

# Chapter 7 — Using and Interpreting Multiple Regression

February 15, 2005

## 1 Another Formal Perspective on Holding Constant

Considered in the last chapter was the residualization process by which partial regression coefficients are constructed. Equation 1 addresses the same issues from a different vantage point. Again, there are only two terms, and we assume that in the population the usual regression model holds or that the model used corresponds to the way nature generated the data. In the population, the regression hyperplane goes through the conditional means of  $y$ . Stated to allow for data generated by nature, the errors are uncorrelated with the two terms  $x$  and  $z$ .

The regression coefficient for variable  $x$  is

$$\eta_x = \frac{r_{yx} - r_{xz}r_{yz}}{(1 - r_{xz}^2)} \times \frac{\text{SD}(y)}{\text{SD}(x)}, \quad (1)$$

where  $r_{(.,.)}$  is the Pearson correlation coefficient and  $\text{SD}(\cdot)$  refers to the standard deviation. From Equation 1, we learn several important things about the partial regression coefficient for  $x$ .

- If all of the correlations are nonzero, the marginal response will almost certainly differ from the conditional response. So omitted variables can be a very serious problem.
- If the terms  $x$  and  $z$  are uncorrelated, conditioning does not change things; hence, the virtue of randomized experiments. In principle, you

get the same estimate of the regression coefficient even if you leave out terms related to  $y$ . (But including them can improve the fit and reduce the standard errors—see below.)

- If  $x$  and  $z$  have a correlation of 1.0, the regression coefficient for  $x$  holding  $z$  constant is undefined. That is an example of linear dependence between terms.
- If  $x$  and  $y$  are uncorrelated, the regression coefficient for  $x$  may still be nonzero, depending on the correlation between  $x$  and  $z$  and the correlation between  $y$  and  $z$ . When the bivariate regression coefficient is zero and the partial regression coefficients are not zero, the term “suppressor effect” is sometimes applied.
- If a constant is added to any of the variables, nothing in Equation 1 changes. The correlations are all unit free and the standard deviations are unaffected when a constant is added.
- If any of the variables are multiplied by a constant, the changes in the partial regression coefficient that occur stem from changes in the standard deviations. When  $y$  is multiplied by  $c$ , the standard deviation for  $y$  is multiplied by  $c$ , and so the regression coefficient is multiplied by  $c$ . When  $x$  is multiplied by  $c$ , the standard deviation for  $x$  is multiplied by  $c$ , and so the regression coefficient is divided by  $c$ . Because the correlations are all unit free and because the standard deviation of  $z$  is nowhere to be found, multiplying  $z$  by  $c$  changes nothing. Linear transformations also have no effect on the fit or the standard errors because everything just scales up or down.
- Nonlinear transformations, in contrast, can change everything.
- Each of the above conclusions essentially generalize if  $z$  is a set of terms rather than a single term.

In short, the regression coefficient for the marginal relationship between  $y$  and  $x$  will typically be different from the conditional relationship between  $y$  and  $x$ . Linear transformations do not change anything fundamental, although the metric of the regression coefficients will change. But nonlinear transformations can make an enormous difference.

## 2 When Does Holding Constant Make Sense?

1. If the purpose of an analysis is to examine how the conditional mean of the response varies as  $x_j$  varies with all other terms fixed, one can proceed with multiple regression without much worry about what sense it makes to think of the other terms as fixed.
2. Moving to causal inference is an enormous step that needs to be thoroughly considered. To begin whether the causal variable of interest can be usefully conceptualized as an intervention within a response schedule framework. Recall what the econometric formulation of causal effects requires. It must be possible to manipulate each “input” *independently*. In Equation 1, for instance,  $x$  can be manipulated while  $z$  is unchanged and  $z$  can be manipulated with  $x$  unchanged. Consider two different situations: grant seed money and a one quarter sabbatical versus class size and course content. What about the price of water and water conservation education programs?
3. Now, we revisit the role of fixed attributes such as gender and race. Although these are not manipulable and hence, are not included within the definition of an intervention, one could still imagine that the response schedule for determining income, for instance, could differ depending on race and/or gender.

Consider the response schedule formulation again.

$$Y_{i,x} = a + bx + \delta_{i,x}, \quad (2)$$

where the  $\delta_{i,x}$ , with an expected value of zero, is drawn independently and at random from some invariant distribution, given a value for  $x$ . The coefficients  $a$  and  $b$  could differ depending on whether the subject is male or female.

In practice, one could apply simple linear regression separately to a sample of males and a sample of females to estimate the different values of  $a$  and  $b$ . Alternatively, the different response schedules for men and women could be formulated within a single multiple regression equation. How one would do so will be discussed in the next chapter.

### 3 Standardized Regression Coefficients: Once More With Feeling

One of the policy questions that social scientists like to ask is which terms are “most important” or “more important.” In response, some applied researchers transform each term into  $z$  scores (subtract the mean, divide by the standard deviation) before applying regression. The result is a “standardized” regression coefficient defined as

$$\hat{\eta}_j^* = SD(u_j)\hat{\eta}_j. \quad (3)$$

Now one can speak about the change in the mean of  $y$  for a 1- $SD$  change in  $u_j$ , other terms held constant. The new units for all of the terms are standard deviation units. Comparability seems to have been achieved.

But not so fast.

1. To begin, the process can be very close to tautological.
2. It also depends on the relevant variances.
3. The lesson is that “importance” must be defined outside of the data and in such a way that the standardized coefficients are responsive. When such an effort is made, importance is often defined as “variance explained.” So let’s examine that.
4. In the social sciences, is it common to find the standardized regression coefficient, or “beta,” defined as follows:

$$\hat{\beta}_j = \frac{SD(u_j)}{SD(y)}\hat{\eta}_j. \quad (4)$$

In effect,  $y$  and each of the terms have been transformed into  $z$  scores. Thus, for a 1- $SD$  change in  $u_j$ , there is an average change of  $\hat{\beta}_j$   $SD$  units in  $y$ . In some sense, the relationships are even more standardized.

If one squares  $\hat{\beta}_j$ , the result is one definition of the variances “uniquely explained” by  $u_j$ . So if  $\hat{\beta}_j$  is larger than  $\hat{\beta}_k$ , it means that  $u_j$  by itself accounts for more variance in  $y$  than  $u_k$ . In that sense, the former is more important than the latter. However, there are other equally defensible definitions of “unique” variance explained. For example, one

can define such an entity for  $u_j$  as the amount that the  $R^2$  declines if  $u_j$  is removed from the regression equation. This will generally not be the same as  $\hat{\beta}_j^2$ .

5. Yet another form of standardization is the elasticity, usually approximated by working with logs of all of the variables. Recall that they translate metric coefficients into units of percentage change and have the clear advantage of being useful theoretical concepts in economics. But when they are used solely for purposes of manufacturing comparability, they too raise the question of what the gains really are.
6. Finally, what is an applied researcher to do if the relative importance of predictors (or terms) really needs to be assessed? The answer is that importance is probably best defined in subject-matter or policy terms.

## 4 Variances of the Coefficient Estimates

When there is more than one term in the regression model, the degree of linear dependence among the terms affects the estimated standard errors. Even though this may seem to be only a technical issue, there are important implications for interpretation.

The standard errors for the regression coefficients in multiple regression are affected by the full set of terms, just as the coefficients themselves are. Thus,

$$\text{SE}(\hat{\eta}_j) = \frac{\hat{\sigma}}{\text{SD}(u_j)\sqrt{n-1}} \sqrt{\frac{1}{1-R_j^2}}, \quad (5)$$

where  $R_j^2$  is the  $R^2$  for the linear regression of the  $j$ th term  $u_j$  on all of the other terms in the mean function.

A brief examination of Equation 5 will indicate the following:

1. As in the simple regression case, the standard error is larger if the estimate of the residual variance (or standard deviation) is larger.
2. As in the simple regression case, the standard error is smaller if the term in question has more variance.
3. As in the simple regression case, the standard error is smaller if the sample size is larger.

4. This is new: The larger the value of  $R_j^2$ , the larger the standard error.
5. It is sometimes not appreciated, however, that Equation 5 also has important implications when statistical inference is not undertaken. A large standard error implies that small changes in the values of  $u_j$  lead to large changes in  $\hat{\eta}_j$ . For example, if the observational units included were changed a bit or if because of measurement error the values analyzed were a bit different,  $\hat{\eta}_j$  could be substantially different. Moreover, minor changes in which terms are included or in how those terms are defined also can lead to large changes in the estimated value of  $\hat{\eta}_j$ .
6. One popular index is the variance inflation factor (VIF), which is essentially the far right term in Equation 5:

$$\text{VIF} = \frac{1}{1 - R_j^2}. \quad (6)$$

The variance inflation factor is a positive number that multiplies the standard error. The greater the linear dependence between terms, the larger the multiplier. But there is no line in the sand.

7. It is also important to keep in mind that such concerns need only be raised for estimated regression coefficients of central importance to the research. It is only the  $R_j^2$ 's for those terms that matter.
8. The various strategies one may employ are discussed in any number of books on regression analysis.
  - (a) Dropping from the model some of the highly correlated terms (because as an empirical matter they are very similar anyway) item Imposing constraints on the regression coefficients to allow for more stable estimation
  - (b) Applying nonlinear transformations to the offending terms to reduce the problematic correlations.
  - (c) Employing ridge or Stein estimators.
  - (d) The best strategy usually is to collect more and/or better data, but in many situations, that is not possible.

9. Finally, this discussion of the variances of the estimated regression coefficients implies that doing power analyses for regression to help determine appropriate sample sizes is a daunting exercise. The problem is that one needs to anticipate reasonably well all of the  $R_j^2$ 's to determine the sample size required for a given level of precision