# This presentation introduces the topic of the General Linear Model,

known as GLM for short. This presentation is the first in a series about GLM. We are less concerned with calculation this video and more about the concept of the line and how we apply it in statistical analysis.

The General Linear Model is widely used in statistical analysis, especially in multivariate statistics. At least four additional videos in this series are about using the General Linear Model to determine relationships. This count does not include videos on the development of diagnostic techniques and corrections based on diagnostics. This count also not include the rise of classes of statistics to correct for specific situations that the General Linear Model does not handle well—situations where we are not using the General Linear Model but where the General Linear Model is still in the back of our heads as the reason why we are doing what we do. With the General Linear Model, all four elements of an inferential statistic are present: direction, magnitude, significance, performance of the model

Let’s start with a simple example. I want to determine whether the poverty rate might influence the violent crime rate.

Here’s a simple scatterplot, which can be produced easily in Excel or SPSS, showing the relationship. The violent crime rate in this analysis is actually logged because of skewness in the distribution. Pearson’s r is .39, meaning we have a moderate, positive relationship. As the poverty rate increases, the violent crime rate also increases. Now that you know the relationship is positive, if you are like me, you are trying to draw a line across the graph to show its positivity. You might even wonder why we just don’t draw a line on the scatterplot to make it easy.

I made this scatterplot in Excel, which will insert the line for me. Here it is. The line shows our moderate, positive relationship. The General Linear Model is the technique we use to find this line as well as determine how well it fits the points.

We all know about lines. Back in junior high or high school in pre-algebra class, maybe earlier, we had an equation that looked like this: y equals mx + b.

X and Y are the coordinates of a point and become the variable values in the equation. M is the slope, which is defined as the change in Y over the change in X, or colloquially, the rise over the run. B is the y-intercept.

We’ve all been taught that if we have two points, we can calculate the line connecting them. Here I will find the line connecting two points on a two-way scatterplot. The points are (3,3) and (1,2). In this picture, the x-axis runs from left to right, and the y-axis runs from top to bottom. The point where there cross in the middle of the screen is called the origin of the axes and has the coordinate (0,0). The coordinate on the x-axis is mentioned first, and the coordinate on the y-axis is mentioned second. Thus, (1,2) is one unit to the right along the x-axis and two units up on the y-axis.

Here is the calculation for the value m, the slope. In the numerator of the fraction is the distance between the two Y values and in the denominator is the distance between the two X values. The value of the slope in the example is ½, which means that when X increases by 2 units, y increases by 1unit. Another way to say this is that y increases 1 unit for every 2 units of X.

Once the value for the slope is known, the variable m can be replaced with the slope’s value, and we can put the coordinates of any known point on the line in the equation. These substitutions leave b as the only unknown in the equation. By simple algebra, we can solve for b, which gives us the y intercept. That is, we know the value of y when x is 0. In this example, the y intercept is 3/2, which means that the line crosses the y-axis at the point (0, 3/2).

By simple substitution, the equation of the line is now known: y = (1/2)x + (3/2).

# There are differences in the symbols used to describe the line in statistics.

Some of these differences are merely cosmetic, while others reflect the unique situation posed by running the line through a cluster of points.

First, in statistics the y-intercept is often presented first. Here I simply reverse the terms on the right side of the equation. It means the same thing, but a version of this form is used in stats to discuss the General Linear Model.

Now we change the symbols around. In statistics, the slope is represented by the letter B or sometimes the Greek letter beta. The y-intercept is sometimes symbolized by the letter a or the Greek letter alpha. Y is symbolized by y-hat, which is also referred to as the predicted value.

The reason why we refer to the dependent variable as y-hat rather than y is because our line does not necessarily have to fall on any of the actual values of the distribution. There’s a gap, or error, between y-hat and y.

Sometimes we want to include the error in our equations so that we can see how everything totals up. When we do this, we add e to the equation to account for the error . We add it to each side of the equation so that the equation is balanced.

y-hat plus e is the same as y, so we change the left hand of the equation to get the result seen here. This fourth equation is regarded as equivalent to the third—it’s just the symbolization that’s different.

Here’s our final alteration: The intercept is sometimes referred to as beta-naught, and the slope is referred to as beta-subscript-one. We can use this notion with or without the error term, adjusting the symbol for the dependent variable accordingly.

The effective use of the General Linear Model is accompanied by assumptions, some of which are highlighted here. First, the technique is used in situations where the dependent variable is numeric. There is an exception to this rule and I will highlight it when appropriate. Normally, though, the dependent variable is numeric.

Second, the numeric variables have normal distributions. When we look at the equations in other videos, we will note that the parameters of a line are calculated using the mean and standard deviation. If there is skewness or outliers in the underlying variables, the linear model will be distorted.

Third, the error terms are homoscedastic. Homo means same. It means that the size of the error is same throughout the model. Another way to express this idea is to say that the error is randomly distributed. This issue is discussed in the video on regression diagnostics.

Finally, we can have more than one independent variable in the General Linear Model. When we do, the independent variables should not be correlated with each other. We call this property independence, and it will also be dealt with in subsequent videos in this series.

The General Linear Model is the foundation of a variety of statistical techniques. Subsequent presentations cover these techniques and more.