# A RANK-INVARIANT METHOD OF LINEAR AND POLYNOMIAL REGRESSION ANALYSIS

## I

BY

H. THEIL

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

## 0. INTRODUCTION

0.0 Regression analysis is usually carried out under the hypothesis that one of the variables is normally distributed with constant variance, its mean being a function of the other variables. This assumption is not always satisfied, and in most cases difficult to ascertain.

In recent years attention has been paid to problems of estimating the parameters of regression equations under more general conditions (see the references at the end of this paper: A. WALD (1940), K. R. NAIR and M. P. SHRIVASTAVA (1942), K. R. NAIR and K. S. BANERJEE (1942), G. W. HOUSNER and J. F. BRENNAN (1948) and M. S. BARTLETT (1949)). Confidence regions, however, were obtained under the assumption of normality only; to obtain these without this assumption will be the main object of this paper.

0.1. In section 1. confidence regions will be given for the parameters of linear regression equations in two variables. In the sequel of this paper we hope to deal with equations in more variables, polynomial equations, systems of equations and problems of prediction.

## 1. CONFIDENCE REGIONS FOR THE PARAMETERS OF LINEAR REGRESSION EQUATIONS IN TWO VARIABLES

*The probability set.*

1.0. Throughout this section the probability set $\Gamma$ ("Wahrscheinlichkeitsfeld" in the sense of A. KOLMOGOROFF) underlying the probability statements will be the $3n$-dimensional Cartesian space $R_{3n}$ with coordinates $u_1, \ldots, u_n, v_1, \ldots, v_n, w_1, \ldots, w_n$. Every random variable mentioned is supposed to be defined on this probability set.

In the first place we suppose 3 $n$ random variables $\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i$ $(i = 1, \ldots, n)$ [1]) to be defined on $\Gamma$, i.e. we suppose $\mathbf{u}_i$, $\mathbf{v}_i$, $\mathbf{w}_i$ to have a simultaneous probability distribution on $\Gamma$.

---

[1]) The distinction between a stochastic variable and the value it takes in a given observation (or system of observations) will be indicated by bold type for the former one.

If we now put:

$$\left.\begin{array}{ll} (1) & \theta_i = \alpha_0 + \alpha_1 \xi_i \\ (2) & \eta_i = \theta_i + \mathbf{w}_i \\ (3) & \mathbf{x}_i = \xi_i + \mathbf{u}_i \\ (4) & \mathbf{y}_i = \eta_i + \mathbf{v}_i \end{array}\right\} \quad i = 1, \ldots, n$$

then, for any set of values of the $(n+2)$ parameters $\xi_i$, $\alpha_0$ and $\alpha_1$, the varables $\mathbf{x}_i$ and $\mathbf{y}_i$ have a simultaneous distribution on $\Gamma$, and are therefore random variables.

We shall call $\xi_i$ the parameter values of the variable $\xi$. The equation (1) is the regression equation; this equation contains no stochastic variables. Furthermore we shall call $\mathbf{w}_i$ the "true deviations from linearity"; hence the variable $\eta$ is a linear function of $\xi$, but for the deviations $\mathbf{w}$. Finally $\mathbf{u}_i$ and $\mathbf{v}_i$ are called the "errors of observation" of the "true" values $\xi_i$ and $\eta_i$ respectively.

The problem then is, under certain conditions for the probability distribution of $\mathbf{u}_i$, $\mathbf{v}_i$, $\mathbf{w}_i$, to determine confidence intervals for the parameters $\alpha_0$ and $\alpha_1$, given a sequence of observations $x_1, \ldots, x_n, y_1, \ldots, y_n$ of the random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n$.

*Incomplete method: confidence interval for* $\alpha_1$. [2])

1.1. We suppose that the following conditions are satisfied:

*Condition I:* The $n$ triples $(\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i)$ are stochastically independent.

*Condition II:* 1. Each of the errors $\mathbf{u}_i$ vanishes outside a finite interval $|\mathbf{u}_i| \leqq g_i$.

2. For each $i \neq j$ we have: $|\xi_i - \xi_j| > g_i + g_j$.

From condition II it follows that either

$$P[\mathbf{x}_i < \mathbf{x}_j] = 1 \quad \text{and} \quad \xi_i < \xi_j$$

or

$$P[\mathbf{x}_i > \mathbf{x}_j] = 1 \quad \text{and} \quad \xi_i > \xi_j.$$

This condition means that the errors $\mathbf{u}_i$ are sufficiently small in order that arrangement of the observed values $x_i$ according to increasing magnitude be identical with the arrangement according to increasing values of $\xi_i$ (cf. also A. WALD (1940), p. 294, seq., where a similar (weaker) condition is imposed). The arrangement of the $\mathbf{x}_i$ is therefore uniquely determined. We therefore suppose the $\mathbf{x}_i$ as well as the $\xi_i$ to be arranged according to increasing order.

Put $n_1 = n - [\frac{1}{2} n]$; if $n$ is odd, the observation with rank $\frac{1}{2}(n+1)$ is not used. We therefore omit this observation and write $n = 2 n_1$.

---

[2]) The author is indebted to Mr J. HEMELRIJK for his constructive criticism concerning some points of this section.

We determine the following $n_1$ statistics:

$$\Delta(i, n_1+i) = \frac{\mathbf{y}_{n_1+i} - \mathbf{y}_i}{\mathbf{x}_{n_1+i} - \mathbf{x}_i} = \alpha_1 + \frac{\mathbf{z}_{n_1+i} - \mathbf{z}_i}{\mathbf{x}_{n_1+i} - \mathbf{x}_i},$$

in which $\mathbf{z}_i = -\alpha_1 \mathbf{u}_i + \mathbf{v}_i + \mathbf{w}_i$.

We now impose:

*Condition III*, which states:

$$P[\mathbf{z}_i < \mathbf{z}_{n_1+i}] = P[\mathbf{z}_i > \mathbf{z}_{n_1+i}] = \tfrac{1}{2}.$$

As all denominators $\mathbf{x}_{n_1+i} - \mathbf{x}_i$ are positive, it follows that

$$P[\Delta(i, n_1+i) < \alpha_1] = P[\Delta(i, n_1+i) > \alpha_1] = \tfrac{1}{2},$$

i.e. that $\Delta(i, n_1+i)$ has a median $\alpha_1$ and that its distribution function is continuous in the median.

The following conditions III*a* and III*b* are each sufficient in order that $P[\mathbf{z}_i < \mathbf{z}_{n_1+i}] = P[\mathbf{z}_i > \mathbf{z}_{n_1+i}] = \tfrac{1}{2}$:

*Condition IIIa:* the random variables $\mathbf{z}_i$ $(i = 1, \ldots, n)$ have the same continuous distribution function.

*Condition IIIb:* the random variables $\mathbf{z}_i$ have continuous distribution functions which are symmetrical with equal medians med $(\mathbf{z})$.

Proof: In case III*a* the simultaneous distribution of $\mathbf{z}_i$ and $\mathbf{z}_{n_1+i}$ is symmetrical about the line $z_i = z_{n_1+i}$, which proves the statement. In case III*b* it is symmetrical about the lines $z_i =$ med $(\mathbf{z})$ and $z_{n_1+i} =$ med $(\mathbf{z})$; hence the simultaneous distribution of $\mathbf{z}_i -$ med $(\mathbf{z})$ and $\mathbf{z}_{n_1+i} -$ med $(\mathbf{z})$ is symmetrical with respect to the origin, which proves the statement.

We now arrange the $n_1$ statistics $\Delta(i, n_1+i)$ in increasing order:

$$\Delta_1 < \Delta_2 < \ldots < \Delta_{n_1},$$

in which

$$\Delta_j = \Delta(i_j, n_1 + i_j).$$

The probability that exactly $r$ among the $n_1$ values $\Delta(i, n_1+i)$ are $< \alpha_1$, i.e. that $\Delta_r < \alpha_1 < \Delta_{r+1}$, is $2^{-n_1}\binom{n_1}{r}$ because of the conditions I and III. Hence:

$$P[\Delta_{r_1} \leqq \alpha_1 \leqq \Delta_{n_1-r_1+1} \mid \alpha_1] =$$

$$= 1 - 2^{1-n_1} \sum_{s=0}^{r_1-1} \binom{n_1}{s}$$

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

in which

$$I_{\frac{1}{2}}(r_1, n_1 - r_1 + 1) = \frac{\int_0^{\frac{1}{2}} x^{r_1-1}(1-x)^{n_1-r_1}\, dx}{\int_0^1 x^{r_1-1}(1-x)^{n_1-r_1}\, dx}$$

is the incomplete Beta-function for the argument $\tfrac{1}{2}$.

So we have proved:

*Theorem I:* under conditions I, II and III a confidence interval for $\alpha_1$ is given by the largest but $(r_1 - 1)$ and the smallest but $(r_1 - 1)$ among the values $\Delta(i, n_1 + i)$, the level of significance being $2 I_{\frac{1}{2}}(r_1, n_1 - r_1 + 1)$.

We shall call this method an "incomplete method" because a limited use is made of the $\binom{n}{2}$ statistics

$$\Delta(ij) = \frac{\mathbf{y}_i - \mathbf{y}_j}{\mathbf{x}_i - \mathbf{x}_j} \qquad (i < j).$$

*Incomplete method: confidence region for $\alpha_0$ and $\alpha_1$.*

1.2. If the median of $\mathbf{z}_i$ $(i = 1, \ldots, n)$ is numerically known, a confidence region for $\alpha_0$ and $\alpha_1$ can be found. We suppose that the following condition is satisfied:

*Condition IV:* the median of each $\mathbf{z}_i$ $(i = 1, \ldots, n)$ is zero:

$$P[\mathbf{y}_i - \alpha_1 \mathbf{x}_i > \alpha_0] = P[\mathbf{y}_i - \alpha_1 \mathbf{x}_i < \alpha_0] = \tfrac{1}{2}.$$

For any value of $\alpha_1$ we can arrange the $n$ quantities $Z_i = y_i - \alpha_1 x_i$ according to increasing magnitude:

$$Z_1(\alpha_1) < Z_2(\alpha_1) < \ldots < Z_n(\alpha_1).$$

Under the condition that $\alpha_1$ has the value used in this arrangement, we can state that

$$P[\alpha_0 \in (\mathbf{Z}_{r_0}(\alpha_1),\ \mathbf{Z}_{n-r_0+1}(\alpha_1)) \mid \alpha_0, \alpha_1] =$$
$$= 1 - 2 I_{\frac{1}{2}}(r_0, n - r_0 + 1) = 1 - \varepsilon_0.$$

On the other hand, if we write $\boldsymbol{I}_1$ for the interval $(\Delta_{r_1}, \Delta_{n_1 - r_1 + 1})$, we can state:

$$P[\alpha_1 \in \boldsymbol{I}_1 \mid \alpha_1] =$$
$$= 1 - 2 I_{\frac{1}{2}}(r_1, n_1 - r_1 + 1) = 1 - \varepsilon_1.$$

If we denote by $\boldsymbol{I}_0$ the interval bounded by the lowest of the values $\mathbf{Z}_{r_0}(\alpha_1)$ if $\alpha_1$ varies through $\boldsymbol{I}_1$ and by the largest of the values $\mathbf{Z}_{n-r_0+1}(\alpha_1)$ if $\alpha_1$ varies through $\boldsymbol{I}_1$ we have

$$P[\alpha_0 \in \boldsymbol{I}_0 \wedge \alpha_1 \in \boldsymbol{I}_1 \mid \alpha_0, \alpha_1] \geqq (1 - \varepsilon_0)(1 - \varepsilon_1).$$

So we have proved:

*Theorem 2*: under conditions I, II, III and IV a rectangular confidence region in the $\alpha_0$, $\alpha_1$ — plane is given by the intervals $\alpha_0 \in \boldsymbol{I}_0$ and $\alpha_1 \in \boldsymbol{I}_1$, the level of significance being $\leqslant \varepsilon_0 + \varepsilon_1 - \varepsilon_0 \varepsilon_1$.

If all observed points $(x_i, y_i)$ obey the inequality $x_i \geqslant 0$, all quantities $y_i - \alpha_1 x_i$ are decreasing functions of $\alpha_1$. It follows that $I_0$ is bounded by $Z_{r_0}(\Delta_{n_1 - r_1 + 1})$ and by $Z_{n-r_0+1}(\Delta_{r_1})$. The converse holds if every point satisfies the inequality $x_i \leqq 0$.

*Complete method.*

1.3. We suppose that the conditions I, II and III$a$ are satisfied and consider two arrangements of the points $(x_i, y_i)$: the arrangement according to increasing values of $x$ and that according to $z = y - \alpha_0 - \alpha_1 x$.

The arrangement according to $z$ is possible for any assumed value of $\alpha_1$. The hypothesis that this value is the true one is rejected if and only if there is a significant rank correlation between the arrangements.

Consider the statistics

$$\Delta(ij) = \frac{y_i - y_j}{x_i - x_j} = \alpha_1 + \frac{z_i - z_j}{x_i - x_j},$$

in which $i < j$, so that (if the ordering is according to $x$) $x_i < x_j$ and $\xi_i < \xi_j$. It follows that $\Delta(ij) > \alpha_1$, if and only if $z_i < z_j$.

Now, under the null hypothesis that the arrangements of the points according to $x$ and according to $z$ are independent, the distribution of Kendall's "*rank correlation coefficient*"

$$\frac{S}{\binom{n}{2}}$$

is known, in which $S$ is the number of cases in which the ordering according to $z$ is the same as the ordering according to $x$ ($z_k < z_{k'}$, and $x_k < x_{k'}$) minus the number of cases in which the ordering according to $z$ is the inverse as compared with the one according to $x$ ($z_k > z_{k'}$, while $x_k < x_{k'}$).

For any value of $\alpha_1$ the number of cases $z_i > z_j$ can be found. Suppose this to be $q$; it will be clear that

$$S = \binom{n}{2} - 2q\,.$$

The distribution function of $S$ for any value of $n$ has been given by M. G. Kendall (see M. G. KENDALL (1947), p. 403—407 and (1948), p. 55—62) by means of a recurrence formula. So the probability $P[q|n]$ that $q' \leqq q$ cases $z_i > z_j$ are found can be determined. If this probability is below the level of significance chosen, we reject the hypothesis that $\alpha_1$ has the value used in the arrangement according to $z$.

Hence, if we arrange the statistics $\Delta(ij)$ in increasing order:

$$\Delta_1 < \Delta_2 < \ldots < \Delta_{\binom{n}{2}}$$

we find by symmetry

$$P[\Delta_q \leqq \alpha_1 \leqq \Delta_{\binom{n}{2}-q+1} \mid \alpha_1] = 1 - 2P[q-1 \mid n]$$

so that we have proved:

*Theorem* 3: under conditions I, II and III$a$ a confidence interval for $\alpha_1$ is given by the largest but $(q-1)$ and the smallest but $(q-1)$ among the values $\Delta(ij)$, the level of significance being $2P[q-1|n]$.

The method of 1. 2. can be applied here to find a simultaneous confidence region for $\alpha_0$ and $\alpha_1$, $I_1$ now being the interval $(\Delta_q, \Delta_{\binom{n}{2}-q+1})$.

*A comparison.*

1. 4. The second method may be called a "complete method", because all statistics $\Delta(i\,j)$ are used. It requires only 5 points in order to reach the level of significance 0,05 whereas the limited method needs 12 points. However, if the number of points is large, the computational labor of the complete method is considerably greater than that of the incomplete method. Moreover, the conditions under which the complete method is valid are more stringent; the fact that the set of conditions I, II and III is sufficient for the incomplete method is important in view of the general occurrence of "heteroscedastic" distribution, i.e. distributions in which the variance (if finite) of $\eta_i$ is larger for higher values of $\xi$ than for lower ones if $\alpha_1 > 0$ and conversely if $\alpha_1 < 0$.

*Testing linearity.*

1. 5. Suppose that the set of conditions I, II and III$a$ is valid. Then the hypothesis that the regression curve for two variables is linear can be tested against the alternative composite hypothesis that it is either positive- or negative-convex, [3]) i.e. in the set of equations (1), (2), (3), (4) the equation $\theta_i = \alpha_0 + \alpha_1 \xi_i$ is tested against any equation $\theta_i = \theta(\xi_i)$ with

either $$\frac{d^2\theta}{d\xi^2} > 0 \quad \text{for all } \xi$$

or $$\frac{d^2\theta}{d\xi^2} < 0 \quad \text{for all } \xi,$$

the equations (2), (3), (4) remaining unchanged.

Consider the $n_1$ statistics

$$\Delta(1, n_1+1), \ldots, \Delta(n_1, 2n_1)$$

in this arrangement. If this ordering has a significant rank correlation with the ordering of these statistics according to increasing magnitude, we reject the hypothesis that the regression curve is linear.

## REFERENCES

BARTLETT, M. S., Fitting a straight line when both variables are subject to error. *Biometrics* **5**, 207–212 (1949).

DANTZIG, D. VAN, Capita selecta der waarschijnlijkheidsrekening, caput II, (stenciled) (1947).

---

[3]) A function $f(x)$ is positive-convex (cf. e.g. D. VAN DANTZIG, 93–94 (1947)) in an interval if for every $x_1$ and $x_2$ of this interval and for every real positive number $a < 1$ the following inequality is satisfied

$$a f(x_1) + (1-a) f(x_2) > f(a x_1 + \overline{1-a}\, x_2).$$

HOUSNER, G. W. and J. F. BRENNAN, The estimation of linear trends. *Annals of Mathematical Statistics*, **19**, 380—388 (1948).

KENDALL, M. G., *The advanced theory of statistics*, London, **1**, 3rd edition (1947).

———, *Rank correlation methods*, London (1948).

NAIR, K. R. and K. S. BANERJEE, A note on fitting of straight lines if both variables are subject to error. *Sankhya*, **6**, 331 (1942).

——— and M. P. SHRIVASTAVA, On a simple method of curve fitting. *Sankhya*, **6**, 121—132 (1942).

WALD, A., The fitting of straight lines if both variables are subject to error. *Annals of mathematical statistics*, **11**, 284—300 (1940).

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