Using Ratios in Regression

Richard Goldstein, Ph.D.

July 15, 1999

Ratios are often used in regression to "adjust" or "standardize" for some factor such as size. One can divide the ways this is used into two classes, one of which is acceptable and the other of which is (generally) not acceptable.

1. Acceptable: If every variable in the regression is divided by the same factor there is no problem. This is done for example, when turning everything into a "per capita" measurement; another example is weighted regression. One needs, however, to be clear regarding what is meant by "every variable". Say your regression has two predictors (X and Z) and you want to control for population size (POP); the basic regressions looks like (suppressing the subscript for individual observations):

$$Y = b_0 + b_1X + b_2Z + e$$

When adjusted for population size, the regression should look like:

$$Y/POP = b_0/POP + b_1(X/POP) + b_2(Z/POP) + e/POP$$

Leaving out any of these terms will cause problems. (Note that inclusion of a constant in the second model, above, is called for if the first model also includes a term for POP (e.g., b_3POP). Note that this is what is usually called a "weighted" regression model.

2. Unacceptable: Sometimes it makes no sense to divide all variables by the denominator of the ratio; for example, in many health studies there is a desire to control for the size of the individual by using BMI (body mass index: wt/(ht^2)) as a predictor; another example occurs in the study of strength where the desire is to adjust strength by the size of the muscle (or muscle fiber); note in the latter case that the ratio will now be the response variable. If the set of predictors include any demographic variables (e.g., sex, age), then clearly one will not want to divide the demographic predictor by the denominator of the ratio. The issue here is mostly easily, I think, seen by observing that the ratio is an interaction term, but that the regression does not (usually) include the accompanying main effect terms; this is, among other things, a violation of the "marginality" principle[1]. In general, one does not want to automatically include an interaction term without its component parts. Further, the inclusion of an interaction term has implications about the form of the adjustment: use of BMI without either height or weight has implications for the way that size is adjusted and these implications may be wrong. The answer is to multiply out the ratio; e.g., if the ratio is in the response variable, multiply everything on the right by the denominator; if the ratio is in a predictor, add the component main effects to the model and see if the interaction (ratio) adds anything. A good discussion of this case, with explicit advice, can be found in Kronmal, R.A. (1993), "Spurious correlation and the fallacy of the ratio standard revisited," *Journal of the Royal Statistical Society, series A,* 156: 379-392.

---

1 For example, including one main effect but not the other implies that the intercept but not the slope is independent of the other main effect. For more, see Nelder, J.A. (1998), "The selection of terms in response-surface models – How strong is the weak-heredity principle?", *The American Statistician,* 52: 315-8.