15.1: Introduction- Line of Best Fit

In correlations, we referred to a linear trend in the data. That is, we assumed that there was a straight line we could draw through the middle of our scatterplot that would represent the relation between our two variables. Regression involves an equation of that line, which is called the Line of Best Fit.

The line of best fit can be thought of as the central tendency of our scatterplot. The term “best fit” means that the line is as close to all points (with each point representing both variables for a single person) in the scatterplot as possible, with a balance of scores above and below the line. This is the same idea as the mean, which has an equal weighting of scores above and below it and is the best singular descriptor of all our data points for a single variable.

We have already seen many scatterplots in the chapter on graphs and the chapter on correlations, so we know that the dots on a scatterplot never form a perfectly straight line. Because of this, when we plot a straight line through a scatterplot, it will not touch all of the points, and it may not even touch any! This will result in some distance between the line and each of the points it is supposed to represent, just like a mean has some distance between it and all of the individual scores in the dataset.

The distances between the line of best fit and each individual data point go by two different names that mean the same thing: errors or residuals. The term “error” in regression is closely aligned with the meaning of error in ANOVAs (standard error); it does not mean that we did anything wrong! In statistics, "error" means that there was some discrepancy or difference between what our analysis produced and the true value we are trying to get at it. The term “residual” is new to our study of statistics, and it takes on a very similar meaning in regression to what it means in everyday parlance: there is something left over. In regression, what is “left over” – that is, what makes up the residual – is an imperfection in our ability to predict values of the Y variable using our line. This definition brings us to one of the primary purposes of regression and the line of best fit: predicting scores.

Predicting Y from X

If we know that there is a linear relationship between two variables, we can use one variable to predict the other. We use this regression line to “predict” one score/variable from one other score/variable. Remember, we can't say that one variable causes the other variable to change (see Correlation versus Causation)! We are merely saying that because of the statistically significant correlation, we can use a simple equation to use one variable to predict the other.

The most common reasons to predict scores:

- Time: You want an estimate of an event that hasn’t happened yet.
  - You base the prediction on a variable that is available now.
- Expense: You want an estimate of an expensive measure.
  - You base the prediction on a variable that is cheaper.

An example related to time is the SAT. SAT scores were designed to predict successfully finishing the first year of college. A bunch of students were given a bunch of questions, then followed for a year to see who were able to pass their first year of college. The questions that were correlated with finishing the first year of college were refined to create the SAT, then used for future students to predict finishing the first year of college.

An example related to expense are the personality tests that you might complete for certain jobs. High-quality personality assessments are expensive; most companies can't pay that kind of money for each and every applicant. However, what they can do is find a cheaper personality test that is statistically significantly correlated with the expensive test. That way, the company can use the cheaper test to predict scores on the high-quality (but expensive) test. Sure, it's not as good of an assessment of personality and fit with the company or job, but the company can afford to test lots of candidates!
