KONINKLIJKE NEDERLANDSE AKADEMIE
VAN WETENSCHAPPEN

A rank-invariant method of linear and polynomial regression analysis

III

BY

H. THEIL

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

BY

H. THEIL

3. CONFIDENCE REGIONS FOR THE PARAMETERS OF POLYNOMIAL REGRESSION EQUATIONS

The probability set

3. 0.   The probability set Γ 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 $u_{\lambda i}$, v_i, w_i $(\lambda = 1, \ldots, \nu;$ $i = 1, \ldots, n)$ to have a simultaneous probability distribution on Γ. 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])

(10) $$\theta_i = \sum \alpha_{p_1 \ldots p_\nu} \xi_{1i}^{p_1} \ldots \xi_{\nu i}^{p_\nu}$$

(11) $$\eta_i = \theta_i + w_i \qquad \left\{ \begin{array}{l} i = 1, \ldots, n \\ \lambda = 1, \ldots, \nu \end{array} \right.$$

(12) $$x_{\lambda i} = \xi_{\lambda i} + u_{\lambda i}$$

(13) $$y_i = \eta_i + v_i.$$

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

The parameters $\xi_{\lambda i}$ $(i = 1, \ldots, n)$ are interpreted as values assumed by the variable ξ_λ. The equation (10) is the polynomial regression equation. The random variables 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]) Σ in equation (10) denotes summation over all sets $p_1, \ldots, p_\nu$.

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

Conditions; approximation

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

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

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

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 $\boldsymbol{u}_{\lambda i}$ vanishes outside a finite interval $|\boldsymbol{u}_{\lambda i}| < g_{\lambda i}$.

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

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

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

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

Approximation: For any positive s the quantities

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

are neglected. [4]

Confidence regions

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

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

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 $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ cannot be separated in one sample of observations; if, however, the experiment is repeated for the same "true" values $\xi_{\lambda i}$, η_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 $\boldsymbol{v}_i$ can be mitigated by averaging, whereas the deviations $\boldsymbol{w}_i$ cannot.

[4]) The approximation implies that the errors $\boldsymbol{u}_{\lambda i}$ are sufficiently small. This restriction is not very serious, because, unless the number of points n is very large, large v·lues of $\boldsymbol{u}_{\lambda i}$ 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:

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

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

$$\vdots$$

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

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

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

$$\varDelta_1^{(h)} < \ldots < \varDelta_{n_h}^{(h)},$$

in which

$$\varDelta_j^{(h)} = \varDelta^{(h)}\,(i_j, \ldots, hn_h + i_j).$$

3. 3.   Then we have the following theorem:

Theorem 6: Under conditions I, IIa, and III the interval $(\varDelta_{r_h}^{(h)}, \varDelta_{n_h - r_h + 1}^{(h)})$ 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

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

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

Then we have

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

Proof of the lemma: We have

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

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

---

[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

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

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

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

Proof of Theorem 6: The relation between x_i and y_i is given by

$$y_i = \sum_{p=0}^h a_p (x_i - u_i)^p + v_i + w_i$$

$$\approx \sum_{p=0}^h a_p x^p_i - u_i (a_1 + 2 a_2 \xi_i + \ldots + h a_h \xi^{h-1}_i) + v_i + w_i,$$

in which we neglected (in accordance with the Approximation) u^s_i for $s > 1$. Putting $z_i = \varrho_i u_i + v_i + w_i$, in which

$$-\varrho_i = a_1 + 2 a_2 \xi_i + \ldots + h a_h \xi^{h-1}_i,$$

we get

$$y_i \approx \sum_{p=0}^h a_p x^p_i + z_i.$$

Now we have according to the lemma:

$$\Delta(i, n_h+i) \approx a_1 + a_2 \frac{x^2_i - x^2_{n_h+i}}{x_i - x_{n_h+i}} + \ldots + a_h \frac{x^h_i - x^h_{n_h+i}}{x_i - x_{n_h+i}} + \frac{z_i - z_{n_h+i}}{x_i - x_{n_h+i}}$$

$$= a_1 + a_2 P^1_{i,n_h+i} + \ldots + a_h P^{h-1}_{i,n_h+i} + \frac{z_i - z_{n_h+i}}{x_i - x_{n_h+i}}.$$

$$\Delta^{(2)}(i, n_h+i, 2 n_h+i) \approx a_2 + a_3 P^1_{i,n_h+i,2n_h+i} + \ldots + a_h P^{h-2}_{i,n_h+i,2n_h+i} +$$

$$+ \frac{\dfrac{z_i - z_{n_h+i}}{x_i - x_{n_h+i}} - \dfrac{z_{n_h+i} - z_{2n_h+i}}{x_{n_h+i} - x_{2n_h+i}}}{x_i - x_{2n_h+i}}.$$

$$\Delta^{(h)}(i, n_h+i, \ldots, hn_h+i) \approx a_h + Z_i,$$

in which Z_i is a random variable depending on

$$x_i, \ldots, x_{hn_h+i} \quad , \quad z_i, \ldots, z_{hn_h+i}.$$

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

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

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

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

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

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

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 ε_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 IIb is valid, which is a more stringent condition than condition IIa. Consider the equation

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

in which $\varSigma$ 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

$$P\left[\boldsymbol{S}_{ij} < a_{p_1^1\ldots p_\nu^1} \mid a_{p_1^1\ldots p_\nu^1}\right] = P\left[\boldsymbol{S}_{ij} > a_{p_1^1\ldots p_\nu^1} \mid a_{p_1^1\ldots p_\nu^1}\right] = \tfrac{1}{2},$$

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

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

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 $\varGamma$ will be the $n(\nu + 2\tau)$-dimensional Cartesian space $R_{n(\nu+2\tau)}$ with coordinates

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

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

(14) $\qquad f_t\left(\eta_{1i}, \ldots, \eta_{\tau i}, \xi_{1i}, \ldots, \xi_{\nu i}, a_{t1}, \ldots, a_{tm}\right) = \boldsymbol{w}_{ti}$ $\left.\begin{array}{l} \\ \\ \\ \end{array}\right\}$ $\left\{\begin{array}{l} i = 1, \ldots, n \\ \lambda = 1, \ldots, \nu \\ \varkappa, t = 1, \ldots, \tau \end{array}\right.$

(15) $\qquad\qquad\qquad \boldsymbol{x}_{\lambda i} = \xi_{\lambda i} + \boldsymbol{u}_{\lambda i}$

(16) $\qquad\qquad\qquad \boldsymbol{y}_{\varkappa i} = \eta_{\varkappa i} + \boldsymbol{v}_{\varkappa i}.$

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

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

the "true deviations" in the stochastic regression equations. Finally, the random variables $u_{\lambda i}$ and $v_{\varkappa i}$ 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 a_{tj}.

Confidence regions

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

$$y_{\varkappa i} = g_{\varkappa}(x_{1i},\ldots,x_{\nu i}, u_{1i},\ldots,u_{\nu i}, v_{1i},\ldots,v_{\tau i}, w_{1i},\ldots,w_{\tau i},$$
$$a_{11},\ldots,a_{1m},\ldots,a_{\tau 1},\ldots,a_{\tau m}). \qquad (\varkappa = 1,\ldots,\tau)$$

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 ξ's. Suppose that the errors $u_{\lambda i}$ are sufficiently small in order that terms containing $u_{\lambda i} u_{\lambda' i}$ ($\lambda, \lambda' = 1,\ldots,\nu$) can be neglected (cf. the Approximation of section 3. 1.); then we have

$$H_t(x_{1i},\ldots,x_{\nu i}) + \sum_{\varkappa=1}^{\tau} \beta_{t\varkappa} y_{\varkappa i} \approx z_{ti}, \qquad (t = 1,\ldots,\tau)$$

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

(17)
$$y_{\varkappa i} \approx \sum_{t=1}^{\tau} \frac{B_{t\varkappa}}{B}\{-H_t(x_{1i},\ldots,x_{\nu i}) + z_{ti}\},$$

in which

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

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 τN 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 τN 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 τN parameters of the original equations is $\leq \sum_{1}^{\tau} \varepsilon_{\varkappa}$.

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:

$$\left.\begin{aligned}
\eta_{1i} - \beta\,\eta_{2i} &= w_{1i} \\
\alpha\,\xi_{1i} + \eta_{1i} - \alpha\,\eta_{2i} &= w_{2i} \\
x_{1i} &= \xi_{1i} \\
y_{1i} &= \eta_{1i} + v_{1i} \\
y_{2i} &= \eta_{2i}\,,
\end{aligned}\right\} i = 1,\ldots,n$$

in which α and $-\beta$ are positive.

We obtain:

$$y_{1i} = \frac{\alpha\beta}{\alpha-\beta}\,x_{1i} + \frac{\alpha\,w_{1i} - \beta\,w_{2i}}{\alpha-\beta} + v_{1i}$$

$$y_{2i} = \frac{\alpha}{\alpha-\beta}\,x_{1i} + \frac{w_{1i} - w_{2i}}{\alpha-\beta}\,.$$

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

$$a_1 \leqq \frac{\alpha\beta}{\alpha-\beta} \leqq a_2$$

$$b_1 \leqq \frac{\alpha}{\alpha-\beta} \leqq b_2$$

with levels of significance ε_1 and ε_2 respectively. Then we obtain two confidence regions in the α, β-plane, bounded by hyperbolas and by straight lines respectively. The probability that the common part contains the "true" point (α, β) 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.):

$$\left.\begin{aligned}
\theta_i &= a_0 + a_1\,\xi_{1i} + a_2\,\xi_{2i} \\
\eta_i &= \theta_i + w_i \\
x_{\lambda i} &= \xi_{\lambda i} + u_{\lambda i} \\
y_i &= \eta_i + v_i\,.
\end{aligned}\right\} \left\{\begin{aligned} i &= 1,\ldots,n \\ \lambda &= 1, 2. \end{aligned}\right.$$

Hence

$$y_i = a_0 + a_1\,x_{1i} + a_2\,x_{2i} + z_i$$

with

$$z_i = v_i + w_i - a_1\,u_{1i} - a_2\,u_{2i}\,.$$

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

$$\frac{x_{1i} - x_{1j}}{x_{2i} - x_{2j}}\ (i \neq j)$$
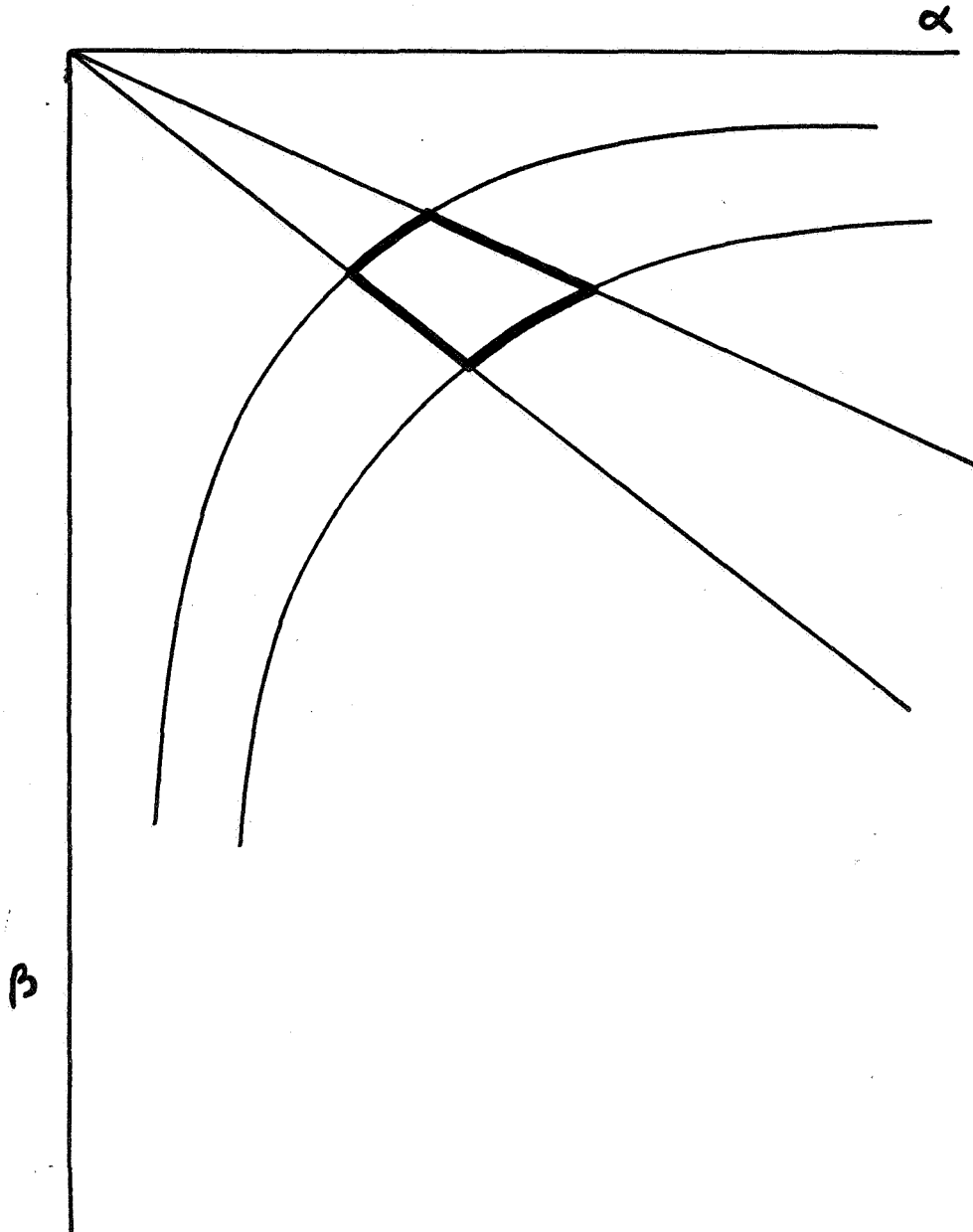
has the same sign.

α

β

Fig. 2

This condition implies that, apart from the above-mentioned linear relation between x_{1i}, x_{2i} and 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 a_1, given a_2, and of a_2, given a_1, respectively are represented by the same set of straight lines in the a_1, a_2-plane:

$$(x_{1i} - x_{1j}) K^{(1)}(ij) + (x_{2i} - x_{2j}) a_2 = y_i - y_j$$

$$(x_{1i} - x_{1j}) a + (x_{2i} - x_{2j}) K^{(2)}(ij) = y_i - y_j.$$

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 $u_{\lambda i}$, $v_{\varkappa i}$ 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 τ-uples $(w_{1i}, \ldots, w_{\tau i})$ are distributed independently of each other.

Condition IIIa: All n τ-uples $(w_{1i}, \ldots, w_{\tau i})$ 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 ν parameters are given:

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

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

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

12

Confidence regions

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

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

so that we have

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

Putting $i = n + 1$ we can write

$$\eta_{\varkappa, n+1} = g_\varkappa\,(\xi_{1, n+1}, \ldots, \xi_{\nu, n+1}, \beta) + h_\varkappa\,(w_{1, n+1}, \ldots, w_{\tau, n+1}, \beta),$$

in which

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

$$h_\varkappa\,(w_{1, n+1}, \ldots, w_{\tau, n+1}, \beta) = \sum_{t=1}^{\tau} \frac{B_{t\varkappa}}{B}\,w_{ti}$$

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

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

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

in which

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

We have the following

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

$$\Big(g_\varkappa\,(\xi_{1, n+1}, \ldots, \xi_{\nu, n+1}, \beta) + h_{\varkappa s},\; g_\varkappa\,(\xi_{1, n+1}, \ldots, \xi_{\nu, n+1}, \beta) + h_{\varkappa, n-s+1}\Big)$$

if β 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 $g_\varkappa\,(\xi_{1, n+1}, \ldots, \xi_{\nu, n+1}, \beta)$ is ex hypothesi a known quantity, the problem is to determine a confidence interval for

13

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

5. 4.   Generally, however, β is unknown, and we can only calculate a confidence region **R** for β. Let now β vary through **R**, and denote by $J_\varkappa$ 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 **R** is ε, the following theorem immediately follows:

Theorem 9:

$$P\left[\eta_{1,n+1} \in J_1, \ldots, \eta_{\tau,n+1} \in J_\tau, \beta \in R\right] \geq (1-\varepsilon)\left(1 - \frac{2\tau s}{n+1}\right).$$
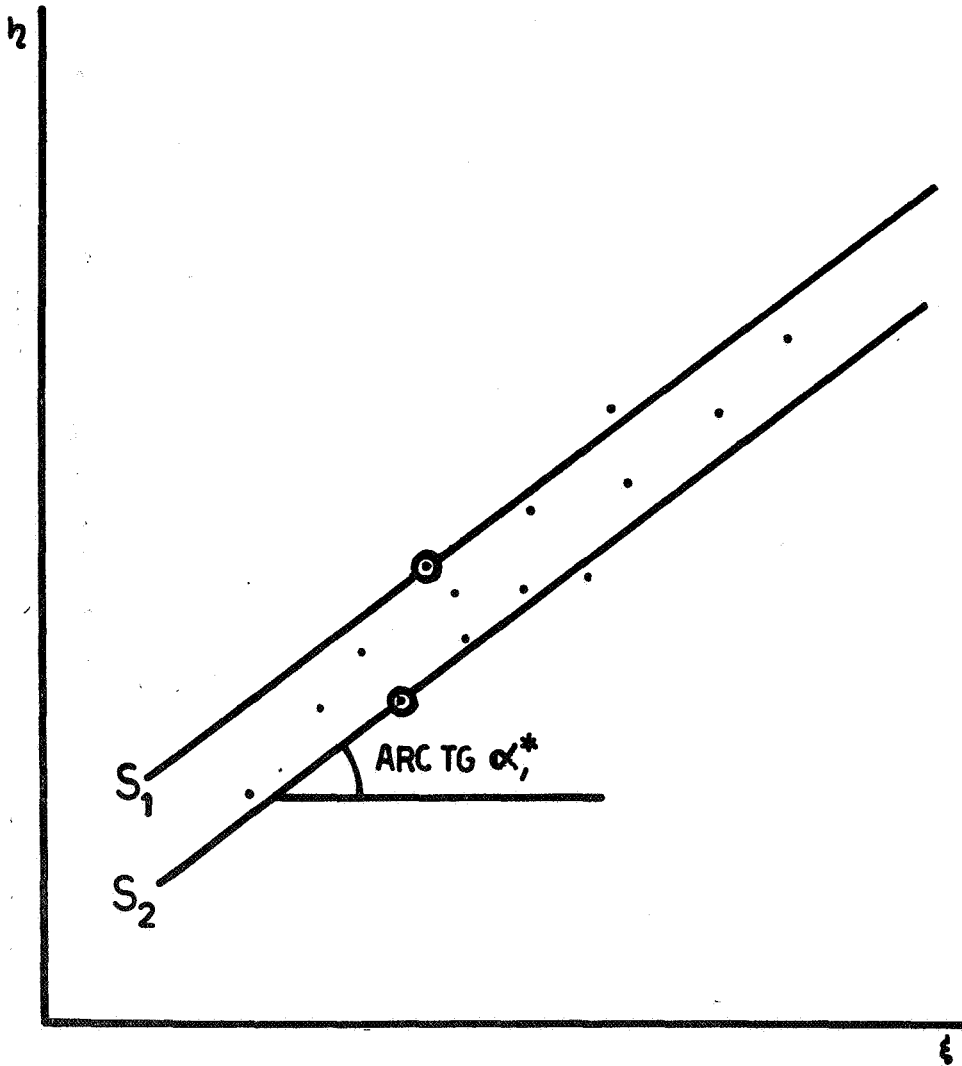

Fig. 3. $n = 14$ $s = 2$

14

The linear case in two variables

5. 5.   For the linear case of two variables

$$\eta_i = a_0 + a_1\,\xi_i + \mathbf{w}_i$$

a simple graphical representation can be given. Suppose that a_1^* is the "true" a_1; after arranging the sample values

$$w_i\,(a_1^*) = \eta_i - a_0 - a_1^*\,\xi_i$$

in order we find two straight lines:

S_1: $\qquad\qquad\qquad \eta = a_0 + a_1^*\,\xi + w_s\,(a^*)$

and S_2: $\qquad\qquad \eta = a_0 + a_1^*\,\xi + w_{n-s+1}\,(a^*).$
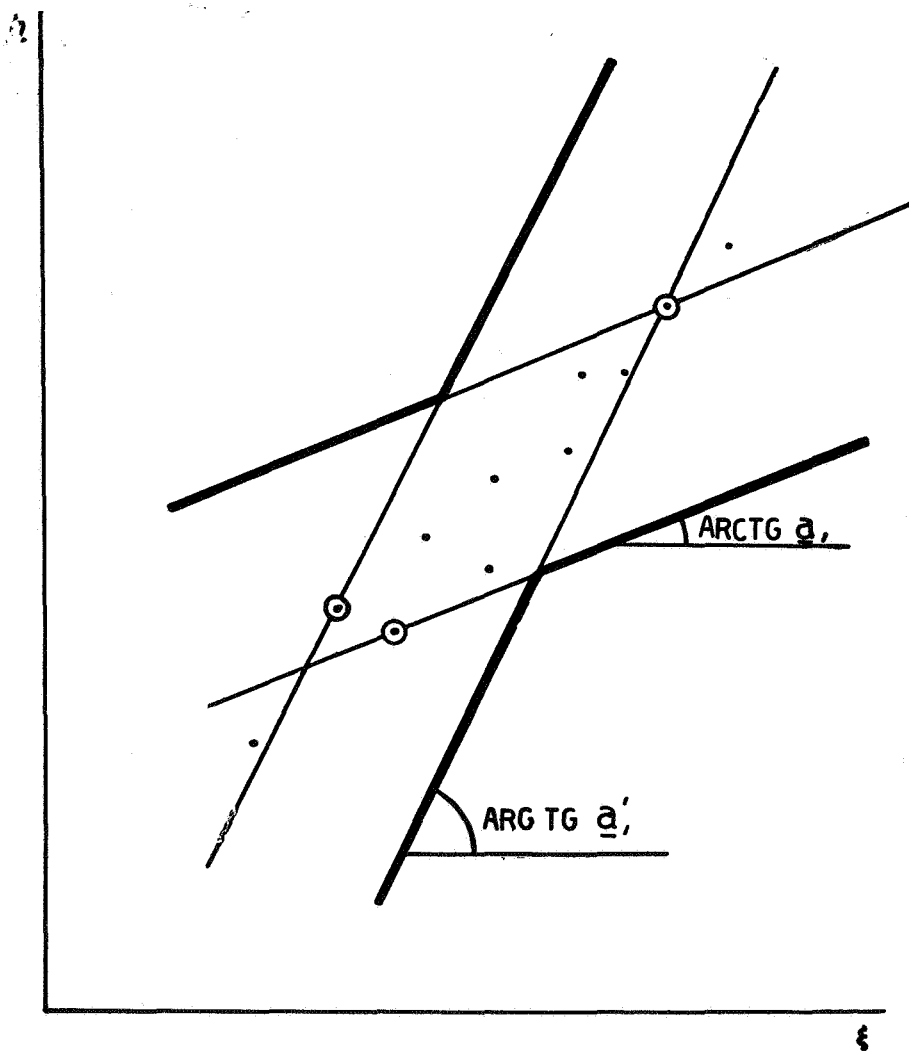

Fig. 4. $n = 11$ $s = 2$

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 α_1 is (a_1, a_1'). When α_1^* varies through this interval the lines S_1 and S_2 revolve around the observed points (ξ_i, η_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 α_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.
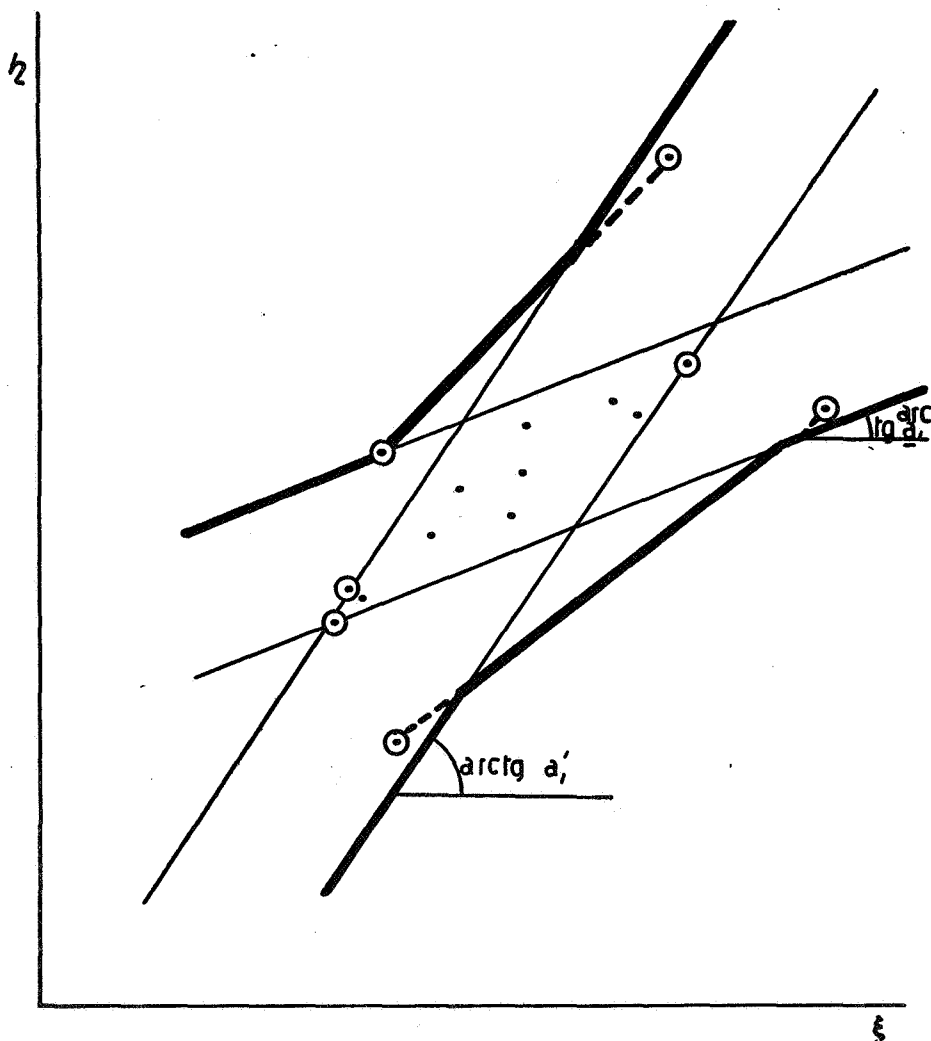

Fig. 5. $n = 15$ $s = 2$

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 a in the stochastic regression equation

$$\eta_i = A\,\xi_i^a + w_i, \qquad (i = 1, ..., n)$$

in which all ξ_i are positive, and in which

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

holds for $i = 1, ..., n$. $\xi_1, ..., \xi_n$ are known, A and a are unknown parameters, and $w_1, ..., w_n$ 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 (ξ_i, η_i) according to increasing magnitude of ξ and define

$$d_i = \frac{\log|\eta_{n_1+i}| - \log|\eta_i|}{\log\xi_{n_1+i} - \log\xi_i}, \qquad (i = 1, ..., n_1)$$

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:

$$d_{(1)} < ... < d_{(n_1)}$$

we have the following

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

Proof.   We have

$$\frac{\eta_{n_1+i} - w_{n_1+i}}{\eta_i - w_i} = \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^a$$

or:

$$\eta_{n_1+i} - \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^a \eta_i = w_{n_1+i} - \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^a w_i.$$

It follows from condition (2), that

$$w_{n_1+i} - \left(\frac{\xi_{n_1+i}}{\xi_i}\right)^a w_i$$

has a continuous distribution function with zero median. Hence:

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

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

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

17

if $\eta_{11}, \ldots, \eta_n$ are positive; if they are negative we have to replace η_i and η_{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 α is identical with the incomplete method for α in the linear equation

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

(which can be written as

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

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 $\varDelta(ij)$; each of these $\binom{n}{2}$ statistics has the property that its sampling median is equal to α_1 (cf. section 1. 3.). Hence one can use the sample median of the observed quantities $\varDelta(ij)$ as an estimate of α_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).