

## Making Sense of Biostatistics: Regression Analysis

By Ronald E. Dechert

In the last column, we examined measuring correlation to evaluate association between two continuous variables. In this column, we will examine another method: regression analysis. Like correlation, linear regression analysis is based on the following formula:

$$Y = a + BX$$

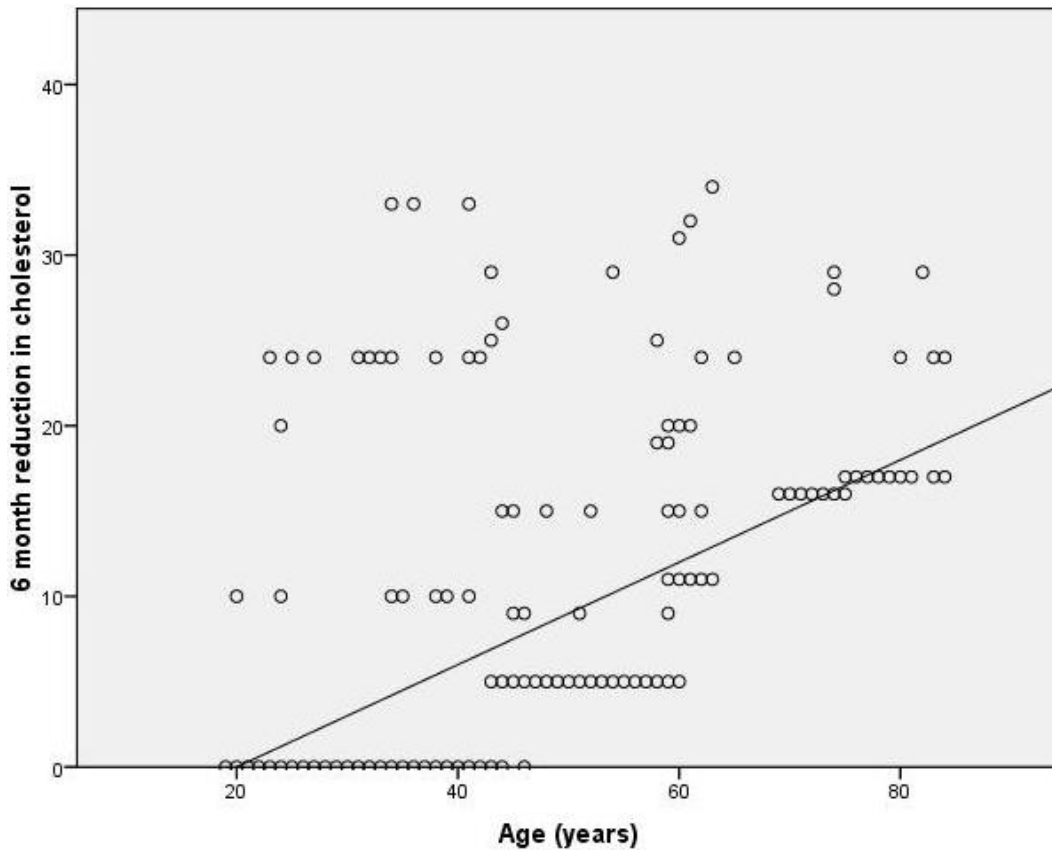
Francis Galton coined the statistical term "regression" in 1885 to describe his observation relating the height of parents and their children. The height of a child tends to be the average of the parents' heights and the population's average height; i.e., the child's height regresses to the mean of the population. Galton's formula turns out to be very useful in fitting a line to a set of bivariate data.

Let us examine a theoretical problem. The Dechert Pharmaceutical Corporation has developed Xema, a new statin that may offer more benefit than existing statin products. Investigators test the new statin using a cohort of clinical volunteers who take 20 mg of the new statin daily for six months. The volunteers have their cholesterol measured at the beginning and end of the six-month period. The investigators want to know if Xema was effective in lowering the cholesterol of the test subjects over the six-month period and, if so, was there an association between the age of the test subject and the degree of change in cholesterol.

To answer the second question (i.e., association between age and cholesterol change), we need to employ regression analysis. Before applying regression analysis, we will want to examine the graphical representation of the data using a scatter plot. Figure 1 demonstrates the scatter plot of change in cholesterol against age of the test subjects.

Although the scatter plot in Figure 1 shows a lot of variation, there appears to be a general association between reduction in the subjects' cholesterol observed following six months of trial and the age of the subject. Specifically, the change in cholesterol is greater in the older population compared to the younger population. Although there are many approaches to fitting lines of various shapes through the data, Figure 1 suggests that a linear (straight line) regression model would be appropriate. There are several reasons to select a straight-line model over a curvilinear model. The most common reasons are: (1) a straight-line model is standard and easy to interpret; (2) selection of the correct curvilinear model is difficult; and (3) straight-line models produce additional information that allows estimates for the dependent variable (Y axis) based on the value of the independent value (X axis) and the degree of co-relation that exist between Y and X.

**Figure 1: Scatter Plot of Six-Month Change in Cholesterol vs. Age of Test Subject (Years, mg/dL)**



Prior to testing Xema in the test cohort, the investigators hoped to address two basic questions: (1) does Xema effect cholesterol levels after six months of use, and (2) if there is an effect, is the strength of the effect associated with the age of the test subject? Having conducted the linear regression analysis, the investigators obtained the results in Figure 2:

**Figure 2. Xema Linear Regression results**

Result	Value	Significance (p=)
Mean reduction in cholesterol ( <b>u</b> )	8.91	0.001
r= (correlation coefficient with age)	0.570	0.001
r-squared=	0.325	0.001
Constant ( <b>a</b> )	-5.912	0.001
Coefficient ( <b>B</b> )	0.302	0.001

From these results, we see that Xema did, in fact, result in a statistically significant reduction in cholesterol when taken for six months ( $u=8.91$ ,  $p=0.001$ ). In addition, there is a significant association between age and cholesterol reduction in the study cohort ( $r=0.570$ ,  $p=0.001$ ) and approximately 32% of the variability observed in the cholesterol reduction is explained by a similar variability in age. However, regression analysis allows us to say more about the magnitude of the changes we observed in the test subjects' cholesterol. Using the linear equation ( $Y = a + Bx$ ), we are able to state that for each year of age we can expect to observe a 0.302 reduction in the subject's cholesterol after taking Xema for six months.

For example, suppose a patient (age 50) asks how much reduction in her cholesterol she should expect to see if she takes Xema for six months. Her physician can answer that, on average, she can expect to see a reduction of 10 mg/dL. Further inspection of Figure 1 demonstrates that the observed range of reduction for her age group is between 6 and 28 mg/dL in the population studied.

Correlation analysis produces an index ( $r$ ) that allows the investigator to interpret the relationship between bivariate data. Regression analysis provides additional information that allows estimates for the dependent data. As such, it is a useful tool for clinical investigation and clinical practice. Linear regression provides a foundation to understand more complex regression modeling, such as multivariate and logistic regression modeling.

### **Author**

Ronald E. Dechert, DPH is Associate Director of the Mott Respiratory Care at the University of Michigan Medical Center. Contact him at 734.936.5237 or rdechert@umich.edu.