STANDARD ERROR OF FORECAST IN MULTIPLE REGRESSION: PROOF OF A USEFUL RESULT

Joseph S. DeSalvo

April 1970

P-4365

STANDARD ERROR OF FORECAST IN MULTIPLE REGRESSION: PROOF OF A USEFUL RESULT

Joseph S. DeSalvo*

The RAND Corporation, Santa Monica, California

ABSTRACT

Proof that the standard errors of forecasting the dependent variable or the expected value of the dependent variable in a multiple regression reduce to very simple formulas when evaluated at the sample means of the independent variables. These simple formulas involve only knowledge of sample size and the standard error of estimate, the latter of which is typically printed out in computer regression routines. By using these results, one avoids the necessity of calculating the more complicated general formulas for the standard errors in those cases for which evaluation at the mean will suffice. Although the results are not surprising, the author has been unable to find a published proof.

---

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

STANDARD ERROR OF FORECAST IN MULTIPLE REGRESSION: PROOF OF A USEFUL RESULT

MOTIVATION

One often wishes to make predictions or forecasts based on a linear statistical model. Perhaps the most common technique uses a least-squares regression of a dependent variable on one or more independent variables. The estimated least-squares function is used to predict a value of the dependent variable for given values of the independent variable(s). In assessing the accuracy of the predictions thus obtained, one should calculate the standard error of forecast and use this statistic in hypothesis tests or confidence intervals about the prediction.

Now, in the simple linear regression model where there is only one independent variable, the standard error of forecast reduces to a very simple formula when it is evaluated at the mean of the independent variable. It is the purpose of this paper to show that in the multiple linear regression model the standard error of forecast reduces to the same simple formula when evaluated at the means of the independent variables. Although this is not a surprising result, the author has been unable to find a published proof.*

Since the standard error of forecast can be calculated for any value of the independent variable(s), why should one be interested in its value at the means of the independent variables? There are several reasons for this. As will be seen below, the general formula for the standard error of forecast is fairly complicated, particularly in the multivariate case. It requires the computation of a quadratic form involving the inverse of the matrix of sums and cross products of observations on the independent variables. Moreover, standard computer regression routines do not always provide the requisite matrix. However, the standard error of forecast reduces to a simple formula involving the standard error of estimate and the sample size when evaluated at

---

*A similar result holds for the standard error of predicting the expected value of the dependent variable, as will also be shown.

the sample means of the independent variables. Since the standard error of estimate is typically printed out in standard computer regression routines, it is an easy matter to calculate the required standard error of forecast evaluated at the means of the independent variables. Moreover, one may be more interested in predictive accuracy at the means of the independent variables than elsewhere. This is especially true if one is trying to compare the predictive accuracy of different estimated functions in an attempt to find the best predicting equation. Also, since the standard error of forecast is a minimum at the sample means of the independent variables, if it is used in forming the limits of an acceptance region for hypothesis testing of the prediction and if the test statistic falls within this truncated region, it is then unnecessary to use the larger region.* Finally, the simple formula is a good approximation to the correct formula other than at the means of the independent variables when the observed values of the independent variables are dispersed widely from their respective means; it becomes a better approximation the wider the dispersion.

PROBLEM

In the simple linear regression model

$$(1) \quad Y_i = \alpha + \beta X_i + u_i \qquad i = 1, \ldots, n$$

$$(2) \quad E(u_i) = 0 \qquad \text{all } i$$

$$(3) \quad E(u_i u_j) = \begin{cases} \sigma^2 & i = j;\ i,j = 1, \ldots, n \\ 0 & i \neq j;\ i,j = 1, \ldots, n \end{cases},$$

the best linear unbiased estimate of Y for a given X is given by

$$(4) \quad \hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0 \quad ,$$

where $\hat{\alpha}$ and $\hat{\beta}$ are least-squares estimators of α and β in (1), X_0 is the particular value of X for which we wish to predict Y, and $\hat{Y}_0$ is that prediction.

---

*See, on this point, Carl F. Christ, Econometric Models and Methods (New York: John Wiley, 1966), p. 557.

It is well known* that the error variance of predicting Y, $\text{Var}(Y_0 - \hat{Y}_0)$, and the error variance of predicting the mean of Y, $\text{Var}\ (EY_0 - \hat{Y}_0)$, are -- in the simple model of (1), (2), and (3) -- given by

$$(5) \qquad \text{Var}(Y_0 - \hat{Y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} \right]$$

$$(6) \qquad \text{Var}(EY_0 - \hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} \right] .$$

When the standard error of estimate, in place of σ,

$$\hat{\sigma} = \left[\frac{\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2}{n-2} \right]^{\frac{1}{2}}$$

is inserted in (5) and square root taken, the resulting expression is known as the standard error of forecasting or predicting Y. The standard error of forecasting the expected value of Y may be obtained similarly. It is these standard errors that are used in hypothesis testing and in obtaining confidence intervals about the predictions.

Now, it is easy to see that, if we evaluate each of these variances at the sample mean of X, we get

$$(7) \qquad \text{Var}(Y_0 - \hat{Y}_0)\Big|_{X_0 = \bar{X}} = \sigma^2 (1 + \frac{1}{n})$$

and

$$(8) \qquad \text{Var}(EY_0 - \hat{Y}_0)\Big|_{X_0 = \bar{X}} = \sigma^2 (\frac{1}{n}) \quad .$$

If we again substitute $\hat{\sigma}$ for σ in (7) and (8), we may use the square root of these expressions for hypothesis tests and confidence intervals. They involve knowing only the sample size and the standard error of estimate, the latter of which is typically printed out in standard computer regression routines.

---

*See, for example, J. Johnston, Econometric Methods (New York: McGraw-Hill, 1963), pp. 36-37.

In the general linear regression model

$$(9) \qquad Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i , \qquad i = 1, \ldots, n$$

for which (2) and (3) hold and in addition the matrix X of sample observations on the k independent variables with $X_{1i} \equiv 1$,

$$X = \begin{bmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{bmatrix}$$

has rank $k < n$, the error variances are

$$(10) \qquad \mathrm{Var}(Y_0 - \hat{Y}_0) = \sigma^2 \left[1 + X_0'(X'X)^{-1}X_0\right]$$

$$(11) \qquad \mathrm{Var}(EY_0 - \hat{Y}_0) = \sigma^2 \left[X_0'(X'X)^{-1}X_0\right] ,$$

where $\hat{Y}_0 = X_0'\hat{\beta}$,

$$X_0 = \begin{bmatrix} 1 \\ X_{02} \\ \vdots \\ X_{0k} \end{bmatrix} , \text{ and } \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} .$$

Again, (10) and (11) are well-known results.* Also, as before, the standard error of estimate

$$\hat{\sigma} = \left[\frac{\sum_{i=1}^{n} (Y_i - \sum_{i=1}^{k} \hat{\beta}_i X_i)^2}{n - k}\right]^{\frac{1}{2}}$$

may be inserted in (10) and (11) and square root taken to get the multivariate versions of the standard error of forecasting Y and the standard error of forecasting the expected value of Y.

---

*See, for example, Johnston, op. cit., pp. 131-132 or Arthur S. Goldberger, Econometric Theory (New York: John Wiley, 1964), pp. 168-170.

We want to prove, for the general linear model,

$$(12) \qquad \mathrm{Var}(Y_0 - \hat{Y}_0)\Big|_{X_0 = \bar{X}} = \sigma^2\left(1 + \frac{1}{n}\right)$$

$$(13) \qquad \mathrm{Var}(EY_0 - \hat{Y}_0)\Big|_{X_0 = \bar{X}} = \sigma^2\left(\frac{1}{n}\right) ,$$

where $\bar{X}$ is a vector of sample means

$$\bar{X} = \begin{bmatrix} 1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_k \end{bmatrix} .$$

PROOF

In order to prove (12) and (13), we need only prove

$$(14) \qquad \bar{X}'(X'X)^{-1}\bar{X} = \frac{1}{n} .$$

First, we write $\bar{X}'(X'X)^{-1}\bar{X}$ as

$$\frac{1}{|X'X|}\bar{X}'C\bar{X} ,$$

where

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1k} \\ \vdots & & \vdots \\ c_{k1} & \cdots & c_{kk} \end{bmatrix}$$

is a matrix of cofactors $c_{ij} = (-1)^{i+j}|X'X_{ij}|$ and $|X'X_{ij}|$ is the minor of the ij^{th} element of $(X'X)$. Note that since C is symmetric $C = C'$.

Next, we write out the quadratic form $\bar{X}'C\bar{X}$, noting that

$\bar{X}_r = \left(\sum_{i=1}^{n} X_{ri}/n\right)$, $\bar{X}_j = \left(\sum_{i=1}^{n} X_{ji}/n\right)$, and $\sum_{i=1}^{n} X_{1i} = n$ since $X_{1i} \equiv 1$.

$$\frac{1}{|X'X|}\bar{X}'C\bar{X} = \frac{1}{|X'X|}\sum_{r=1}^{k}\sum_{j=1}^{k}\bar{X}_r\bar{X}_j c_{rj} .$$

$$= \frac{1}{n^2|X'X|}\sum_{r=1}^{k}\sum_{j=1}^{k}\left(\sum_{i=1}^{n}X_{ri}\right)\left(\sum_{i=1}^{n}X_{ji}\right)c_{rj}$$

$$= \frac{1}{n^2|X'X|}\left\{\sum_{j=1}^{k}\left(\sum_{i=1}^{n}X_{1i}\right)\left(\sum_{i=1}^{n}X_{ji}\right)c_{1j} + \sum_{j=1}^{k}\left(\sum_{i=1}^{n}X_{2i}\right)\left(\sum_{i=1}^{n}X_{ji}\right)c_{2j}\right.$$

$$\left. + \ldots + \sum_{j=1}^{k}\left(\sum_{i=1}^{n}X_{ki}\right)\left(\sum_{i=1}^{n}X_{ji}\right)c_{kj}\right\}$$

$$= \frac{1}{n^2|X'X|}\left\{n\sum_{j=1}^{k}\left(\sum_{i=1}^{n}X_{ji}\right)c_{1j} + \sum_{i=1}^{n}X_{2i}\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{n}X_{ji}\right)c_{2j}\right]\right.$$

$$\left. + \ldots + \sum_{i=1}^{n}X_{ki}\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{n}X_{ji}\right)c_{kj}\right]\right\} .$$

Now, note that $\sum_{j=1}^{k}(\sum_{i=1}^{n}X_{ji})c_{1j} = |X'X|$ since it is the expansion of $(X'X)$ by cofactors of the first row. Note also that $\sum_{j=1}^{k}(\sum_{i=1}^{n}X_{ji})c_{rj}$ $= 0$, $r \neq 1$, since it is the expansion of $(X'X)$ by alien cofactors, i.e., elements of the first row by cofactors of the r^{th} row, $r \neq 1$. Then

$$\bar{X}'(X'X)^{-1}\bar{X} = \frac{1}{n^2|X'X|}\left\{n|X'X| + \sum_{i=1}^{n}X_{2i}(0) + \ldots + \sum_{i=1}^{n}X_{ki}(0)\right\}$$

$$= \frac{1}{n^2|X'X|}\,n|X'X|$$

$$= \frac{1}{n} ,$$

which was to be proved.