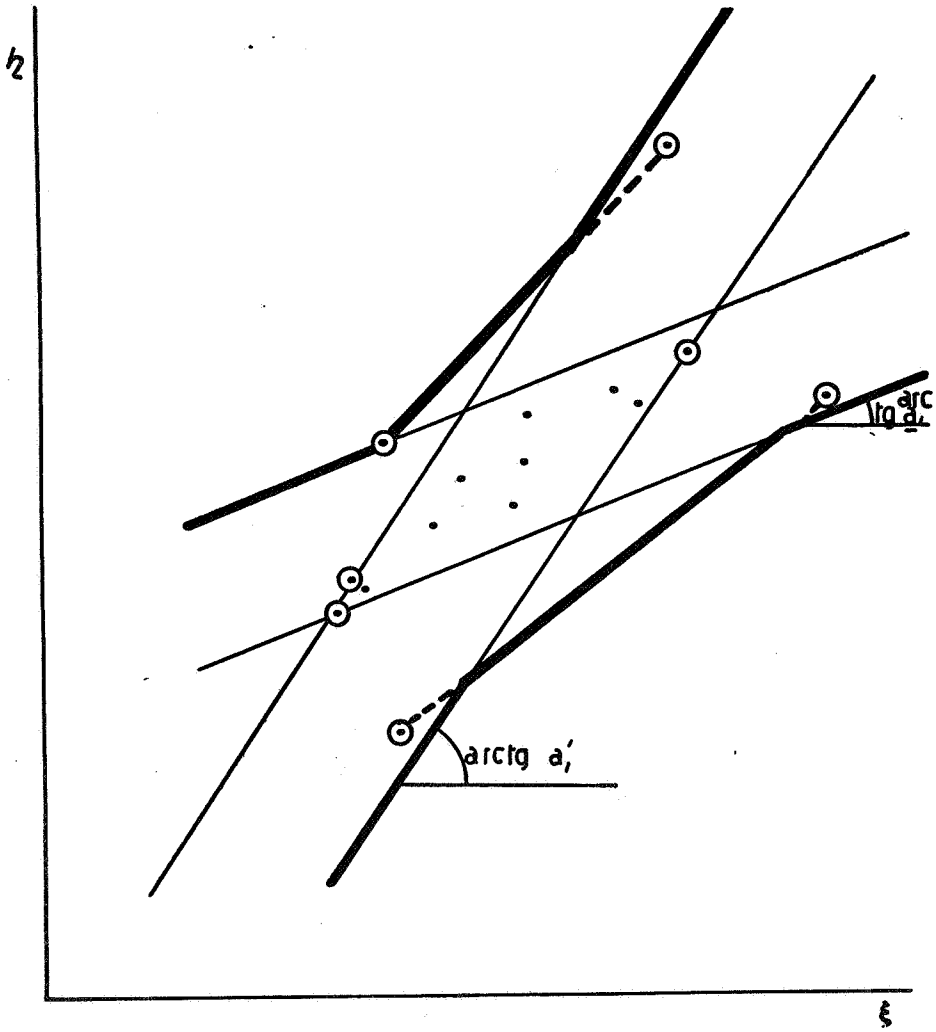


Fig. 5. $n = 15$ $s = 2$

## 6. CONCLUDING REMARKS

6. 0. The methods of determining confidence regions which may be derived from this kind of analysis have not been exhaustively treated. In order to elucidate this statement we shall give a confidence interval for $\alpha$ in the stochastic regression equation

$$\eta_i = A\,\xi_i^{\alpha} + \mathbf{w}_i, \qquad (i = 1, \ldots, n)$$

in which all $\xi_i$ are positive, and in which

$$P\,[A\,\eta_i \leqq 0] = 0$$

holds for $i = 1, \ldots, n$. $\xi_1, \ldots, \xi_n$ are known, $A$ and $\alpha$ are unknown parameters, and $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are random variables, which are supposed (1) to be distributed stochastically independent, (2) to have continuous symmetrical distribution functions with zero median.

We arrange the observed points $(\xi_i, \eta_i)$ according to increasing magnitude of $\xi$ and define

$$\mathbf{d}_i = \frac{\log|\eta_{n_1+i}| - \log|\eta_i|}{\log \xi_{n_1+i} - \log \xi_i}, \qquad (i = 1, \ldots, n_1)$$

in which $n_1 = \frac{1}{2}n$ (if $n$ is odd the point $(\xi_{\frac{1}{2}(n+1)}, \eta_{\frac{1}{2}(n+1)})$ is neglected). After arranging the observed quantities $d_i$ according to increasing magnitude:

$$d_{(1)} < \ldots < d_{(n_1)}$$

we have the following

*Theorem* 10. Under conditions (1) and (2) the interval $(\mathbf{d}_{(r_1)}, \mathbf{d}_{(n_1 - r_1 + 1)})$ is a confidence interval for $\alpha$ to the level of significance $2I_{\frac{1}{2}}(n_1 - r_1 + 1, r_1)$.

Proof. We have

$$\frac{\eta_{n_1+i} - \mathbf{w}_{n_1+i}}{\eta_i - \mathbf{w}_i} = \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^{\alpha}$$

or:

$$\eta_{n_1+i} - \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^{\alpha} \eta_i = \mathbf{w}_{n_1+i} - \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^{\alpha} \mathbf{w}_i.$$

It follows from condition (2), that

$$\mathbf{w}_{n_1+i} - \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^{\alpha} \mathbf{w}_i$$

has a continuous distribution function with zero median. Hence:

$$P\left[\eta_{n_1+i} - \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^{\alpha} \eta_i < 0\right] =$$

$$= P\,[\log \eta_{n_1+i} - \log \eta_i < \alpha\,(\log \xi_{n_1+i} - \log \xi_i)] =$$

$$= P\left[\frac{\log \eta_{n_1+i} - \log \eta_i}{\log \xi_{n_1+i} - \log \xi_i} < \alpha\right] = P\left[\frac{\log \eta_{n_1+i} - \log \eta_i}{\log \xi_{n_1+i} - \log \xi_i} > \alpha\right] = \tfrac{1}{2},$$

if $\eta_1, \ldots, \eta_n$ are positive; if they are negative we have to replace $\eta_i$ and $\eta_{n_1+i}$ by $-\eta_i$ and $-\eta_{n_1+i}$ respectively. From this and from condition (1) the theorem follows.

The theorem shows that this method of determining a confidence interval for $\alpha$ is identical with the incomplete method for $\alpha$ in the linear equation

$$\log \eta_i = \log A + \alpha \log \xi_i + w_i'$$

(which can be written as

$$\eta_i = A\, \xi_i^{\alpha}\, e^{w_i'}),$$

if $w_1', \ldots, w_n'$ satisfy the same conditions (1) and (2).

6. 1. Finally we mention that it is possible to find estimates instead of confidence intervals. Consider e.g. the statistics $\Delta(ij)$; each of these $\binom{n}{2}$ statistics has the property that its sampling median is equal to $\alpha_1$ (cf. section 1. 3.). Hence one can use the sample median of the observed quantities $\Delta(ij)$ as an estimate of $\alpha_1$.

It is a pleasure to acknowledge my indebtedness to Professor Dr D. VAN DANTZIG for his stimulating interest and to Mr J. HEMELRIJK for his valuable and constructive criticism.

## REFERENCES

BENTZEL, R. and H. WOLD, On statistical demand analysis from the viewpoint of simultaneous equations. *Skand. Aktuarietidskr.* **29**, 95—114 (1946).

GIRSHICK, M. A. and T. HAAVELMO, Statistical analysis of the demand for food: examples of simultaneous estimation of structural equations. *Econometrica*, **15**, 79—110 (1947).

HAAVELMO, T., The statistical implications of a system of simultaneous equations. *Econometrica*, **11**, 1—12 (1943).

————, The probability approach in econometrics. *Econometrica*, **12**, suppl. (1944).

KOOPMANS, T., Statistical estimation of simultaneous economic relations. *Journal of the Amer. Statist. Assoc.*, **40**, 448—466 (1945).

————, (ed.), *Statistical inference in dynamic economic models* (New York, 1950).

THOMPSON, W. R., On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *Annals of Math. Statist.*, **7**, 122—128 (1936).

*Publication of the Statistical Department of the "Mathematisch Centrum", Amsterdam*