Multiple Regression

Christian DeLucia, Kaleb Pratt, and Ashley Strong

Nova Southeastern University

Researchers use statistical models to test hypotheses about developmental phenomena. The multiple regression model is a powerful and flexible statistical model that can be used to answer a multitude of research questions. For example, a researcher might be interested in testing whether parent substance use, peer substance use, and lower school performance are risk factors for adolescent substance use. By collecting data on a sample of adolescents, a researcher could use a multiple regression model to provide insights into this research question. Below, we provide a brief, non-technical introduction to the multiple regression model.

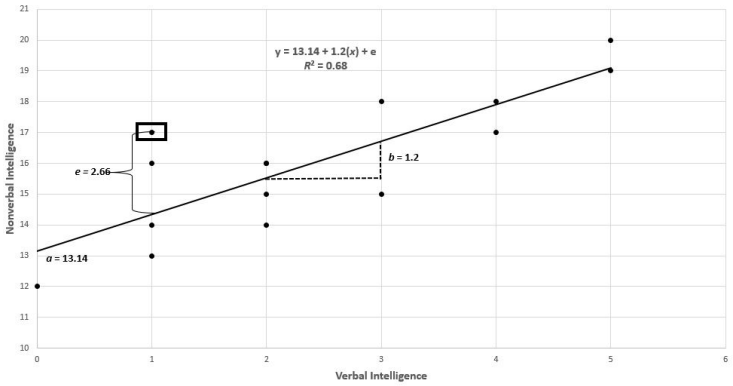
In general, the multiple regression model involves predicting a single outcome variable from two or more predictor variables. The outcome, often designated as the *y* variable, is a continuous variable (suggesting it can take on many different values). The predictor variables, often designated as *x* variables, can be categorical (e.g., diagnosed/non-diagnosed) or continuous (e.g., a measure of intelligence). First, consider a single predictor model in which 15 participants were measured on two IQ test subscales: verbal and nonverbal intelligence. An equation for the single predictor model is:

$$y_i = a + b(x_i) + e_i$$

Figure 1 shows the scatter of points with verbal intelligence (denoted by x_i in the equation above) on the *x*-axis and nonverbal intelligence (denoted by y_i in the equation above) on the *y*-axis. In

this example, verbal intelligence is the predictor and nonverbal intelligence is the outcome. Each point in the plot is the intersection of an *x*-*y* pairing. For example, the boxed point is from a child who scored a 1 on verbal intelligence and a 17 on nonverbal intelligence. The *y*-intercept, denoted by "*a*" in the equation above, is the predicted *y*-value when *x* equals zero. The best-fit line codes the linear association between *x* and *y*. The slope associated with the best-fit line is *b* in the equation above and is referred to as the regression coefficient. Its numerical value, 1.2 in this case, shows that as individuals increase by one unit on *x*, their predicted increase on *y* is 1.2 units. In other words, a one unit increase on verbal intelligence is associated with a 1.2 unit increase on nonverbal intelligence. A positive regression coefficient suggests that increases in *x* are associated with increases in *y* and decreases in *x* are associated with decreases in *y*. A negative regression coefficient suggests that increases in *x* are associated with decreases in *y* (or decreases in *x* are associated with increases in *y*). For positive associations, the variables are said to "change" together; whereas for negative associations, the variables are said to change in opposite directions. The error in prediction is captured by *e* in the equation above, which equals the difference between the observed and predicted *y* scores for each individual. For the boxed point in the plot, the residual is 2.66 (17 – 14.34). Terms that carry the "*i*" subscript in the equation above are free to vary over individuals. In general, regression coefficients are deemed statistically significant if they can be significantly differentiated from zero. A regression coefficient of zero indicates that no linear relationship exists between *x* and *y*. Various squared correlation coefficients can be used to quantify the magnitude or strength of the linear association. Squared correlations reflect the proportion of variance in the outcome that can be uniquely attributed to the predictor(s).

Figure 1. Verbal Intelligence as a Predictor of Nonverbal Intelligence



A far more typical research scenario involves predicting a single outcome from multiple predictor variables, requiring the use of multiple regression. The equation for the multiple regression model is expanded to include as many *b(x)* pairings as there are predictor variables. In the equation below, there are *k* predictors; in a three-predictor model, *k* equals three.

$$y_i = a + b_1(x_{1i}) + b_2(x_{2i}) + ... + b_k(x_{ki}) + e_i$$

The primary strength of the multiple regression model is it allows a researcher to examine the effects of the full set of predictors, of subsets of predictors, and/or of individual predictors. The primary advantage of including multiple predictors in the same model involves being able to look at their unique effects on the outcome while statistically controlling for the other predictor(s) in the model. For example, one might be interested in predicting reading achievement in children from three predictors: verbal aptitude, parent education, and motivation. In this case, simple (zero-order) correlations between each predictor and the outcome might result in three positive and significant correlations. Entered into a multiple regression model, however, one or more predictors might fail to uniquely predict the outcome. In developmental

psychology, predictors are often highly inter-correlated (raising the potential importance of statistical control). Below we present output from our hypothetical example in which reading achievement is predicted from verbal aptitude, parent education, and motivation.

	Statistics		
Predictor	*b*	*SE*	*p*
y-intercept	1.864	0.547	0.006
Verbal aptitude	0.573	0.141	0.002
Parent education	0.174	0.23	0.465
Motivation	0.682	0.208	0.007

The R^2 for the overall model is .89, suggesting that the three predictors account for 89% of the variance in child reading achievement. The *b*s are regression coefficients, the *SEs* are standard errors, and *p* values below .05, suggest individual predictors significant at the .05 level. Only verbal aptitude and motivation are significant unique predictors. Parent education fails to predict unique variance in achievement once the effects of verbal aptitude and motivation have been statistically controlled.

Myriad extensions of the basic model presented above can be implemented to answer interesting developmental questions including whether predictors interact in predicting an outcome and/or whether associations between predictors and outcomes are non-linear. For a comprehensive discussion of the multiple regression model, the reader is referred to Cohen, Cohen, West, & Aiken (2003).

Further Reading

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.