

Multiple Linear Regression

James H. Steiger

The Fundamental Idea

In least squares linear regression, we predict a single criterion, y , from a single predictor, x , with the basic equation being

$$\begin{aligned}y &= b_1x + b_0 + e \\ &= \hat{x} + e\end{aligned}\tag{1.1}$$

The *least squares* criterion minimizes the sum of squared errors in the sample, and the expected sum of squared

errors in the population. The solutions to b_1 and b_0 are well known to all basic statistics students.

In multiple linear regression, we simply add more predictors. The previous equation now becomes

$$\begin{aligned} y &= b_1x_1 + b_2x_2 + \dots + b_kx_k + b_0 + e \\ &= \hat{x} + e \end{aligned} \tag{1.2}$$

Key Initial Question – Which Predictors?

Often in practice, one is confronted with numerous *potential predictors* of an important criterion. Which one(s) should be used? The question is complicated by several factors, not the least of which are (a) sampling error and (b) the problems connected with *post hoc* selection.

Two Statistical Models

Many introductory texts do not emphasize that there are actually *two statistical models* used in standard linear regression. For a normally distributed set of variables in \mathbf{y} , we have

Fixed Predictor Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

$$E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I} \quad (1.3)$$

$$E(\mathbf{e}) = \mathbf{0}$$

which implies

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad (1.4)$$

The scores in \mathbf{X} are *fixed scores*. As we will see, they may represent *design codes* in ANOVA, for example.

Random Predictor Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

$$E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I} \quad (1.5)$$

$$E(\mathbf{e}) = \mathbf{0}$$

where the random variables in \mathbf{y} and \mathbf{X} represent observations from a multivariate normal distribution.

Choosing a Model

Statistical theory is different for the two models. The theory is *much more complicated* for the random predictor model.

The fixed predictor model is reasonable for the *regression treatment of ANOVA*. However, the random predictor model often mirrors, more accurately, what happens in the social sciences, where people sample a large number of observations from a set of predictors and a criterion that jointly have a distribution that is approximately MVN.

Guess What?

However, all the classic treatments of multiple regression in intro textbooks use the fixed predictor model, often without commenting.

This is not as bad as it sounds. The basic test that $\rho^2 = 0$ is the same in both models. However, the non-null distribution is actually different. So classic power calculations given by, for example, Cohen in his power analysis book, actually are approximations. For most applications, these approximations are reasonable.

Multiple, Partial and Semipartial ρ^2

A squared correlation can be expressed as the ratio of the variance accounted for divided by the total variance.

For example, the squared multiple correlation of y with x and w is the ratio of the variance predictable uniquely from x and w divided by the variance of y .

So in T&F figure 5.2a, what would be the squared multiple correlation between DV and the predictor set IV_1, IV_2 ?

The squared *semipartial* (or *part*) correlation is the proportion of variance in the criterion predictable from that part of the predictor remaining after the partial variables have been subtracted out of the predictors.

What would be the squared semipartial correlation between DV , and IV_2 after IV_1 is partialled from IV_2 ?

What would be the squared semipartial correlation between DV , and (IV_1, IV_2) after IV_3 is partialled from (IV_1, IV_2) ?

The squared *partial* correlation is the variance in the criterion predictable from that part of the predictor remaining after the partialled variables have been subtracted out of the predictors *and* the criterion.

What would be the squared partial correlation between DV , and IV_2 after IV_1 is partialled from IV_2 and DV ?

What would be the squared partial correlation between DV , and (IV_1, IV_2) after IV_3 is partialled from (IV_1, IV_2) and DV ?

Multiple regression weights are based on semi-partial variances. Consequently, the regression weight between an independent and dependent variable very much depends on what other independent variables are in the prediction equation.