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.

are given, with $\rho_{ii} = 1$, while the

$$r_i = \rho(x_i, \zeta), \qquad i = 1, \cdots, h,$$

are unspecified. Then if $(u_1, \cdots, u_h)$ are coordinates of h-dimensional Euclidean space, it is known ([2]; cf. also [1], Theorem 12.3.5) that the point $r = (r_1, \cdots, r_h)$ lies inside or on the boundary of the ellipsoid defined by

$$\begin{vmatrix} \rho_{11} & \cdots & \rho_{1h} & u_1 \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{h1} & \cdots & \rho_{hh} & u_h \\ u_1 & \cdots & u_h & 1 \end{vmatrix} = 0. \tag{4}$$

Writing P for the determinant

$$P = \begin{vmatrix} \rho_{11} & \cdots & \rho_{1h} \\ \cdots & \cdots & \cdots \\ \rho_{h1} & \cdots & \rho_{hh} \end{vmatrix}$$

and P_{ij} for its cofactors, we further remark that (b) implies

$$1 - R_i^2 = P/P_{ii} > 0, \qquad i = 1, \cdots, h.$$

It will suffice to verify (3) for $i = 1$. The method of least squares regression gives

$$b_1 - \beta_1 = \sigma(\xi) \sum_1^h r_i P_{1i}/\sigma(x_1) P. \tag{5}$$

Next, $b_1 - \beta_1$ being proportional to the linear form

$$f = P_{11} r_1 + \cdots + P_{1h} r_h$$

we ask for the extremum of f, and this clearly occurs for a point r on the boundary of the ellipsoid (4), that is, for a point which satisfies

$$\sum_{i,j=1}^h P_{ij} r_i r_j = P. \tag{6}$$

We shall accordingly seek the extremum of

$$F = f + \lambda \left[\sum_{i,j=1}^h P_{ij} r_i r_j - P \right],$$

where λ is a Lagrange multiplier. This leads to the system

$$P_{1i} + 2\lambda \sum_{j=1}^h P_{ij} r_j = 0, \qquad i = 1, \cdots, h,$$

which combined with (6) gives the solution

$$r_1 = \pm\sqrt{1 - R_1^2}; \qquad r_2 = \cdots = r_h = 0.$$

Substituting in (5), we infer that the extremum of the specification error is given by the right-hand member of (3). The lemma is proved.

We shall now render the introductory conditions (A)-(B) exact. First, we write

$$\text{(A)} \qquad \sigma(\zeta) \leqq \epsilon \cdot \sigma(x_i), \qquad i = 1, \cdots, h,$$

where $\epsilon \geqq 0$. Then if ϵ is small, the disturbance ζ is small in the sense that its standard deviation is small relative to the standard deviations of the explanatory variables. To give condition (B) a convenient form we observe that for given $r_1, \cdots, r_h$ there is a point $(r_1^*, \cdots, r_h^*)$ on the boundary of the ellipsoid (4) and a proportionality factor ϵ' with $0 \leqq \epsilon' \leqq 1$ such that

$$\text{(B)} \qquad r_i = \epsilon' \cdot r_i^*, \qquad i = 1, \cdots, h.$$

Then if ϵ' is small, the correlations r_i are small in the sense that the point $(r_1, \cdots, r_h)$ lies near the centre of the ellipsoid.

Thus prepared, we obtain the following

COROLLARY. *On conditions* (a) *and* (b) *of the lemma, we have*

$$| b_i - \beta_i | \leqq \epsilon \cdot \epsilon' / \sqrt{1 - R_i^2}, \qquad i = 1, \cdots, h,$$

where ϵ *and* ϵ' *are defined by* (A)-(B).

Hence if ϵ and ϵ' are of small order the specification error of the regression coefficients b_i will at most be of order $\epsilon\epsilon'$.

In the special case of one explanatory variable, $h = 1$, we have $R_1 = 0$ and $| b_1 - \beta_1 | \leqq \epsilon\epsilon'$. For example, if $\sigma(\zeta) = \frac{1}{5}\sigma(x_1)$ and $r_1 = \rho(x_1, \zeta) = \frac{1}{5}$, the specification error of b_1 cannot exceed 0, 04.

REFERENCES

1. H. WOLD IN ASSOCIATION WITH L. JURÉEN, *Demand Analysis: A Study in Econometrics*, Geber, Stockholm, 1952; and John Wiley and Sons, New York, 1953.
2. H. WOLD, "A theorem on regression coefficients obtained from successively extended sets of variables," *Skand. Aktuarietids.*, Vol. 28 (1945), pp. 181–200.

SETS OF MEASURES NOT ADMITTING NECESSITY AND SUFFICIENT STATISTICS OR SUBFIELDS[1, 2]

BY T. S. PITCHER[3]

Let X be the interval from 0 to 1 and F the field of Borel sets on X. For every $x \leqq \frac{1}{2}$, let m_x be the probability measure assigning probability $\frac{1}{2}$ to the point x and probability $\frac{1}{2}$ to the point $(x + \frac{1}{2})$ and let F_x be the subfield of F consisting of all Borel sets which contain both x and $(x + \frac{1}{2})$ or else neither. Then if M is a set of probability measures consisting of all m_x, $0 \leqq x < \frac{1}{2}$ and some measures assigning probability 0 to every point, the only set of m-measure zero

Received May 4, 1956; revised June 12, 1956.

[1] The research in this document was supported jointly by the Army, Navy, and Air Force, under contract with the Massachusetts Institute of Technology.

[2] For definitions of these concepts, see references [1] and [2].

[3] Staff Member, M.I.T., Lincoln Laboratory.