The Center for Research and Outreach
REACH
Leveraging Evidence and Science into Practice

UNDERSTANDING REGRESSION ANALYSIS

A HOW-TO GUIDE

USDA

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Developed in collaboration with the Department of Defense's Office of Family Policy, the National Institute of Food and Agriculture, and the U.S. Department of Agriculture under Award No. 2009-48667-05833

Understanding Regression Analysis: A How-To Guide

Why are some youth program instructors more effective than others? Why are some children more engaged in youth programming while others seem to pay less attention? Why do some youth attend programs regularly while others go rarely or not at all? These questions and more can be answered (at least partially) through the use of regression analysis.

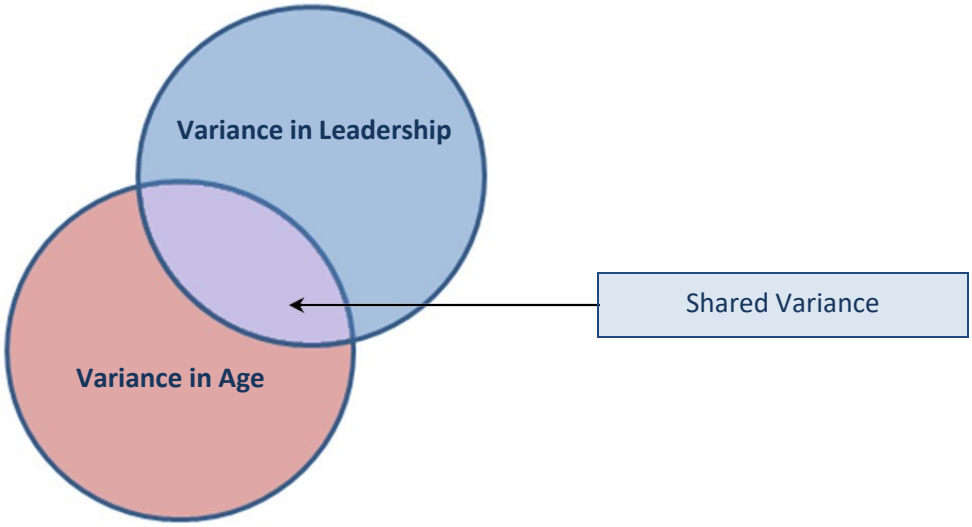
The differences between instructors and between children described above refer to variance, or the differences between different individuals. Variance is the degree to which scores on some measure differ. For example, suppose you sampled 100 adolescents about the amount of leadership they take in their youth program. Some adolescents would score very high while other adolescents would score very low; others would fall somewhere in between. Thus, scores from these 100 adolescents would vary quite a bit, hence the term variance. Regression analysis is a way of explaining variance, or the reason why scores differ within a surveyed population.

The Link Between Correlation and Regression

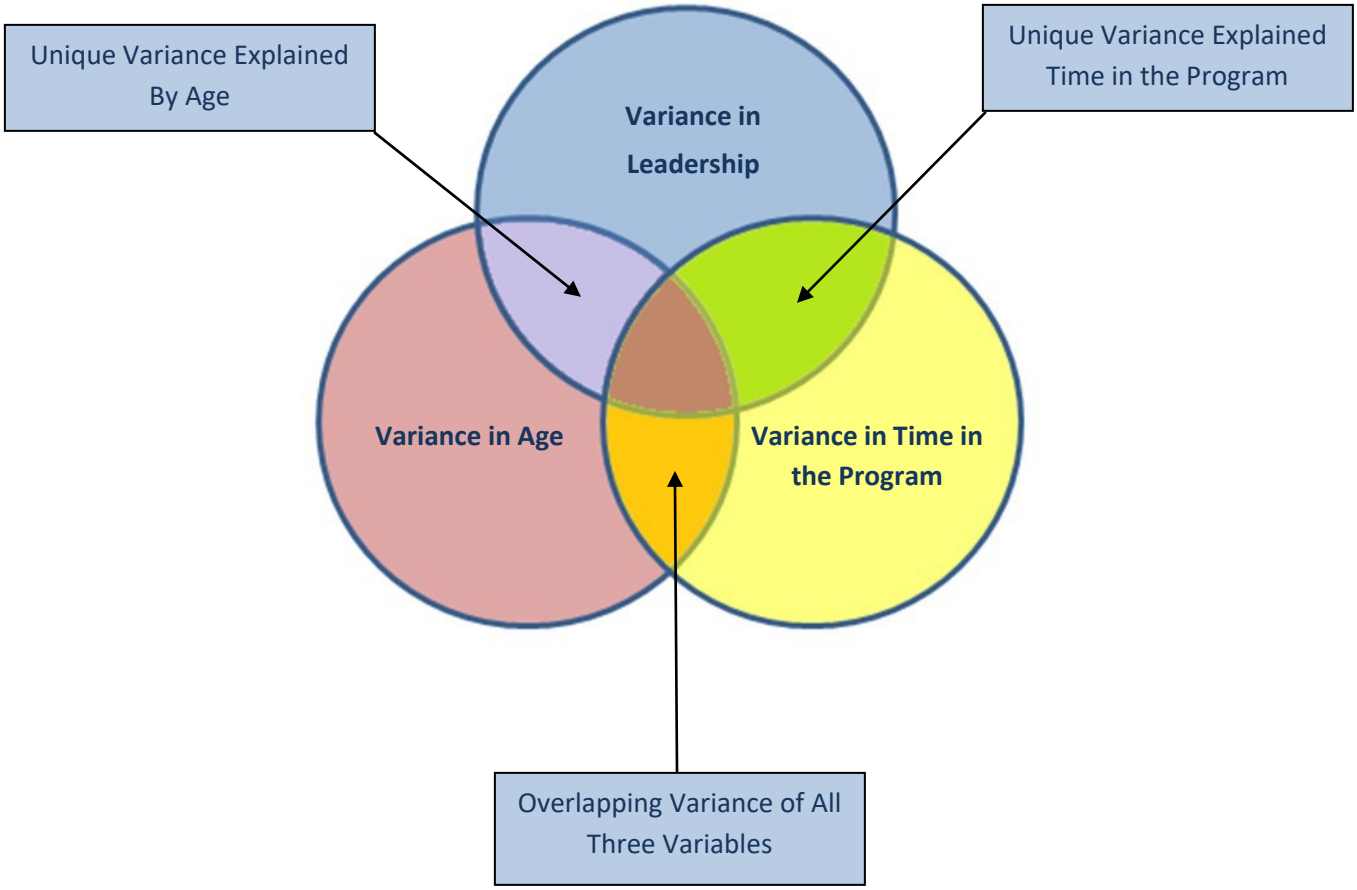
Regression can be thought of as a more advanced correlation analysis (see *Understanding Correlation: A How-To Guide* if you are unfamiliar with correlation). In fact, regression is based on the concept of a correlation. While correlations provide information about the association between two variables, regression often tells you information about the association between three or more variables. Typically, two or more variables are used to explain variance in a third. Along these lines, there are two basic types of variables that are used: (1) outcomes and (2) predictors. Outcomes are the main variables of interest. They are the variables for which you are trying to explain variance. Predictors, on the other hand, are the variables that you use to explain variance in the outcome. For example, suppose we hypothesized that two reasons for variance in adolescent leadership may be due to age (i.e., older adolescents may be better equipped to take on a leadership role) and how long they have been a member of the youth program (i.e., adolescents in the program longer may be better equipped to take on a leadership role). In this example, the outcome would be leadership and the predictors would be age and length of time in the program.

The utility of regression, compared to a correlation, is its ability to examine different elements that explain variance in the outcome. Imagine that the contents of this circle represent all the variance in leadership we obtained from our sample of 100 adolescents. We want to try to explain this variance using our two predictors, age and length of time in the program.

Let's start by considering using only the first predictor: adolescent age. Regression with only one predictor is the same as a correlation. In the diagram below, age is the predictor and leadership is the outcome. Each circle represents variance in each of these variables. The purple part of the diagram indicates the shared variance between these two variables. In other words, the purple part represents the amount of variance that age explains in adolescent leadership.



When a second predictor is added, the figure gets a bit more complicated. The utility of regression is its ability to divide variance in the outcome among the different predictors. In the diagram below, the purple section is the amount of unique variance explained by age, the green area is the amount of unique variance that is explained by length of time in the program, and the brown area is the amount of overlapping variance that all three variables share. The remaining blue area in variance in leadership is the amount of variance left over after accounting for the variance explained by age and length of time in the program; we refer to this as "unexplained variance".



Tying It All Together

Now that we have reviewed the basics of regression analysis, we can talk about how regression might be reported and interpreted in a research article. There are two key terms that are important in reporting regression analysis: (1) regression coefficients symbolized by the Greek letter beta (*β* or *B*) and (2) the R^2 coefficient. First lets cover the regression coefficients. These coefficients refer to the size of the unique association between the predictors and the outcome. The value of the *β* (or *B*) represents the purple area for age and the green area for time in the program in the above diagram. The larger the areas, the greater the beta-weight will be. Thus, the larger the regression coefficient, the greater the association (variance explained) between the predictor and outcome. The R^2 coefficient represents the percentage of variance in the outcome accounted for by <u>all</u> predictors included. In the present example, the purple, brown, and green areas are represented by the R^2. R^2 is an important coefficient to know as it provides overall information about the ability of the regression model to explain variance in the outcome.

Below is an example of how a regression analysis might be reported in a research article:

> *Participant age and the length of time in the youth program were used as predictors of leadership behavior using regression analysis. Results indicated that both age (β = .32, p < .05) and length of time in the program (β = .27, p < .05) uniquely predicted variance in leadership. The overall model explained 26% of the variance in leadership (R^2 = .26).*

From this example, several pieces of information can be derived.

1. **Predictors**: In this example, the predictors are identified as age and length of time in the youth program.

2. **Outcome:** Because leadership is what is being predicted, we can deduce that it is the outcome.

3. **Regression Coefficients:** Two regression coefficients are reported in this example, one for age (β = .32) and one for length of time in the program (β = .27). The size of these coefficients indicates the degree of the unique relation between each predictor and the outcome. Interpretation of the size of the relationship (small, medium, large) is similar to interpreting the size of correlations.

4. $\boldsymbol{R^2}$: The R^2 value from this regression analysis was .26. This number provides the percentage of variance that the predictors explain in the outcome. Overall, the regression model explained 26% of the variance in leadership. This essentially means that 26% of the reason why adolescents differ in terms of their leadership is due to age and length of time in the program.

A Cautionary Note on Causality

In regression choices are made regarding what variables are predictors and what variables are outcomes. The simplest case of regression where there is one predictor and one outcome is really the same as a correlation. The fact that one variable is chosen as the predictor and one is chosen as the outcome does not imply that the predictor *causes* the outcome but rather that they have an *association*. This is also true when there are several predictors and one outcome. Just because one variable or a set of variables are designated predictors does not mean that they cause the outcome. Instead, regression gives us useful information about the amount of variance that one variables accounts for in another by virtue of the unique associations that the predictors have with the outcome.

Summary

In this How-To Guide we have gone over the basics of regression analysis. At this point you should be equipped to read and understand the results of a regression model in research articles. On the website there are several other How-To Guides that cover other statistical topics such as *Analysis of Variance*, *Correlation*, and *Statistical Language*. These guides may be helpful to you as you become a consumer of research.

Developed by
The Arizona Center for Research and Outreach (AZ REACH)

Lynne M. Borden, Principal Investigator
Gabriel Schlomer, Ph.D. (Primary Author)

Updated March 2021

The Center for Research and Outreach (REACH Lab)
University of Minnesota
Email: REACHLab@umn.edu
Website: https://reachfamilies.umn.edu