Cornell University
Cornell Statistical Consulting Unit

StatNews #66

Meaningful Regression Parameters Through Centering
December 2004

When we run regression analyses, we want to ensure that the regression parameters have meaningful interpretations. But sometimes parameters are not inherently meaningful, even in simple situations. Consider an example where we predict weight from age for a data set with children between 6 and 12 years old. We fit a simple linear regression model to these data, where weight in pounds is assumed to be related to age in years as a straight line:

WEIGHT = 30.6 + 3.6*AGE

Here, the regression parameter for age is 3.6 lb/year, and it indicates that weight differs by 3.6 lb for a one-year difference in age. The intercept value of 30.6 lb represents the value of weight when the child is 0 years old, which is not very meaningful, as age varies in the data set only between 6 and 12 years.

Centering can make regression parameters more meaningful. Centering involves subtracting a constant from every observation's value of a predictor variable and then running the model on the centered data. Many times, it is helpful to center the data around the mean of the variable, although any logical constant can be used. In the above example, if we center AGE at some value within its range (for instance, 6 years), the following equation is obtained:

WEIGHT = 52.4 +3.6 (AGE - 6)

The intercept now represents the weight of child at the centered value AGE= 6. The regression coefficient for AGE remains the same. Only the intercept differs from its original value, and this is because its meaning has changed. In the centered form, the intercept value of 52.4 represents the weight of the child at age of 6 years, which provides a more meaningful interpretation.

Centering is particularly useful when regression equations include multiplicative terms to specify interactions or curvature. For example, polynomial regression may involve quadratic or higher order terms. In the above example, suppose we evaluate a model in which we want to relax the assumption that the relationship between weight and age is a straight line. We might then predict weight by age and age-squared:

WEIGHT = 32.4 + 3.2*AGE + 0.025*AGE 2

The intercept of 32.4 again represents the weight at AGE=0, which is not in the range of the data. Furthermore, because the model has a quadratic term, the parameter of 3.2 for AGE now represents the slope between WEIGHT and AGE only at AGE=0, outside the range of the data.

Instead, we could center AGE at 6. We would then obtain the following result:

WEIGHT + 52.5 +3.5(AGE-6) + 0.025(AGE-6)2

The centered model provides a more meaningful interpretation, where the intercept value of 52.5 represents the value of weight when the child is 6 years old. The linear term of 3.5 indicates the linear trend (i.e., slope) in the relationship between weight and age at 6 years. In other words, this is the tangent line for the curve at age 6 years. This is advantageous because the interpretations of the intercept and slope are now in the range of the observed data.

Note that in the model without the quadratic term (i.e., age-squared), the interpretation of the regression parameter for AGE is that weight differs by 3.6 lb for a one-year difference in age throughout the range for age. When the model includes the quadratic term, the meaning of the regression parameter for AGE is different because it now represents a difference in weight for a one-year difference in age only at one point, i.e., one single age. Centering is particularly helpful in this model because it moves that one point within the range of the data for age.

Furthermore, when multiplicative terms are added to a regression model, the original variables and the multiplicative terms can be highly correlated. Centering of predictor variables may reduce these correlations and avoid computational difficulties that may occur in extreme situations. Nevertheless, the main advantage of centering the predictor variables is that it often results in regression parameters that are easier to interpret.

Author: Simona Despa sd249@cornell.edu

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).