Journal of Econometrics 4 (1976) 393–397.

THE USE OF DUMMY VARIABLES TO COMPUTE PREDICTIONS, PREDICTION ERRORS, AND CONFIDENCE INTERVALS

David S. SALKEVER*

Johns Hopkins University, Baltimore, MD 21205, U.S.A.

Received September 1975, revised version received January 1976

This paper presents a method for computing predictions, prediction error variances, and confidence intervals, which can be implemented with any regression program. It demonstrates that a regression estimated for an augmented data set, obtained by (1) combining n sample points with r forecast points, and (2) including r dummy variables (each equalling one only for the corresponding forecast point), will yield r dummy variable coefficients and variances which equal the corresponding prediction errors and prediction error variances. Since most programs lack special routines to calculate these magnitudes, while manual computation is cumbersome, the proposed method is of considerable practical value.

Multiple linear regression analysis is often applied to either time-series or cross-section data for the purpose of generating and testing predictions. Typically, this process involves three steps: (1) estimation of regression coefficients from a set of original data points, (2) computation of predictions and prediction errors for one or more additional data points, and (3) computation of estimated prediction error variances and the associated confidence intervals. With the aid of a suitable computer program, these tasks can be easily accomplished. Surprisingly, however, most packaged regression programs do not provide specific procedures for obtaining estimated prediction error variances or confidence intervals.[1] And the manual calculation of these variances is obviously cumbersome when the number of independent variables is large. (For example, with ten independent variables more than fifty terms must be calculated and summed to obtain the prediction error variance for a single data point.) Moreover, in the case of some packaged programs it is not even possible to calculate predictions or prediction errors without additional manual arithmetic.

*The author is assistant professor in the Department of Health Care Organization and the Department of Political Economy. The Johns Hopkins University. He is indebted to Richard Royall and Carl Christ for helpful suggestions during the preparation of this paper. Financial support for this research was provided by U.S. Department of Health, Education and Welfare Grant HS00110.

[1]For example, such widely used packages as SPSS, OSIRIS, RAPE, TSP, and BMD do not contain this feature.

In this note, I propose an alternative method for computing predictions, prediction errors, and their variances. The advantage of this method is that, in contrast to the usual approach, it can easily be implemented with any packaged regression program.

The proposed method is to estimate a single regression equation, including data from both the original data points and the additional data points, and including among the regressors a dummy variable for each additional data point which takes on a value of 1 for that data point and zero otherwise. The coefficients for each of these dummy variables will equal the prediction error for the corresponding data point, while the variances of these coefficients will equal the prediction error variances. The coefficients and associated variances for all other regressors are identical to those obtained from a regression employing only the original data points.[2]

The proof of these assertions is straightforward. In the standard regression model,

$$y = \sum_{i=1}^{k} B_i x_i + u,$$

the least-squares estimate of the coefficient vector B, obtained from m original data points, is given by

$$\hat{B} = (X'X)^{-1}X'Y,$$

where X is the $m \times k$ matrix of observations on the original regressors (the first column being a vector of 1's) and Y is the $m \times 1$ vector of observations on the dependent variable. The estimated variance–covariance matrix of $\hat{B}$ is $s^2(X'X)^{-1}$, where s^2, the unbiased estimate of the variance of u, is equal to the sum of squared residuals divided by $(m-k)$. Also, letting Z denote the $r \times k$ matrix of observations from r additional data points on the k original regressors, we note that the estimated prediction error variance for the jth additional data point is $s^2[1+Z_j(X'X)^{-1}Z_j']$, where Z_j is the jth row of Z.[3]

In the expanded regression, including the r additional data points and the r additional dummy variables, the $(m+r) \times (k+r)$ matrix of observations on the regressors may be written as

$$Q = \left[\begin{array}{c|c} X & 0 \\ \hline Z & I \end{array}\right], \tag{1}$$

[2]After the preparation of this paper, I discovered that the same method proposed here has also appeared in the statistical literature dealing with missing observations in the analysis of experimental data. [See, for example, Wilkinson (1960).] However, the contributors to this literature do not consider the application of this method in a prediction context. Moreover, they do not demonstrate (or even mention) the most practically significant of our results namely, the equivalence of the dummy variable coefficient variances and the prediction error variances.

[3]See Goldberger (1964, p. 169).

where 0 is an $m \times r$ null matrix, I is an $r \times r$ identity matrix, and X and Z have previously been defined. To obtain the $(m+r) \times 1$ vector of values for the dependent variable, denoted by P, we augment the $m \times 1$ vector of original observations, Y, by an $r \times 1$ vector of arbitrarily assigned values for the additional observations, denoted by T. (Reasons for choosing particular values for T are discussed below.) Finally, the least-squares coefficients in the expanded regression are

$$C = (Q'Q)^{-1}Q'P,$$

and their estimated variance–covariance matrix is $\bar{s}^2(Q'Q)^{-1}$, where $\bar{s}^2$ equals the sum of squared residuals divided by $(m+r)-(k+r)$.

From (1) it follows that

$$Q'Q = \left[\begin{array}{c|c} X'X+Z'Z & Z' \\ \hline Z & I \end{array}\right], \tag{2}$$

and employing the standard formula for the inverse of a partitioned matrix,[4] we have

$$(Q'Q)^{-1} = \left[\begin{array}{c|c} D^{-1} & -E(I-ZE)^{-1} \\ \hline -IZD^{-1} & I+IZD^{-1}Z'I \end{array}\right], \tag{3}$$

where

$$D = (X'X+Z'Z-Z'IZ) \quad \text{and} \quad E = (X'X+Z'Z)^{-1}Z'.$$

But then

$$D^{-1} = (X'X)^{-1},$$

and from the symmetry of $(Q'Q)^{-1}$ and $(X'X)^{-1}$, we may rewrite (3) as

$$(Q'Q)^{-1} = \left[\begin{array}{c|c} (X'X)^{-1} & -(X'X)^{-1}Z' \\ \hline -Z(X'X)^{-1} & I+Z(X'X)^{-1}Z' \end{array}\right]. \tag{4}$$

Finally, since

$$Q'P = \left[\begin{array}{c|c} X' & Z' \\ \hline 0' & I \end{array}\right]\left[\begin{array}{c} Y \\ \hline T \end{array}\right] = \left[\begin{array}{c} X'Y+Z'T \\ \hline T \end{array}\right], \tag{5}$$

it is obvious that

$$C = \left[\begin{array}{c} (X'X)^{-1}X'Y \\ \hline T-Z\hat{B} \end{array}\right]. \tag{6}$$

Thus,

$$C_i = \hat{B}_i, \qquad \text{for} \quad i = 1, \ldots, k.$$

[4]See Graybill (1969, p. 165).

And since the residuals for the r additional observations are identically zero while

$$(m+r)-(k+r) = m-k,$$

it must be true that

$$\bar{s}^2 = s^2.$$

It then follows from (4) that

$$\text{var}\,(C_i) = \text{var}\,(\hat{B}_i), \qquad \text{for} \quad i = 1, \ldots, k,$$

and that

$$\text{var}\,(C_{k+1}), \ldots, \text{var}\,(C_r),$$

which are the diagonal elements of the matrix $s^2[I+Z(X'X)^{-1}Z']$, are equal to the prediction error variances for the r additional data points. The off-diagonal elements of this matrix are the estimated covariances of the prediction errors while the covariance between $\hat{B}$ and the prediction errors is given by the matrix $s^2[-Z(X'X)^{-1}]$.

Finally, if T is defined as the actual dependent variable values for the r additional observations, then $C_{k+1}, \ldots, C_{k+r}$ are the prediction errors for these additional observations. In a forecasting situation, where these dependent variable values are not yet known, it may be desirable to set each element of T equal to the mean dependent variable value for the original data points, since the expanded regression will then yield the R^2 for the original regression. It might even be desirable to follow this procedure in all cases since the actual prediction errors (and predictions) can easily be calculated once the dummy variable coefficients have been obtained.

In addition to avoiding the use of specialized features which are not available in many regression programs, the method proposed here is very convenient. For any number of additional data points, a single regression will yield prediction errors (which can be subtracted from actual dependent variable values to obtain predictions) and their variances, from which confidence intervals can easily be computed. It will also yield estimated prediction error covariances, which can be used to test the hypothesis that the original and additional data points were both generated by the same structural relationship.[5] Finally, it will yield all of the standard results for the regression on the original data points except for the value of R^2 (which may also be obtained if T is chosen appropriately).

In short, it is clear that use of the proposed method would save time and effort in many circumstances.

[5]The details of this test are presented in Christ (1966, pp. 557–560).

References

Christ, C., 1966, Econometric models and methods (Wiley, New York).
Goldberger, A., 1964, Econometric theory (Wiley, New York).
Graybill, F., 1969, Introduction to matrices with applications in statistics (Wadsworth, Belmont, CA).
Wilkinson, G. N., 1960, Comparison of missing value procedures, The Australian Journal of Statistics 2, no. 2, Aug., 53–65.