Notes

Errors of Prediction in Multiple Regression with Stochastic Regressor Variables

D. KERRIDGE

University of Aberdeen

This paper investigates the errors of prediction of multiple regression equations, in which the regressor variables are regarded as drawn at random from a multivariate normal population.

In most theoretical treatments of multiple regression it is assumed that one or more "dependent" variables are to be predicted, given the observed values of a number of "independent" variables. The independent or regressor variables are treated as constants. However, in many applications, it is more reasonable to regard them as random variables, drawn, for example, from a multivariate normal population. In this case the term "independent" variables is confusing, and it seems better to call them regressor variables. It has been suggested by Ehrenberg (1963) that regression or stochastic regressor variables is useless. This paper shows that although such regression has it limitations, especially if the number of regressor variables is large, it may be useful if the limitations are understood.

The particular case to be discussed is that in which the regressor variables are drawn from a multivariate normal population.

The k-dimensional vectors $\mathbf{X}_1 , \mathbf{X}_2 , \cdots \mathbf{X}_i \cdots \mathbf{X}_n$ are given as a random sample from the distribution $N(\mathbf{\Lambda}, \mathbf{\Sigma})$. Corresponding to these vectors we observe $y_1 , y_2 , y_3 \cdots y_n$. We assume that for each:

$$y_i = \beta + \mathbf{X}_i'\mathbf{A} + \epsilon_i$$

where β, $\mathbf{A}$ are unknown constants, and the ϵ_i are distributed $N(0, \sigma^2)$ independently of each other and of the $\mathbf{X}_i$. This is exactly equivalent to assuming that the partitioned vector $\{y_i : X_i'\}$ is drawn from a multivariate normal population.

Let a bar over any variable denote averaging over $i = 1, 2 \cdots n$ only, and let

$$\mathbf{S} = \sum_{i,j} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})$$

Then if a further vector $\mathbf{X}_{n+1}$, from the same multivariate normal population is observed, we predict y_{n+1} just as we would if the $\mathbf{X}_i$ were fixed, that is, by

$$\hat{y}_{n+1} = \bar{y} + (\mathbf{X}_{n+1} - \bar{\mathbf{X}})'\hat{\mathbf{A}}$$

Received December 1965.

where

$$\hat{\mathbf{A}} = \mathbf{S}^{-1} \sum_{i,j} y_i(\mathbf{X}_i - \bar{\mathbf{X}}).$$

The error of prediction is just

$$\begin{aligned} y_{n+1} - \hat{y}_{n+1} &= \beta + \mathbf{X}'_{n+1}\mathbf{A} + \epsilon_{n+1} - \bar{y} - (\mathbf{X}_{n+1} - \bar{\mathbf{X}})'\hat{\mathbf{A}} \\ &= (\mathbf{X}_{n+1} - \bar{\mathbf{X}})'(\mathbf{A} - \hat{\mathbf{A}}) + \epsilon_{n+1} + \bar{\epsilon} \end{aligned}$$

Each term is independent of each other term.

If we argue conditionally on $\mathbf{X}_1, \cdots, \mathbf{X}_{n+1}$ it is known that $\mathbf{A} - \hat{\mathbf{A}}$ is distributed according to the multivariate normal distribution $N(\mathbf{0}, \sigma^2\mathbf{S}^{-1})$. Hence $y_{n+1} - \hat{y}_{n+1}$ is distributed unconditionally like

$$z\sigma\left([\mathbf{X}_{n+1} - \bar{\mathbf{X}}]'\mathbf{S}^{-1}[\mathbf{X}_{n+1} - \bar{\mathbf{X}}] + 1 + \frac{1}{n}\right)^{\frac{1}{2}}$$

where z is $N(0, 1)$

But $(1 + 1/n)^{-1}(n - 1)[\mathbf{X}_{n+1} - \bar{\mathbf{X}}]'\mathbf{S}^{-1}[\mathbf{X}_{n+1} - \bar{\mathbf{X}}]$ is distributed like Hotelling (1931) T'^2, and so has the distribution of

$$(n - 1)\frac{k}{n - k}F_{k,n-k}$$

where $F_{k,n-k}$ is an F variate with k, $n - k$ degrees of freedom. Hence we may write $y_{n+1} - \hat{y}_{n+1}$ in the form

$$\begin{aligned} &z\sigma\left(1 + \frac{1}{n}\right)^{\frac{1}{2}}\left(\frac{\chi_k^2}{\chi_{n-k}^2} + 1\right)^{\frac{1}{2}} \\ &= z\sigma\left(1 + \frac{1}{n}\right)^{\frac{1}{2}}\left(\frac{\chi_{n-k}^2 + \chi_k^2}{\chi_{n-k}^2}\right)^{\frac{1}{2}} \end{aligned}$$

where χ_{n-k}^2 and χ_k^2 are independent χ^2 variables with $n - k$, k degrees of freedom.

Thus we have expressed the error of prediction as the ratio of a normally distributed random variable and the square root of an independent Beta-variate. From this the explicit form of the distribution may be deduced if required.

For practical purposes the expectation of the mean squared error may be of greater interest than its exact distribution. This may easily be seen to be given by

$$\sigma^2\left(1 + \frac{1}{n}\right)\left(\frac{n - 2}{n - k - 2}\right)$$

This provides a useful guide as to the number of regressor variables which can usefully be allowed in a multiple regression equation.

Similarly, if we consider the multivariate prediction problem, based on the model

$$\mathbf{Y}_i = \mathbf{B} + \mathbf{X}'_i\mathbf{A} + \mathbf{E}_i$$

where $\mathbf{E}_i$ is $N(\mathbf{0}, \mathbf{V})$, and $\mathbf{B}$, $\mathbf{A}$ are matrices, it may be shown without difficulty that the expected error of prediction is given by

$$E\{(Y_{n+1} - \hat{Y}_{n+1})(Y_{n+1} - \hat{Y}_{n+1})'\} = \mathbf{V}\left(1 + \frac{1}{n}\right)\left(\frac{n - 2}{n - k - 2}\right)$$

We may also determine the covariance between the errors of prediction of y_{n+1} and a still further value y_{n+2}, corresponding to an observed vector $\mathbf{X}_{n+2}$.

Adapting the previous argument we may show that

$$E\{(y_{n+1} - \hat{y}_{n+1})(y_{n+2} - \hat{y}_{n+2})\} = \sigma^2 E\{(X_{n+1} - \bar{X})'S^{-1}(X_{n+2} - \bar{X})\}$$

$$+ \frac{1}{n} = \sigma^2 E(\mathbf{X}_{n+1} - \Lambda)'\mathbf{S}^{-1}(\mathbf{X}_2 - \bar{\mathbf{X}})$$

$$- (\bar{\mathbf{X}} - \Lambda)'\mathbf{S}^{-1}(\mathbf{X}_{n+2} - \Lambda) + (\bar{\mathbf{X}} - \Lambda)'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \Lambda) + \frac{1}{n}$$

Owing to the independence of $\bar{\mathbf{X}}$, $\mathbf{X}_{n+1}$, $\mathbf{X}_{n+2}$, and $\mathbf{S}$, the first two terms have zero expectation. Further, the last term may be written in the form

$$\frac{k}{n(n - k)} F_{k,(n-k)}$$

where F is an F-variate with k, $n - k$ degrees of freedom. It follows that the covariance between the errors of prediction of y_{n+1} and y_{n+2} is just

$$\frac{\sigma^2}{n}\left(\frac{n - 2}{n - k - 2}\right)$$

Similarly in the case of a multivariate prediction equation, the expression is the same except that $\mathbf{V}$ replaces σ^2.

The errors of prediction of a multiple regression equation with constant regressor variables have been treated by Hooper and Zellner (1961).

Gorman and Toman (1966) have discussed the problem of selecting variables for a multiple regression equation using an approximate expression for the expected mean squared error of prediction due to C. L. Mallows (unpublished). The result of this paper provides an exact alternative to this, provided that multivariate normality may be assumed.

References

EHRENBERG, A. S. C., 1963. Bivariate regression analysis is useless, *Applied Statistics, 12*, 161–179.

GORMAN, T. W., and TOMAN, R. J., 1966. Selection of variables for fitting equations to data, *Technometrics, 8*, 27–51.

HOOPER, J. W. and ZELLNER, A., 1961. The error of forecast for multivariate regression models, *Econometrica, 29*, 544–555.

HOTELLING, H., 1931. The generalisation of student's ratio. *Ann. Math. Statist., 2*, 360–378.