ON THE SPECIFICATION ERROR IN REGRESSION ANALYSIS

By H. Wold and P. Faxér

University of Uppsala, Sweden

In the difference between a statistical estimate and the corresponding theoretical value it is customary to distinguish between the sampling error, which arises because the estimate is based on a finite sample from a specified population, and the specification error, which arises if the population is not correctly described in the assumptions that form the basis of the estimation method. It is easy to see that the specification error of a least squares regression coefficient will be small if (A) the disturbance term is small, *or* if (B) the disturbance is nearly uncorrelated with the explanatory variables. The proximity theorem ([1], Theorem 12. 1.3; see also p. 37) states the simple fact that conditions (A) and (B) strengthen each other, to the effect that if they are fulfilled up to magnitudes of the first order, the specification error will be small of the second order. The present note gives limits for the unspecified constant that is involved in the proximity theorem.

We shall first prove an auxiliary lemma which contains the proximity theorem, and from which the limits sought for will be deduced by way of a corollary. It is sufficient for our purpose to consider large samples, so as not to place emphasis on the difference between observed and theoretical values for variances, correlation coefficients, etc.

Lemma. *Given the theoretical relation*

$$y = \beta_1 x_1 + \cdots + \beta_h x_h + \zeta \tag{1}$$

suppose: (a) *the disturbance* ζ *has zero expectation and finite variance* $\sigma^2(\zeta)$, *but is otherwise arbitrary, and* (b) *none of the explanatory variables* $x_1, \cdots, x_h$ *is identically linear in the other ones. Let*

$$y = b_1 x_1 + \cdots + b_h x_h + z \tag{2}$$

be the least squares regression of y *on* $x_1, \cdots, x_h$. *Then*

$$|b_i - \beta_i| \leqq \frac{\sigma(\zeta)}{\sigma(x_i)\sqrt{1 - R_i^2}}, \tag{3}$$

where $R_i = R_{i(1,2,\cdots,i-1,i+1,\cdots,h)}$ *is the multiple correlation coefficient of* x_i *and* $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_h$.

Proof. The assumptions of the lemma lead us to regard the joint distribution of $x_1, \cdots, x_h$ as given and the distribution of ζ as unspecified. Hence if $\rho(\xi, \mu)$ denotes the correlation coefficient of ξ and μ, the coefficients

$$\rho_{ij} = \rho(x_i, x_j), \qquad i, j = 1, \cdots, h,$$

Received May 18, 1956.