A MEASURE OF PREDICTIVE PRECISION IN REGRESSION ANALYSIS

By H. Linhart

South African Council for Scientific and Industrial Research, Johannesburg

1. Introduction and summary. It has been suggested ([5], [6]) to use the expected value, $E(l)$, of the length, l, of a confidence interval for the variable to be predicted as a measure of precision of prediction in a regression analysis. The measure is relevant only if the predictor variables are random.

Criticism of this particular choice of a measure of precision usually centres about the following questions:

a. Why is the measure based on *this* and no other system of confidence intervals; what is known about optimality of this system. (The system referred to will be described in Section 2.)

b. Why is it based on the physical length of the intervals and not on Neyman's "shortness"?

c. If it is agreed that it is based on l, what justifies the choice of $E(l)$?

In the following a few points are raised which are, of course, not sufficient to prove that $E(l)$ is the only possible choice for a measure of precision, but which indicate that the intuitive choice is not altogether unreasonable.

Not much can be said about a. It turns out, in Section 3, that the confidence limits discussed here are unbiased, but nothing about optimality in any sense is known to the author.

To throw some light on b, Neyman's shortness of the system of confidence intervals used is calculated in Section 3. A parameter enters the problem which makes it impossible to use Neyman's shortness as an overall measure of precision.

With regard to c, one may argue that l is a random variable and if a single measure of precision is needed a single characteristic of its distribution must be used. Obvious possibilities are the mean and the median. The distribution of l is obtained in Section 4, and it becomes apparent that the use of the median would involve heavy numerical calculations.

2. A system of confidence intervals for y_0. In all $n + 1$, independent vectors enter the problem. All vectors have the same $(k + 1)$-dimensional normal distribution with mean vector μ—here without loss of generality assumed to be the zero vector—and with unknown positive definite covariance matrix Λ. It is assumed that $n > k + 1$. The first n vectors, $(x_{0\nu}, x_{1\nu}, \cdots, x_{k\nu})$, $\nu = 1, 2, \cdots, n$, are used to estimate μ and Λ. These n vectors are called "the sample". Only the last k components of the $(n + 1)$st vector, $(y_0, y_1, \cdots, y_k)$, are known. Using the information on the structure of the distribution of $(y_0, y_1, \cdots, y_k)$, which was provided by the first n vectors, a confidence interval, corresponding to the confidence coefficient $1 - \alpha$, for a hypothetical future observation y_0 is given. The length of the confidence interval is a random variable; its expectation can be used as

Received December 26, 1958; revised October 22, 1959.

measure of precision of prediction. A system of confidence intervals will now be described and studied; $E(l)$ for this system will be computed and discussed.

The following notation will be used: $L \equiv (l_{ij})$, is the $(k+1) \times (k+1)$ covariance matrix in the sample,

$$nl_{ij} = \sum_{\nu=1}^{n} (x_{i\nu} - \bar{x}_i)(x_{j\nu} - \bar{x}_j), \qquad i, j = 0, 1, \cdots, k.$$

The $k \times k$ submatrix of L comprising only the elements l_{ij} with $i, j = 1, 2, \cdots k$, is denoted by L_0 and $L_0^{-1} \equiv (l_0^{ij})$ is written for $(L_0)^{-1}$. A completely analogous notation is used for the covariance matrix in the population $\Lambda \equiv (\lambda_{ij})$. Determinants are written with vertical strokes. The double tailed 100α per cent point of Student's distribution with ν degrees of freedom will be denoted by $t_\alpha^{(\nu)}$. The predicted value of y_0 is $\hat{y}_0 = \hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i(y_i - \bar{x}_i)$, where $\hat{\alpha}$ and $\hat{\beta}_i$, $i = 1, 2, \cdots, k$, are the least square estimates of the regression coefficients ([2], p. 552).

The confidence intervals for y_0 which are used here are those which are usually obtained under the assumption that $x_{1\nu}, x_{2\nu}, \cdots, x_{k\nu}$ and $y_1, y_2, \cdots, y_k$ are *fixed* (*nonrandom*) variables ([8], p. 305). Under that assumption, $\hat{y}_0 - y_0$ is $N[0,(1 + n + T) \mid \Lambda \mid / n \mid \Lambda_0 \mid]$, where

$$T = \sum_{i,j=1}^{k} (y_i - \bar{x}_i)(y_j - \bar{x}_j) l_0^{ij}; \tag{1}$$

and

$$u = \frac{n \mid \Lambda_0 \mid \mid L \mid}{\mid \Lambda \mid \mid L_0 \mid}, \qquad 0 \leqq u, \tag{2}$$

is, independently of $\hat{y}_0 - y_0$, distributed as χ^2 with $n - k - 1$ degrees of freedom. It follows that

$$t = (\hat{y}_0 - y_0) \left[\frac{\mid L_0 \mid (n - k - 1)}{\mid L \mid (1 + n + T)} \right]^{\frac{1}{2}} \tag{3}$$

has Student's distribution with $n - k - 1$ degrees of freedom, and that a confidence interval for y_0 is given by

$$\hat{y}_0 - t_\alpha^{(n-k-1)} \left[\frac{\mid L \mid (1 + n + T)}{\mid L_0 \mid (n - k - 1)} \right]^{\frac{1}{2}} \leqq y_0 \leqq \hat{y}_0 + t_\alpha^{(n-k-1)} \cdot \left[\frac{\mid L \mid (1 + n + T)}{\mid L_0 \mid (n - k - 1)} \right]^{\frac{1}{2}}. \tag{4}$$

The distribution of t, (3), is independent of L_0 and T, and remains therefore unchanged if L_0 is a random matrix and T a random variable. The probability that the intervals (4) cover y_0 is therefore [3] even in the random case equal to $1 - \alpha$.

In Section 3 it will be shown that these confidence intervals are unbiased, but nothing about optimality in any sense is known to the author.

It might also be noted here that the intervals (4) are not confidence intervals in the classical sense, as y_0 is a random variable and not a parameter. Sometimes intervals of this kind are called "prediction intervals" or "quasi confidence intervals".

3. The shortness of the confidence intervals used. The shortness, in the sense of Neyman, is measured by the probability that the confidence intervals cover a value which is different from the true value of the parameter ([9], p. 371). One must therefore find the probability that the intervals (4) cover $y_0 + \delta$, where δ is any constant.

Recalling how the confidence intervals (4) have been obtained, one may see that $\hat{y}_0 - y_0 - \delta$ is, conditionally, given L_0 and T,

$$N[-\delta, (1 + n + T)|\Lambda|/n|\Lambda_0|].$$

It follows ([1], p. 113) that $(\hat{y}_0 - y_0 - \delta)^2 n|\Lambda_0|/(1 + n + T)|\Lambda|$ has the noncentral χ^2-distribution with 1 degree of freedom and noncentrality parameter

$$\tau^2 = \delta^2 n|\Lambda_0|/(1 + n + T)|\Lambda|. \tag{5}$$

The square of

$$t' = (\hat{y}_0 - y_0 - \delta)\left[\frac{|L_0|(n-k-1)}{|L|(1+n+T)}\right]^{\frac{1}{2}}, \tag{6}$$

has then ([1], p. 114), again conditionally, the noncentral F-distribution with $f_1 = 1$ and $f_2 = n - k - 1$ degrees of freedom,

$$f(t'^2) = \exp\{-\tau^2/2\}\frac{f_1}{f_2}\sum_{\nu=0}^{\infty}\frac{(\tau^2/2)^\nu (f_1 t'^2/f_2)^{f_1/2+\nu-1}}{\nu!\mathrm{B}(f_1/2+\nu, f_2/2)(1+f_1 t'^2/f_2)^{(f_1+f_2)/2+\nu}}, \tag{7}$$

$$0 \leqq t'^2.$$

The conditional probability, given T, that the confidence intervals (4) cover $y_0 + \delta$, or the conditional shortness, is therefore

$$s(\delta \mid T) = \int_0^{[t_\alpha(n-k-1)]^2} f(t'^2)\, dt'^2. \tag{8}$$

The unconditional shortness, $s(\delta)$, is the expectation of $s(\delta \mid T)$. It is thus necessary to obtain at first the density function of T.

Hsu has shown ([4], p. 235, equation 12); see also ([1], p. 114)) that $(n - k)T/k$ has, conditionally, given $(y_1, y_2, \cdots, y_k)$, the noncentral F-distribution (7) with $f_1 = k$ and $f_2 = n - k$ degrees of freedom, and with noncentrality parameter $\tau^2 \equiv 2\lambda$,

$$\lambda = (n/2)\sum_{i,j=1}^{k} y_i y_j \lambda_0^{ij}. \tag{9}$$

The density function of T is the expectation over the variables $y_1, y_2, \cdots, y_k$, of the conditional density function. Now the vector $(y_1, y_2, \cdots, y_k)$ is $N(0, \Lambda_0)$,

and it is well known, that $2\lambda/n$ has a χ^2-distribution with k degrees of freedom ([2], p. 319, example 15). It is then easily verified that

$$E(\lambda^\nu \exp\{-\lambda\}) = n^\nu \Gamma(k/2 + \nu)/(n + 1)^{k/2+\nu}\Gamma(k/2). \tag{10}$$

The density function of T is found to be

$$f(T) = \frac{(1 + n)^{(n-k)/2} T^{k/2-1}}{\mathrm{B}[k/2, (n - k)/2](1 + n + T)^{n/2}}, \qquad 0 \leqq T. \tag{11}$$

The substitution

$$v = (1 + n)/(1 + n + T), \qquad 0 \leqq v \leqq 1, \tag{12}$$

shows that v has a Beta-distribution with k and $n - k$ degrees of freedom.

For the expectation of $s(\delta \mid T)$ with respect to T one needs, as may be seen from (7) and (8), the expectation of $(\tau^2/2)^\nu \exp\{-\tau^2/2\}$, where τ^2, as given by (5), is a multiple of v. Using a well known integral representation for the confluent hypergeometric function ([7], p. 87), one obtains easily

$$\begin{aligned} &E[(\tau^2/2)^\nu \exp\{-\tau^2/2\}] \\ &= \frac{\eta^\nu \Gamma[(n - k)/2 + \nu]\Gamma(k/2)\Gamma(n/2)}{\Gamma(n/2 + \nu)\Gamma[(n - k)/2]} F[(n - k)/2 + \nu; n/2 + \nu; -\eta], \end{aligned} \tag{13}$$

where

$$\eta = \delta^2 n \mid \Lambda_0 \mid /2(1 + n)\mid \Lambda \mid. \tag{14}$$

The shortness may then be calculated. It is

$$\begin{aligned} s(\delta) = \int_0^{[t_\alpha(n-k-1)]^2} &\frac{\Gamma(n/2)}{(n - k - 1)\Gamma[(n - k - 1)/2]\Gamma[(n - k)/2]} \\ \cdot \sum_{\nu=0}^{\infty}\sum_{\mu=0}^{\infty} &\frac{(-1)^\mu \eta^{\nu+\mu}[t'^2(n - k - 1)]^{\nu-\frac{1}{2}}}{\nu!\mu![1 + t'^2/(n - k - 1)]^{(n-k)/2+\nu}} \\ &\cdot \Gamma[(n - k)/2 + \nu]\Gamma[(n - k)/2 + \nu + \mu] \\ &\cdot \Gamma(1/2 + \nu)\Gamma(n/2 + \nu + \mu). \end{aligned} dt'^2 \tag{15}$$

The function $s(\delta)$ is difficult to study, and, in addition to that, besides $\mid \Lambda \mid/\mid \Lambda_0 \mid$, k and n, another parameter, δ, enters the problem; it seems impossible to base an overall measure of precision of prediction on $s(\delta)$, unless it is weighted by some arbitrary weight function of δ.

Remembering that $s(\delta)$ is the expectation of $s(\delta \mid T)$, and realising that $1 - s(\delta \mid T)$ is nothing else than the power function of Student's t-test, one may draw some conclusions about the properties of $s(\delta)$.

The symmetrical double tailed t-test is unbiased; $s(\delta \mid T)$ has, therefore, for fixed k, n and $\mid \Lambda \mid/\mid \Lambda_0 \mid$, for each finite T an absolute maximum at $\delta = 0$; $s(\delta)$ has, therefore, for fixed k, n and $\mid \Lambda \mid/\mid \Lambda_0 \mid$, also a maximum at $\delta = 0$. The system of confidence intervals used is therefore unbiased in Neyman's sense.

If $n_1 > n_2$, the t-test with n_1 degrees of freedom is uniformly more powerful

than the corresponding test with n_2 degrees of freedom. For each finite value of T, and for fixed n, δ and $|\Lambda|/|\Lambda_0|$, $s(\delta \mid T)$ is thus strictly monotonically increasing with k. For fixed n, δ and $|\Lambda|/|\Lambda_0|$, $s(\delta)$ is, therefore, strictly monotonically increasing with k. This means that, if $s(\delta)$ were used as a measure of precision, the inclusion of more variables in a regression analysis which is *not* accompanied by a decrease in the residual variance $|\Lambda|/|\Lambda_0|$ would result in a uniform (in δ) deterioration of precision of prediction. This is a property which a measure based on $s(\delta)$ would share with the measure $E(l)$.

To defend the use of $E(l)$, work by S. S. Wilks [10] should be mentioned here, which uses "average shortness" as a large sample optimum property of systems of confidence intervals. Shortness is here physical shortness, and not Neyman's shortness; average shortness is thus completely analogous to $E(l)$.

Neyman ([9], p. 370) writes a few sentences about physical shortness as an optimum property and remarks in conclusion that, "The above statement may appeal to intuition, but it is obviously too vague to be used in practice." It would appear that he does not condemn the idea as a whole, but only stresses practical difficulties.

One must, however, also mention two drawbacks of physical shortness: it is not invariant under monotone transformations; and it covers only precision, but not accuracy, a physically short interval may be relatively bad if it contains values which are very different from the true value of the estimated parameter.

4. The distribution of the length of the confidence intervals. From (4) one may see that the length of the confidence interval, corresponding to the confidence coefficient $1 - \alpha$, is given by

$$l = 2t_\alpha^{(n-k-1)} \left[\frac{|L|\ (1 + n + T)}{|L_0|\ (n - k - 1)}\right]^{\frac{1}{2}}. \tag{16}$$

It is convenient to obtain the distribution of

$$z = \tfrac{1}{2} \cdot \frac{n\,|\Lambda_0|\,|L|}{|\Lambda|\,|L_0|} \cdot \frac{1 + n + T}{1 + n} = \tfrac{1}{2} \cdot u \cdot \frac{1}{v} \tag{17}$$

at first, where u is given by (2) and v by (12). In Section 2 it was mentioned that u has, conditionally, given L_0, a χ^2-distribution with $n - k - 1$ degrees of freedom. It follows immediately, for the unconditional case, that u is distributed independently of T according to the same distribution. The simultaneous density function of u and v is, therefore,

$$\begin{aligned} &f(u, v) \\ &\propto u^{(n-k-1)/2-1} \exp\{-u/2\} v^{(n-k)/2-1}(1 - v)^{k/2-1}, \quad 0 \leqq u, \quad 0 \leqq v \leqq 1. \end{aligned} \tag{18}$$

Substituting $z = u/2v$, $u/2 = t$ one has

$$f(z) \propto z^{-(n-k)/2-1} \int_0^z \exp\{-t\} t^{n-k-\frac{3}{2}}(1 - t/z)^{k/2-1}\, dt. \tag{19}$$

Using the integral representation of the confluent hypergeometric function ([7], p. 87) it is easily deduced that

$$f(z) = \frac{B(n-k-1/2, k/2)\exp\{-z\}}{B[(n-k)/2, k/2]\Gamma[(n-k-1)/2]} z^{(n-k-1)/2-1} F(k/2; n-k/2-\tfrac{1}{2}; z), \qquad 0 \leqq z. \tag{20}$$

The distribution of l may be obtained by substituting

$$l = 2^{\frac{1}{2}} t_\alpha^{(n-k-1)} z^{\frac{1}{2}} \left[\frac{(1+n)\,|\,\Lambda\,|}{n(n-k-1)\,|\,\Lambda_0\,|}\right]^{\frac{1}{2}}. \tag{21}$$

One may see that this density is not of a very convenient form; tables of its distribution function do not exist. Using the median of this distribution would thus involve heavy numerical calculations. One may, for a comparison, note the comparatively simple form of the mean

$$E(l) = 2^{\frac{1}{2}} t_\alpha^{(n-k-1)} \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]} \left[\frac{(1+n)\,|\,\Lambda\,|}{n(n-k-1)\,|\,\Lambda_0\,|}\right]^{\frac{1}{2}}. \tag{22}$$

5. Acknowledgment. I am indebted to the referee and to Prof. W. Kruskal for many critical remarks and helpful suggestions which, in particular, led to a much shorter proof of the distribution of l.

REFERENCES

[1] T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, New York, John Wiley and Sons, 1958.

[2] HARALD CRAMÉR, *Mathematical Methods of Statistics*, Princeton, Princeton University Press, 1946.

[3] EDWIN L. CROW, "Generality of confidence intervals for a regression function," *J. Amer. Stat. Assn.*, Vol. 50 (1955), pp. 850–853.

[4] P. L. HSU, "Notes on Hotelling's generalized T," *Ann. Math. Stat.*, Vol. 9 (1938), pp. 231–243.

[5] H. LINHART, "Critère de sélection pour le choix des variables dans l'analyse de régression," *Revue suisse d'Economie politique et de Statistique*, Vol. 94 (1958), pp. 202–232.

[6] H. LINHART, "A criterion for selecting variables in a regression analysis," *Psychometrika*. Vol. 25 (1960), pp. 45–58.

[7] WILHELM MAGNUS AND FRITZ OBERHETTINGER, *Formulas and theorems for the functions of mathematical physics*," Chelsea, New York, 1954.

[8] ALEXANDER MCFARLANE MOOD, *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950.

[9] J. NEYMAN, "Outline of a theory of statistical estimation based on the classical theory of probability," *Phil. Trans. Roy. Soc. London, Ser. A*, Vol. 236 (1937), pp. 333–380.

[10] S. S. WILKS, "Shortest average confidence intervals from large samples," *Ann. Math. Stat.*, Vol. 9 (1938), pp. 166–175.