13.1: 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, X and Y. Regression involves solving for the 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 chapter 2 and chapter 12, so we know by now that no scatterplot has points that form a perfectly straight line. Because of this, when we put 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 and residuals. The term “error” in regression is closely aligned with the meaning of error in statistics (think standard error or sampling error); it does not mean that we did anything wrong, it simply 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.