CHAPTER 20

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

HENRI THEIL
Economic Research Institute
University of Amsterdam
Amsterdam, The Netherlands

PART I

1. Introduction

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.

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.

2. Confidence Regions for the Parameters of Linear Regression Equations in Two Variables

2.1 THE PROBABILITY SET

Throughout this section the probability set Γ ("Wahrscheinlich- keitsfeld" in the sense of A. Kolmogoroff) underlying the probability statements will be the $3n$-dimensional

* This article first appeared in the *Proceedings of the Royal Netherlands Academy of Sciences* **53** (1950) Part I: 386-392, Part II: 521-525, Part III: 1397-1412. Reprinted here with the permission of the Royal Netherlands Academy of Arts and Sciences.

B. Raj et al. (eds.), *Henri Theil's Contributions to Economics and Econometrics*

Cartesian space R_{3n} with coordinates $u_1, ..., u_n, v_1, ..., v_n, w_1, ..., w_n$. Every random variable mentioned is supposed to be defined on this probability set.

In the first place we suppose $3n$ random variables u_i, v_i, w_i $(i = 1, ..., n)$[1] to be defined on Γ, i.e. we suppose u_i, v_i, w_i to have a simultaneous probability distribution on Γ.

If we now put:

$$\left.\begin{aligned} \theta_i &= \alpha_0 + \alpha_1 \xi_i \\ \eta_i &= \theta_i + w_i \\ x_i &= \xi_i + u_i \\ y_i &= \eta_i + v_i \end{aligned}\right\} \quad i = 1,...,n \qquad \begin{aligned} &(1)\\ &(2)\\ &(3)\\ &(4) \end{aligned}$$

then, for any set of values of the $(n+2)$ parameters ξ_i, α_0 and α_1, the variables x_i and y_i have a simultaneous distribution on Γ, and are therefore random variables.

We shall call ξ_i the parameter values of the variable ξ. The equation (1) is the regression equation; this equation contains no stochastic variables. Furthermore we shall call w_i the "true deviations from linearity"; hence the variable η is a linear function of ξ, but for the deviations w. Finally u_i and v_i are called the "errors of observation" of the true values ξ_i and η_i respectively.

The problem then is, under certain conditions for the probability distribution of u_i, v_i, w_i, to determine confidence intervals for the parameters α_0 and α_1, given a sequence of observations $x_1, ..., x_n, y_1, ..., y_n$ of the random variables $x_1, ..., x_n, y_1, ..., y_n$.

2.2 INCOMPLETE METHOD: CONFIDENCE INTERVAL FOR α_1.[2]

We suppose that the following conditions are satisfied:

Condition I: The n triples (u_i, v_i, w_i) are stochastically independent.

Condition II: 1. Each of the errors u_i vanishes outside a finite interval $|u_i| \leq g_i$.

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

From condition II it follows that either

---

[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.

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