Examples of Easily Explainable Suppressor Variables in Multiple Regression Research

Franklin T. Thompson and **Daniel U. Levine**
University of Nebraska at Omaha

Multiple regression techniques are a valuable tool in conducting ecological studies, especially when provisions are made to control for problems dealing with the interaction of variables. One problem in multiple regression research, the presence of suppressor variables, has the potential to seriously limit findings that can be reported, and in some cases may cause a researcher to pass over a useful data set. Researchers have long been aware of the presence of suppressors in multiple regression research, but there is little agreement as to why it exists or what to do about it. Several considerations in employing methods to "unsuppress" several data sets are discussed.

Suppressor variables have been defined as variables that substantially improve the prediction of a criterion through the addition of a variable which is uncorrelated or relatively little correlated with the criterion but is related to another predictor or set of predictors. When suppression occurs, addition of the suppressor to the regression equation frequently is associated with a sizable increase in the beta weight(s) of the previously suppressed predictor(s), and, in a forward stepwise analysis, an increase in R-square nearly as large or larger than that contributed by the previously-suppressed predictor. Given this pattern, one might well refer to the variable that thus "kicks up" the prediction as an "unsuppressor".

Although we have been examining and consuming research based on multiple regression for many years, we seldom have encountered studies incorporating or reporting clear (and valid) suppression effects. Analysis of the functioning of suppressor variables and their dynamics is still less frequent, even in research that could be clearly improved by devoting explicit attention to the effects and meaning of suppressor relationships.To illustrate the functioning of suppressors in actual studies, and ways in which analysis of their effects can enhance understanding of relationships in a data set, we will portray and summarize three examples of suppressor variables in multiple regression analysis. We will conclude with suggestions regarding procedures that can help researchers in determining how to proceed in multiple regression studies that examine or should include examination of suppressor relationships.

II. Education and Military Spending in 78 Nations

Our first example of suppression occurs in a data set that examines the relationships between spending for education and for the military (both assessed as percentage of gross national product) and average life expectancy in a diverse group of 78 nations. Using the 2 expenditure variables in a forward stepwise regression analysis to predict life expectancy, education enters first with a standardized coefficient of .3602 and an adjusted r square of .118. Military spending then enters with a standardized coefficient of -.364, the adjusted R square increases to .231, and the coefficient for education spending increases to .462. Thus education now has a stronger relationship with life expectancy than was true before controlling for military spending, and the explained variance has increased by .113 even though the zero-order correlation between military spending and life expectancy is only -.238.The addition of military spending to the analysis has unsuppressed the underlying pattern wherein education spending now is more strongly related to life expectancy than before, and the two predictors together explain more of the criterion variance than might have been expected from an examination of zero-order relationships.

Having noticed the appearance of suppressor dynamics, we examined what was taking place by calculating correlations between education spending and life expectancy in countries high and low in military spending, and by plotting this relationship while portraying the high/low level of military spending (Figure 1). The correlation analysis showed that among 42 nations with military spending below 3.5 percent of GNP, the correlation between education spending and life expectancy was .62; among 36 nations with military spending at or above 3.5 of GNP, the correlation was virtually non-existent at .02. Thus education spending is highly related to life expectancy in countries with relatively low spending devoted to military purposes, but not at all related to life expectancy in countries that have relatively high military expenditures.Given this pattern, it is intuitively easy to understand why taking account of military spending clarifies and enhances the effect of education spending in the regression analysis.

Examination of Figure 1 (which shows only a random .5 sub sample of the nations in the data set) further points to what may be happening. As shown in the plot, few countries that are high in military spending are very low in life expectancy, thus restricting possibilities for a high correlation between expectancy and other variables. Having identified these patterns, we can proceed to try to determine (not discussed in this paper) why nations that are high in military spending as a percent of GNP generally are not low in life expectancy, and how this situation may involve relationships between these and other variables.

III. Family Income and Academic Achievement in Two School Districts

Our second example involves analysis of relationships between a measure assessing family income (i.e., percent of students from low-income families) and average sixth-grade mathematics scores at 55 elementary schools in 2 school districts. The first variable to enter in predicting achievement in a forward stepwise regression analysis was the family income measure, which correlated at -.574 with achievement and accounted for an r square of .329 in the latter criterion. This correlation was not nearly as high as we generally have found in other analyses of achievement in large school districts.

The major reason for this relatively poor prediction became quickly apparent when a dummy variable portraying the 2 districts in the data set entered the multiple regression analysis, and when we plotted family income against achievement taking account of district (Figure 2). Although its zero-order correlation with achievement was only -.242, the dummy variable increased the R square to .625 and pushed up the regression coefficient for family income to -.874. As shown in Figure 2, family income is highly correlated with achievement in both districts but achievement in district 1 is generally higher than achievement in district 2.

Results were even more clear and dramatic when we combined total student achievement scores (combined math, reading, and language sub test scores) of 52 schools from the two districts and plotted them (Figure 3) against a poverty indicator we referred to as "school SES" (i.e., a factor analysis score made up of percent mobility, percent minority, and percent poor students). The zero-order relationship between achievement and school SES was .520, with an adjusted r square of .25. After once again controlling for district differences, the dummy variable increased the R square to .882 with 77% of the variance explained; a dramatic .52 increase in the adjusted r square at the .000 significance level (Table I). In addition to achievement being generally higher in district 1 than achievement in district 2, we are left to speculate that there may be additional influences (not discussed in this paper) differentiating the districts which help to further suppress the relationship between total achievement and our socioeconomic poverty variable.

When district-level achievement and other possible district differences are controlled through multiple regression analysis, the effects of family poverty and socioeconomic status are "unsuppressed", and we can proceed to additional analysis (not discussed in this paper) and research examining reasons for the high correlation with achievement, substantive possibilities for overcoming this association through improved instruction, and causes of differential achievement in the 2 districts.

IV. Percentage of Students Residing Nearby and Math Achievement at 25 Schools in 1 School District

Various considerations led us to expect that schools which mostly enrolled students resident in their respective attendance areas in a school district we were studying would have proportionately lower achievement than schools which enroll higher proportions of students from distant neighborhoods. However, the correlation between percentage of resident students and average sixth-grade mathematics achievement was only -.023. Examination of the plot (Figure 4) suggested that a small group of 3 higher-than-predicted schools (i.e., box symbols with an x in Figure 4) was detracting from a clear relationship. We knew that reading scores accounted for more than 80 percent of the variance in math scores in this data set (as in many others), so we re-examined the relationship controlling for reading, and found that the standardized coefficient for percentage of resident students was now -.148. Inspection of the partial plot (Figure 5) indicated that increase in the size of the relationship between residency and math achievement was due to a reduction in the effects exercised by the three higher-than-predicted schools. Equally or more important, we were now in a better position to proceed with meaningful theoretical and quantitative exploration (not discussed in this paper) of relationships among variables in the analysis.

V. Discussion and Conclusions

The effects of a variable are "unsuppressed" when controlling for another variable indicates an increase in its relationship with the dependent variable. In the example involving family income and achievement described above and portrayed in Figure 2, the influence of poverty on achievement is increased to a multiple regression coefficient of -.874 from a zero-order correlation of -.574 because the latter relationship in a sense is a spuriously-low result of failure to control for district differences. As shown below, taking account of district in a path model helps the analyst understand underlying relationships and computations. Let "D" stand for district, "P" for the poverty/family income measure, and "A" for achievement:

-.448
D P

-.669 -.874
A

In this example, the zero-order correlation between P and A is the sum of the direct effect of P on A controlling for D and its indirect path through D. The calculations are as follows:

r pa = B pa.d + (r pd X B da.p)
-.574 = -.874 + (-.448 X -.669)
-.574 = -.874 + .300.
-.574 = -.574.

It is important to examine underlying interrelationships and even check out the calculations (as illustrated above) when one encounters regression data .indicating that suppression effects are present. For one thing, the data produced by the computer may be invalid: If there is high multicollinearity among predictors or if there are too few cases to sustain valid computations given the number of predictors, multiple correlations and regression coefficients may invalidly indicate whopping increases as new relatively-poorly correlated variables are added to a stepwise multiple regression.

In addition, examining the model and/or the calculations can help the analyst understand the dynamics of forces at work in the data set. For example, examination of the model shown above underlines the fact that on the average, schools in District 1 (coded as 1) have higher poverty and achievement scores than schools in District 2 (coded as 2), even though the "normal" strong relationships between poverty and achievement are apparent within each district. Furthermore, these relationships are clearly visible in and, indeed, clearly suggested by the plot portrayed in Figure 2. These considerations help lead us to the following general conclusions:

1. Plotting relationships can be very helpful in understanding the dynamics of a data set including suppressors, and also in verifying that suppressor relationships actually are present. In some cases, plots can call attention to analytic possibilities not previously apparent that are worth further exploration.

2. Investigation of suppressor variables and relationships can greatly enhance analysis and understanding of what is occurring or may be implied in a researcher's data set. However, researchers should be cautious in identifying suppressors, because statistics pointing toward the presence of suppressors frequently are invalid indicators produced by a sample that is too small or by highly correlated predictors.

Table 1
Multiple Regression Analysis* of School Inputs Using Dependent Variable Achievement: Observing the Effects of a Suppressor Variable for Combined District Data

Step number	Independent variable	N	MR	Adj. R^2	Standard error	Beta	T score	p
1	School SES	52	.52	.25	.86	.52	- 4.27	.000
2	School SES					-.97	-12.29	.000
	District	52	.88	.77	.48	-.85	-10.74	.000

* Probabilities of F for entry = .05, and for removal = .10