Next, let for $n > e$

$$a_n = n \log^{-\eta} n. \tag{12}$$

Then (8) holds and from (9) and (10) we obtain easily for large n

$$b_n = \sum_{k=1}^{n} \{1 - \log^{-\eta} a_k\} < n - (1 - \epsilon)a_n . \tag{13}$$

Substituting into (7) one sees that, again for sufficiently large n,

$$\Pr \{S_n - n + (1 - \epsilon)a_n < \epsilon a_n\} \to 1, \tag{14}$$

or, since $M = 1$,

$$\Pr \{S_n - nM < -(1 - 2\epsilon)a_n\} \to 1. \tag{15}$$

This proves (I).

A NOTE ON RANK, MULTICOLLINEARITY AND MULTIPLE REGRESSION[1]

By Gerhard Tintner

Iowa State College

Let $X_{it} (i = 1, 2 \cdots M)$ be set of M random variables, each being observed at $t = 1, 2 \cdots N$. $X_{it} = M_{it} + y_{it}$. (This is essentially the situation envisaged by Frisch [1]). The systematic part of our variables $M_{it} = EX_{it}$. The y_{it} are normally distributed with means zero. Their variances and covariances are independent of t. The M_{it} and y_{it} are independent of each other. Define $\bar{X}_i = \Sigma_t X_{it}/N$ the arithmetic mean of X_{it} and $x_{it} = X_{it} - \bar{X}_i$ the deviation from the mean. Then $a_{ij} = \Sigma_t x_{it} x_{jt}/(N - 1)$ gives the variances and covariances of the observations. We want to determine the rank of the matrix of the variances and covariances of M_{it}.

Now assume that $||V_{ij}||$ is an estimate of the variance-covariance matrix of the error terms or "disturbances" y_{it}. The elements of this matrix are distributed according to the Wishart distribution and are independent of the M_{it}. They can be estimated as deviations from polynomial trends, as deviations from Fourier series, by the Variate Difference Method, etc. The estimates could also be based upon a priori knowledge if for instance the y_{it} are interpreted as errors of measurement. Assume that the estimate is based upon N' observations.

[1] The author is much obliged to Professors W. G. Cochran (Iowa State College), H. Hotelling (Columbia University), T. Koopmans (University of Chicago) and A. Wald (Columbia University) for advice and criticism with this paper. He has also profited by reading the unpublished paper: "On the Validity of an Estimate from a Multiple Regression Equation" by F. V. Waugh and R. C. Been which deals in part with a problem related to the one presented here.

Journal Paper No. J-1323 of the Iowa Agricultural Experiment Station, Ames, Iowa. Project No. 730.

Form the determinantal equation:

$$| a_{ij} - \lambda V_{ij} | = 0. \tag{1}$$

Apart from sampling fluctuations there should be r solutions $\lambda = 1$ of equation (1) if there are r independent linear relationships between the M_{it}. The rank of the variance-covariance matrix of M_{it} is then $M - r$. Following a suggestion of P. L. Hsu [2] made on the basis of the earlier work of R. A. Fisher [3] we form the test function

$$\Lambda_r = (N - 1)(\lambda_1 + \lambda_2 \cdots + \lambda_r), \tag{2}$$

where λ_1 is the smallest root of (1), λ_2 the next smallest, etc. Hence (2) is the sum of the r smallest roots of equation (1). The hypothesis to be tested is that there are exactly r independent linear relationships between the systematic parts of our variables in the population. This quantity (2) is distributed like χ^2 with $r(N - M - 1 + r)$ degrees of freedom for large samples, i.e. if N' becomes large. It can be used for forming an opinion about the number of independent relationships existing among the systematic parts of our variables (M_{it}).

The importance of the question of the rank lies in the following: Sometimes we are not so much interested in making predictions as to estimate the "true" relationships which exist in the population which corresponds to our sample (Wald) [4]. Practically speaking, these relationships and their estimation are of great importance in economic statistics, as Haavelmo has shown [5]. But a knowledge of the rank i.e. the number of independent relationships existing between the systematic parts of the variables may also be of some significance for the problem of prediction. The inclusion of strongly correlated predictors cuts down on the number of degrees of freedom without contributing significantly to the reduction of the variance.

The remainder of this paper will be concerned with an attempt to estimate the relationships which in the population exist between the systematic parts of the variables. This is an extension of the work of T. Koopmans [6] and the author [7] who dealt with the special case in which there is only one relationship between the systematic parts.

Suppose that we decide that there are R independent relationships among the systematic parts of our variables

$$k_{v0} + \sum_j k_{vj} M_{jt} = f_{vt} = 0; \qquad v = 1, 2, \cdots, R, \; t = 1, 2, \cdots, N. \tag{3}$$

We desire to obtain estimates of these relationships. Our purpose here is not prediction but estimation of the structural coefficients k_{vj}.

The method of maximum likelihood leads to the method of least squares if we treat the V_{ij} as constants. This is again permissible if N' is large and our estimates of the V_{ij} become reasonably accurate. We have to minimize the following sum of squares

$$Q = \sum_t Q_t \tag{4}$$

where

$$Q_t = \sum_i \sum_j V^{ij}(x_{it} - m_{it})(x_{jt} - m_{jt}), \tag{5}$$

where $|| V^{ij} || = || V_{ij} ||^{-1}$, the inverse of the variance-covariance matrix of the errors. We also define $m_{it} = M_{it} - \bar{M}_i$, $(t = 1, 2, \cdots, N)$ where $\bar{M}_i$ is the mean of M_{it}.

If there are R relationships (3) they can be written by using only $R(M - R)$ coefficients $k_{vj}(j = 1, 2 \cdots M)$, if we disregard the constant terms k_{0v}, because we are now dealing with deviations from means. We can for instance express the first $(M - R)$ variables m_{it} in terms of the last R variables m_{it}. Hence, we have to impose R^2 conditions upon the MR coefficients $k_{vj}(j = 1, 2, \cdots, M)$ appearing in (3).

We impose $R(R + 1)/2$ conditions as follows

$$\Sigma_i \Sigma_j k_{vi} k_{wj} V_{ij} = g_{vw} = \delta_{vw}, \tag{6}$$

where δ_{vw} is a Kronecker delta. These conditions orthogonalize and normalize the coefficients k_{vj}. We have now to adjust the Q_t as given in (5) under the conditions (6) by determining appropriate m_{it}. This is a problem of restricted minima.

We introduce a new function

$$F_t = Q_t - \sum_v \mu_{vt} f_{vt}, \tag{7}$$

where the μ_{vt} are Lagrange multipliers. Differentiating with respect to m_{it} and setting equal to zero we get the solution:

$$\sum_j V^{ij}(x_{jt} - m_{jt}) = \sum_v \mu_{vt} k_{vi}; \qquad (i = 1, 2, \cdots, M); \tag{8}$$

or, solving for $x_{it} - m_{it}$

$$x_{it} - m_{it} = \sum_v \sum_j \mu_{vt} V_{ij} k_{vj}; \qquad i = 1, 2, \cdots, M. \tag{9}$$

Multiplying (9) by k_{vi} and summing we get

$$\mu_{vt} = \Sigma_j k_{vj} x_{jt}. \tag{10}$$

Hence we have

$$Q_t = \sum_v \mu_{vt}^2 = \sum_v \left(\sum_j k_{vj} x_{jt} \right)^2. \tag{11}$$

Now we dispose of the remaining $R(R - 1)/2$ conditions

$$\sum_t \mu_{vt} \mu_{wt} = h_{vw} = 0, \qquad v \neq w. \tag{12}$$

We have to maximize Q under the R^2 conditions (6) and (12). This is done by finding the appropriate k_{vj}.

We form a new expression

$$G = Q + \Sigma_v \Sigma_w \beta_{vw} h_{vw} - \Sigma_v \Sigma_w \alpha_{vw} g_{vw} \tag{13}$$

where the α_{vw} and β_{vw} $(v \neq w)$ are again Lagrange multipliers and $\beta_{vv} = 0$. Because of considerations of symmetry we have: $\alpha_{vw} = \alpha_{wv}$ and $\beta_{vw} = \beta_{wv}$ Differentiating with respect to k_{vi} and setting equal to zero we get the condition

$$\begin{aligned} &\Sigma_t(\Sigma_j k_{vj} x_{jt}) x_{it} + \Sigma_w \beta_{vw} \Sigma_t(\Sigma_j k_{wj} x_{jt}) x_{it} \\ &\quad = \Sigma_w \alpha_{vw} \Sigma_j V_{ij} k_{wj}, \qquad v = 1, 2, \cdots, R, \qquad i = 1, 2, \cdots, M. \end{aligned} \tag{14}$$

Multiplying by k_{vi} and summing we get

$$\Sigma_t \mu_{vt}^2 = \alpha_{vv}. \tag{15}$$

Multiplying by k_{zi} $(z \neq v)$ and summing we have

$$\beta_{vz} \Sigma_t \mu_{zt}^2 = \alpha_{vz} \qquad (v \neq z). \tag{16}$$

Both (15) and (16) follow from conditions (6) and (12).

Exchanging the role of v and z in (16) we have also

$$\beta_{vz} \Sigma_t \mu_{vt}^2 = \alpha_{vz} \qquad (v \neq z). \tag{17}$$

Hence we have $\alpha_{vz} = \beta_{vz} = 0$, if $v \neq w$. Inserting these results in (14) we get a system of linear and homogeneous equations in the unknown coefficients k_{vj}. The determinant of the system must be equal to zero in order to yield non-trivial solutions. Trivial solutions are not admitted because of (6). Hence the α_{vv} are simply the roots k of the equation $| \Sigma_t x_{it} x_{jt} - k V_{ij} | = 0$.

Introducing

$$\lambda_v = \alpha_{vv}/(N - 1), \tag{18}$$

expression (14) becomes actually the determinantal equation (1). This expression can be used to find the R smallest latent roots λ_v and the corresponding characteristic vectors k_{vj} by Hotelling's methods [8].

The constants of the equation (3) are finally determined by the condition that the optimum solutions have to go through the means of the variables

$$k_{v0} + \Sigma_j k_{vj} \bar{X}_j = 0. \tag{19}$$

The distribution of the variances and covariances of the observations has recently been established by T. W. Anderson and M. A. Girshick for the cases $R = M - 1$ and $R = M - 2$ [9].

REFERENCES

[1] R. Frisch: *Statistical Confluence Analysis by Means of Complete Regression Systems*, Oslo, 1934.

[2] P. L. Hsu: "On the problem of rank and the limiting distribution of Fisher's test function," *Annals of Eugenics*, Vol. 11 (1941), pp. 39, ff.

[3] R. A. Fisher: "The statistical utilization of multiple measurements," *Annals of Eugenics*, Vol. 8 (1938), pp. 376 ff.

[4] A. WALD: "The fitting of straight lines if both variables are subject to error," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 284 ff.
[5] T. HAAVELMO: "The probability approach in econometrics," *Econometrica*, Vol. 12 (1944), Supplement.
[6] T. KOOPMANS: *Linear Regression Analysis in Economic Time Series*, Haarlem, 1937.
[7] G. TINTNER: "An application of the variate difference method to multiple regression," *Econometrica*, Vol. 12 (1944), pp. 97 ff.
[8] H. HOTELLING: "Simplified calculation of principal components," *Psychometrica*, Vol. 1 (1936), pp. 27 ff.
[9] T. W. ANDERSON AND M. A. GIRSHICK: "Some extensions of the Wishart distribution," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 354 ff.

NOTE ON THE DISTRIBUTION OF THE SERIAL CORRELATION COEFFICIENT[1]

BY WILLIAM G. MADOW

Bureau of the Census

The distribution of the serial correlation coefficient when $\rho = 0$ has been previously obtained.[2] The purpose of this note is to derive the distribution of the serial correlation coefficient, using the circular definition, when $\rho \neq 0$.

Let us assume that the random variables $x_1, \cdots, x_N$ have a joint normal distribution[3] $p(x_1, \cdots, x_N \mid A, B, \mu)$ where

$$\log p(x_1, \cdots, x_N \mid A, B, \mu) = \log K_1 - \tfrac{1}{2}\left[A \sum_i (x_i - \mu)^2 + 2B \sum_i (x_i - \mu)(x_{i+L} - \mu)\right]$$

the term in the bracket is positive definite, K_1 is independent of the x_i and if $i + L > N$ then $x_{i+L} = x_{i+L-N}$. It is then clear that $\bar{x}$, V_N, and ${}_LC_N$, where $\bar{x}$ is the arithmetic mean, $V_N = \sum_i (x_i - \bar{x})^2$ and

$${}_LC_N = \sum_i (x_i - \bar{x})(x_{i+L} - \bar{x})$$

are sufficient statistics with respect to the estimation of μ, A, and B.

Let $V_N \, {}_LR_N = {}_LC_N$ define ${}_LR_N$, the serial correlation coefficient. Then if

[1] Presented at a meeting of the Cowles Commission for Economic Research in Chicago, January 31, 1945.

[2] See R. L. Anderson, "*Distribution of the serial correlation coefficient*", pp. 1–13 and T. Koopmans, "Serial correlation and quadratic forms in normal variables", pp. 14–33, *Annals of Math. Stat.*, Vol. XIII, No. 1, March, 1942.

[3] The expression $p(\xi_1, \cdots, \xi_m \mid \theta_1, \cdots, \theta_q)$ means the probability density or the distribution of the random variables $\xi_1, \cdots, \xi_m$ for the given values of the parameters $\theta_1, \cdots, \theta_q$. When used as an index of summation or multiplication, the letter i will assume all values from 1 through N.