MATHEMATICS

A RANK-INVARIANT METHOD OF LINEAR AND POLYNOMIAL REGRESSION ANALYSIS

II [1])

BY

H. THEIL

(Communicated by Prof. D. VAN DANTZIG at the meeting of March 25, 1950)

2. CONFIDENCE REGIONS FOR THE PARAMETERS OF LINEAR REGRESSION EQUATIONS IN THREE AND MORE VARIABLES.

The probability set.

2. 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 will be supposed to be defined on this probability set.

In this first place we consider $n(\nu+2)$ random variables $\boldsymbol{u}_{\lambda i}$, $\boldsymbol{v}_i$, $\boldsymbol{w}_i$ $(\lambda=1,\ldots,\nu;\ i=1,\ldots,n)$. Furthermore we consider $(n+1)\nu+1$ parameters α_0, α_λ, $\xi_{\lambda i}$ $(i=1,\ldots n;\ \lambda=1,\ldots,\nu)$ and put:

$$\left.\begin{array}{lll}(5) & \theta_i & = \alpha_0 + \sum\limits_{\lambda=1}^{\nu} \alpha_\lambda \xi_{\lambda i} \\ (6) & \eta_i & = \theta_i + \boldsymbol{w}_i \\ (7) & \boldsymbol{x}_{\lambda i} & = \xi_{\lambda i} + \boldsymbol{u}_{\lambda i} \\ (8) & \boldsymbol{y}_i & = \eta_i + \boldsymbol{v}_i \end{array}\right\} \left\{\begin{array}{l} i=1,\ldots,n \\ \lambda=1,\ldots,\nu. \end{array}\right.$$

So the variables $\boldsymbol{x}_{\lambda i}$ and $\boldsymbol{y}_i$ have a simultaneous distribution on Γ, and are therefore random variables.

We call $\xi_{\lambda i}$ the parameter values of the variable ξ_λ. The equation (5) is the multiple regression equation. The random variables $\boldsymbol{w}_i$ are called "the true deviations from linearity", while the random variables $\boldsymbol{u}_{\lambda i}$ and $\boldsymbol{v}_i$ are called "the errors of observation" of the values $\xi_{\lambda i}$ and η_i respectively.

[1]) This paper is the second of a series of papers, the first of which appeared in these Proceedings, 53, 386−392 (1950).

Putting

$$\mathbf{z}_i = - \sum_{\lambda=1}^{\nu} a_\lambda \mathbf{u}_{\lambda i} + \mathbf{v}_i + \mathbf{w}_i$$

we have

$$\mathbf{y}_i = a_0 + \sum_{\lambda=1}^{\nu} a_\lambda \mathbf{x}_{\lambda i} + \mathbf{z}_i,$$

the random variables $\mathbf{z}_i$ being called "the apparent deviations from linearity".

Confidence regions for $a_0, a_1, \ldots, a_\nu$.

2. 1. In order to give confidence regions for the $(\nu + 1)$ parameters a_0, a_λ $(\lambda = 1, \ldots, \nu)$ we impose the following *conditions*:

Condition I: The n $(\nu + 2)$-uples $(\mathbf{u}_{1i}, \ldots, \mathbf{u}_{\nu i}, \mathbf{v}_i, \mathbf{w}_i)$ are stochastically independent.

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

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

Furthermore we impose *for the incomplete method* to be mentioned:

Condition III:

$$P[\mathbf{z}_i < \mathbf{z}_j] = P[\mathbf{z}_i > \mathbf{z}_j] = \tfrac{1}{2} \text{ for } i \neq j$$

and *for the complete method*:

Condition IIIa: Each $\mathbf{z}_i$ has the same continuous distribution function.

2. 2. Secondly we *define* the following quantities:

$$\mathbf{G}^{(\lambda')}(i) = \mathbf{y}_i - \sum_{\substack{\lambda=1 \\ \lambda \neq \lambda'}}^{\nu} a_\lambda \mathbf{x}_{\lambda i} =$$

$$= a_0 + a_{\lambda'} \mathbf{x}_{\lambda' i} + \mathbf{z}_i \qquad (\lambda' = 1, \ldots, \nu;\ i = 1, \ldots, n).$$

Furthermore, after arranging the n observed points $(y_i, x_{1i}, \ldots, x_{\nu i})$ according to increasing values of $x_{\lambda'}$ (which, by condition II, is identical with the arrangement according to increasing values of $\xi_{\lambda'}$):

$$x_{\lambda' 1} < x_{\lambda' 2} < \ldots < x_{\lambda' n}$$

we define the quantities

$$\mathbf{K}^{(\lambda')}(ij) = \frac{\mathbf{G}^{(\lambda')}(i) - \mathbf{G}^{(\lambda')}(j)}{\mathbf{x}_{\lambda' i} - \mathbf{x}_{\lambda' j}} =$$

$$= \frac{\mathbf{y}_i - \mathbf{y}_j}{\mathbf{x}_{\lambda' i} - \mathbf{x}_{\lambda' j}} - \sum_{\substack{\lambda=1 \\ \lambda \neq \lambda'}}^{\nu} a_\lambda \frac{\mathbf{x}_{\lambda i} - \mathbf{x}_{\lambda j}}{\mathbf{x}_{\lambda' i} - \mathbf{x}_{\lambda' j}} =$$

$$= a_{\lambda'} + \frac{\mathbf{z}_i - \mathbf{z}_j}{\mathbf{x}_{\lambda' i} - \mathbf{x}_{\lambda'}} \qquad (i = 1, \ldots, n-1;\ j = i + 1, \ldots, n).$$

For any set of values $a_1, \ldots, a_{\lambda'-1}, a_{\lambda'+1}, \ldots, a_\nu$ we arrange the quantities $K^{(\lambda')}(i\,j)$ according to increasing magnitude; we define $K_i^{(\lambda')}$ as the quantity with rank i in this arrangement:

$$K_1^{(\lambda')} < K_2^{(\lambda')} < \ldots < K_{\binom{n}{2}}^{(\lambda')}$$

Finally we define the intervals $\boldsymbol{I}_{\lambda'}(a_1, \ldots, a_{\lambda'-1}, a_{\lambda'+1}, \ldots, a_\nu)$ as the intervals

$$\left(\boldsymbol{K}_q^{(\lambda')},\ \boldsymbol{K}_{\binom{n}{2}-q-1}^{(\lambda')}\right)$$

with $2\,q \leqq \binom{n}{2}$; $\boldsymbol{A}_{\lambda'}$ as the union of

$$\boldsymbol{I}_{\lambda'}(a_1,\ldots, a_{\lambda'-1}, a_{\lambda'+1},\ldots, a_\nu) \text{ for all } a_\lambda\ (\lambda = 1,\ldots,\nu;\ \lambda \neq \lambda');$$

and $\boldsymbol{A}$ as the union of all $\boldsymbol{A}_{\lambda'}$ $(\lambda' = 1, \ldots, \nu)$.

2. 3. We have the following theorem concerning the *complete method* for three and more variables:

Theorem 4: Under conditions I, II and III*a* the region $\boldsymbol{A}$ is a confidence region for the parameters $a_1, \ldots, a_\nu$, the level of significance being $\leqq 2\,\nu . P[q-1\,|\,n]$ [2]).

Proof: If the set of assumed parameters values $a_1,\ldots, a_{\lambda'-1}, a_{\lambda'+1},\ldots, a_\nu$ is the "true" set, it follows from the analysis in section 1.3., that $\boldsymbol{I}_{\lambda'}(a_1, \ldots, a_{\lambda'-1}, a_{\lambda'+1}, \ldots, a_\nu)$ is a confidence interval for $a_{\lambda'}$ to the level of significance $2\,P[q-1|n]$. Hence it follows that if $(a_1, \ldots, a_\nu)$ represents the "true" point in the $a_1, \ldots, a_\nu$-space, we have

$$P\,[(a_1,\ldots, a_\nu) \in \boldsymbol{A}_{\lambda'}] = 1 - 2\,P\,[q-1\,|\,n], \qquad (\lambda' = 1,\ldots,\nu),$$

which proves the theorem.

2. 4. If condition III (but not necessarily III*a*) is fulfilled, the method mentioned above can be replaced by the following one. We replace the quantities

$$\boldsymbol{K}^{(\lambda')}(i\,j) \qquad (\lambda' = 1,\ldots,\nu;\ i = 1,\ldots, n-1;\ j = i+1,\ldots, n)$$

by

$$\boldsymbol{K}^{(\lambda')}(i, n_1 + i) \qquad (\lambda' = 1,\ldots,\nu;\ i = 1,\ldots, n_1). \text{ [3])}$$

The intervals $\boldsymbol{I}'_{\lambda'}(a_1, \ldots, a_{\lambda'-1}, a_{\lambda'+1}, \ldots, a_\nu)$ are now defined as the intervals bounded by the values of $\boldsymbol{K}^{(\lambda')}(i, n_1 + i)$ with rank r_1 and $(n_1 - r_1 + 1)$ respectively, if they are arranged in ascending order; whereas the definitions of $\boldsymbol{A}'_{\lambda'}$ as the union of all $\boldsymbol{I}'_{\lambda'}$ and of $\boldsymbol{A}'$ as the union of all $\boldsymbol{A}'_{\lambda'}$

---

[2]) For the definition of $P\,[q-1\,|\,n]$ the reader is referred to section 1.3. (part I of this paper).

[3]) $n_1 = \frac{1}{2}\,n$. Cf. section 1.2.

remain unchanged. The following theorem of the *incomplete method* for three and more variables will now be obvious from the analysis of section 1. 1.:

Theorem 5. Under conditions I, II and III the region $\mathbf{A}'$ is a confidence region for the parameters $a_1, \ldots, a_\nu$, the level of significance being $\leqq 2\,\nu.\, I_{\frac{1}{2}}(r_1, n_1 - r_1 + 1)$.

2. 5. A confidence region for the parameters $a_0, a_1, \ldots, a_\nu$ can be constructed, if the median of $\mathbf{z}_i$ is known, e.g. if the following condition is fulfilled:

Condition IV: The median of each $\mathbf{z}_i$ is zero.

The method for the construction of this confidence region is analogous to the one given in section 1. 2.

An illustration for the special case $\nu = 2$.

2. 6. The form of the region $\mathbf{A}_\lambda$ or $\mathbf{A}'_\lambda$ will now be indicated for the case of three variables:

$$\mathbf{y}_i = a_0 + a_1\,\mathbf{x}_{1i} + a_2\,\mathbf{x}_{2i} + \mathbf{z}_i.$$

Using the incomplete method we find n_1 functions of a_2:

$$\mathbf{K}^{(1)}(i, n_1 + i) = \frac{\mathbf{y}_i - \mathbf{y}_{n_1+i}}{\mathbf{x}_{1i} - \mathbf{x}_{1,n_1+i}} - a_2\,\frac{\mathbf{x}_{2i} - \mathbf{x}_{2,n_1+i}}{\mathbf{x}_{1i} - \mathbf{x}_{1,n_1+i}},$$

which are estimates of a_1, given a_2. They are represented by straight lines in the a_1, a_2-plane. For any value of a_2 we can arrange these quantities in ascending order. As long as (under continuous variation of a_2) the numbers i_1 and i_2 for which the statistics $K^{(1)}(i_1, n_1 + i_1)$ and $K^{(1)}(i_2, n_1 + i_2)$

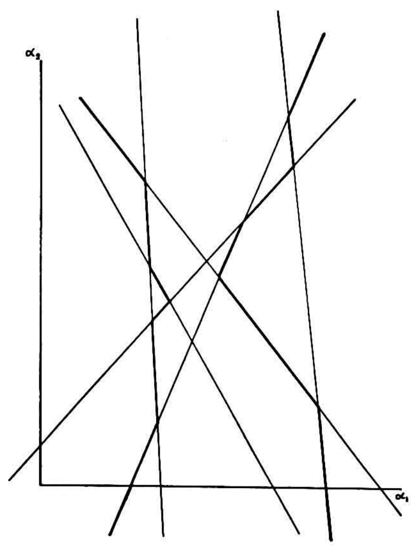

Fig. 1. $n_1 = 6$, $r_1 = 2$.

have the r_1-th and $(n_1 - r_1 + 1)$-th rank according to increasing order (with r_1 as defined in section 2. 4.) remain constant, the extreme points of the confidence intervals vary along straight lines. If, when passing some value a_2^* of a_2 either i_1 or i_2 changes, the corresponding straight line passes into another one, intersecting the first one in a point with $a_2 = a_2^*$.

So a diagram can be constructed, in which the n_1 straight lines are drawn in the a_1, a_2-plane. This gives the stochastic region $\mathbf{A}_1'$ depending on the given observations and bounded to the left and to the right by broken lines.

According to Theorem 5 it contains the true point (a_1, a_2) with the probability

$$1 - 2\, I_{\frac{1}{2}}\, (r_1, n_1 - r_1 + 1).$$

The region $\mathbf{A}_2'$, bounded above and below, can be constructed in a similar way; then the observed points must be arranged in ascending order of $\boldsymbol{x}_2$.

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