MULTIPLE LINEAR REGRESSION VIEWPOINTS
Volume 11, Number 1, September, 1981

UNDERSTANDING PARTIAL REGRESSION COEFFICIENTS IN THE PRESENCE OF CORRELATED REGRESSORS

JEFFREY K. SMITH AND LINDA F. LEARY
RUTGERS UNIVERSITY

ABSTRACT

The interpretation of partial regression coefficients in the presence correlated regressors causes difficulty for students in the social sciences. nce correlation among regressors is the typical case in the social sciences, is presents a considerable instructional problem. This article presents an planation of the partial regression coefficient in the presence of correlated gressors that is a simple and direct extension of the case where regressors e mutually orthogonal. The interpretation presented emphasizes the rela- nship between the partial regression coefficient and the simple regression efficient. An example using SAS computer package is provided.

troduction

The extension of the principles and techniques of simple linear gression to multiple linear regression frequently results in confusion and isunderstanding for students in the social sciences. The major problem ea concerns the understanding of the regression coefficients when regres- s are moderately correlated. In most texts on regression analysis (Cohen d Cohen, 1975; Draper and Smith, 1966; Kerlinger and Pedhazur, 1973) e extension from simple to multiple regression is discussed via the special se where the regressors are uncorrelated. Pedagogically, this is appro- riate since it requires the introduction of a minimum of new concepts. owever, in the actual analysis of data in the social sciences, correlated gressors are far more the rule than the exception. Unfortunately, it is the conceptual leap from independent regressors to correlated regressors at there exists the greatest lack of clarity in explanation. An example of is confusion is the belief displayed not only be beginning students, but practicing researchers, that the order of entry of the variables into a

This work was supported in part by a grant from the Rutgers University Research Council. The authors appreciate editorial comments from an anonymous reviewer.

stepwise regression procedure affects the resultant regression weights when the full model is estimated! This confusion is exacerbated by the treatment of stepwise regression output in statistical computer packages such as SPSS (although, to the credit of the SPSS authors, they provide the best explanation we have found to date on the problem of correlated regressors) (Nie, et al., 1975).

The purpose of this paper is to present a lucid explanation of partial regression coefficients in the presence of correlation among the regressors. Our goal is to bridge the gap between a purely verbal explanation such as ". . . the increase in Y for a unit increase in X holding all other variables constant. . ." and a purely mathematical explanation such as:

$$B_{Y1\cdot 2} = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2}\left(\frac{S_Y}{S_1}\right)$$

Although both of these approaches are technically correct, neither provides a particularly good intuitive understanding of what is involved in multiple regression with correlated regressors.

Simple and Partial Regression Coefficients

It is our experience that the simple regression coefficient is readily comprehended by students as they approach multiple regression, and that an explanation of the partial regression coefficient in terms of a simple regression coefficient is heuristically appealing to students. Such a transition is clear and direct in the case of mutually orthogonal regressors. This multiple regression setting reduces to a series of simple regression equations (as in Draper and Smith, 1966, pp. 107-115). That is, the partial regression coefficient is identical to what it would be in a simple regression.

Our purpose here is to show that a similar reduction can be used even when regressors are correlated. The presentation below demonstrates how this would be done. It might reasonably follow the mutually orthogonal setting in a regression course.

Consider a regression with dependent variable Y, and three moderately correlated regressors X_1, X_2, and X_3:

(1) $$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_0$$

Since the regressors are correlated, it is obvious that the coefficient B_1 will not have the same value as a simple regression coefficient from the regression of Y on X_1 (alone). However, B_1 will be identical to the coefficient obtained from a simple regression of Y on the residuals of X_1 (say, X_1') after the collinearity with X_2 and X_3 has been removed. This can be accomplished by regressing X_1 on X_2 and X_3:

$$\hat{X}_1 = a_2X_2 + a_3X_3 + a_0$$

then $$\hat{X}_{1i}' = X_{1i} - \hat{X}_{1i} \quad .$$

The same procedure is followed for X_2 and X_3. A new equation:

$$\hat{Y} = B_1'X_1' + B_2'X_2' + B_3'X_3' + B_0' \tag{2}$$

can be shown to yield exactly the same regression coefficients as equation (1). That is, $\underset{\sim}{B}' = \underset{\sim}{B}$. The pedagogical advantage gained by creating equation (2) is that the $\tilde{X}_i$'s are mutually orthogonal and the B_i's can be understood as in the mutually orthogonal case. Thus, the partial regression coefficient is the simple regression coefficient of Y on the residuals of X_1 after the effects of X_2 and X_3 have been removed from X_1.

The utility of this approach to understanding regression coefficients is that it allows the student to link his comprehension of the partial regression coefficient to the firmer ground of the simple regression coefficient. This is particularly useful when such concepts as suppressor variables, multicollinearity, and shrinkage in r-squared are discussed.

An Example

An example of this approach with three regressors using the SAS statistical package is presented below:

```
(JCL)

DATA SAMPA;

INPUT  Y  X1  X2  X3;

CARDS;

   (insert data)

PROC GLM; MODEL  Y = X1  X2  X3;

PROC GLM; MODEL  X1 = X2  X3;
```

```
OUTPUT OUT = SAMPB  RESIDUAL = RESID;
DATA SAMPC; MERGE SAMPA SAMPB;
PROC GLM; MODEL Y = RESID;
DATA SAMPA;
PROC GLM; MODEL X2 = X1  X3;
OUTPUT OUT = SAMPD  RESIDUAL = RESID;
DATA SAMPE;  MERGE SAMPA  SAMPD;
PROC  GLM;  MODEL  Y = RESID;
DATA SAMPA;
PROC  GLM;  MODEL  X3 = X1 X2;
OUTPUT OUT = SAMPF  RESIDUAL = RESID;
DATA SAMPG; MERGE SAMPA  SAMPF;
PROC GLM;  MODEL  Y = RESID;
//
```

The first PROC GLM statement results in the standard multiple regression output for the full model. The second PROC GLM regresses X_1 on the remaining independent variables and calculates the residuals, while the third PROC GLM performs the regression of Y on the residual of X_1. Students can now verify that the regression coefficient for residuals is identical to that for X_1 in the original model. The remaining PROC GLM statements calculate the coefficients for X_2 and X_3 in the same manner. Although the layout for calculating all regression coefficients is presented here for completeness, calculation of only one or two of these may be sufficient for instruction.

References

n, J. and Cohen, P. (1975) Applied Multiple Regression/Correlation \nalysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum \ssociates, Publishers.

er, N.R. and Smith, H. (1966) . Applied Regression Analysis. New York: John Wiley and Sons.

nger, F.N. and Pedhazur, E.J. (1973) Multiple Regression in Behavioral Research. New York: Holt, Rinehart, and Winston, Inc.

N.H. et al. (1975) Statistical Package for the Social Sciences, Second Edition. New York: McGraw-Hill Book Company.