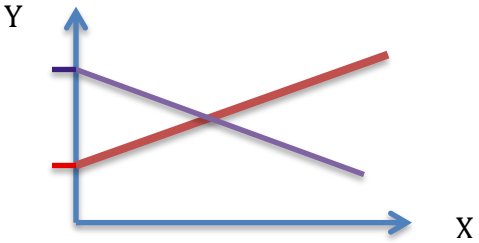
Orientation: Understanding Multiple Regression Results

To understand patterns in the universe scientists often want to know how much a measured variable corresponds to another measured variable from the same cases. For example, if we want to know if people learn more words with each passing day, it would be nice to have the number of days old a person is and how many words the person can say—two measures for each person in the study. (To protect the anonymity of the measures we will give them pseudonyms X and Y).

If there is a general increase in one measure accompanying an increase in the other measure, there is a monotonic relationship. If it is basically proportional (e.g., 3 more words per day), then it is a linear relationship. (Notice that there could be a monotonic relationship that is a square or a log or other increasing but "curved" relationships. The linear analyses discussed here will not capture anything besides linear, so if the actual relationship function had a lot of non-linear components, they would not be analyzed well with linear methods. A sine wave has a lot of up and down linear components, but would average to 0. A circle has, technically, no linear components but there is a perfect non-linear relation between X and Y for a circle).

From elementary algebra you probably remember the formula for the slope of a line:
$Y = mX + b$

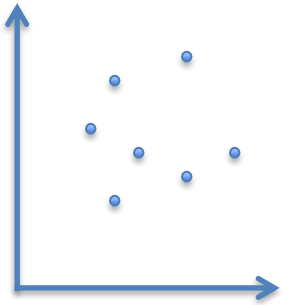


For the two lines on the left, b is what the value of X is where the line hits the Y axis—the intercept. "m" represents the steepness of the slope. The red line shows that Y increases as X does, so the slope is positive and $m > 0$. The purple line (or "purpline" for short) shows a decrease in Y as X increases, so its $m < 0$. A line with $m = 0$ is flat—Y is constant and equals b. More important about a line with no slope: knowing what X is will give you no information whatsoever about what Y is.

Now here's the thing – math is perfect. It is also fiction. It's made up to be perfect. With a linear equation you can ALWAYS know exactly what Y will be if you are told X, or vice versa. (It's like being followed around by somebody who ALWAYS knows and says the punch lines.)

Real life, though, is messy. (Unless you are a very neat person who lives solely with other neat people, imagine a "house" or "apartment." Does your actual home look like that, or is there some clutter in reality that you would have to erase to make it look like the home you imagine? We scientists have to know how to wade through the clutter without throwing out the things that would make a house recognizable, e.g., not the kitchen sink.) In real life, nothing is measured perfectly, but we still have to figure out if there are patterns lurking beneath the clutter and dust and blurry glass despite imperfect measurement. We would like to be able to know how much variable Y might be if we learn the value of variable X, and we also

want to know how much to bet on that exact figure, or leave more latitude for imprecision.



In real life, we get dirt—specks of data that don't have to line up (see left). However imperfect, can X give us a good guess about what Y is?

When we measure all the dirt specks for X and Y, then we have data. Just like the slope *m* in algebra, in statistics, the correlation coefficient *r* tells us how much (if at all) X and Y have a linear slope. If r is 0 (or close enough to it), there is no linear pattern to the relationship. If r is very close to 1.0 (and it can't get bigger than that), that means a one unit increase in X will produce a one unit increase in Y, on average. Likewise, if r is very close to -1.0 (and it can't get smaller than that), that means a one unit increase in X will produce a one unit decrease in Y, on average.

The correlation coefficient ignores the intercept or means. But if you want to know what that is and how strong the correlation is, you can get that too using "regression." (This does not mean turning into a baby so just keep paying attention!) Regression is a minimization technique for taking a bunch of observations of two variables (the eponymous X and Y) and *estimating* what the slope and intercept are. Because they are not perfect predictions, unlike algebra, they have different names than m and b: The intercept is the constant c which is estimated by the mean. The slope is known as the Greek letter beta (β). The minimization technique used to estimate β is to make the distance between each observed data point and the line – totaled all up—to be as small as possible.

Multiple regression just means that instead of having only one X variable, you have more than one (plus the Y). You already know you could correlate X1 with Y, X_2 with Y, X_3 with Y and so forth. Multiple regression means you put all of the predictor variables you have in and let them fight it out for how much each X relates to Y when the others are in the ring. You get a different regression coefficient for each predictor variable (like β_1 for X_1, β_2 for X_2, etc.).

If the predictors are strongly correlated to each other, then putting more than one predictor in might not add more (new) information. So like if X_1 and X_2 are strongly correlated, then maybe only β_1 will be non-zero because X_2 doesn't add much more information to predicting Y than X_1 does.

All this is to help you be able to read not just graphs of results, but also tables of regression coefficients. Just like with correlation coefficients (that only consider 1 X and 1 Y alone), standardized regression coefficients can range from 1 to -1, and the more the value is away from 0, the steeper the slope or stronger the linear relation between X and Y. Just like with correlation coefficients and other inferential statistics, there is usually a p-value given (such as $p < .05$) that indicates the estimated chance that the observed relationship would be a measurement accident rather than a real relationship you can count on.

Application:

In Table 2 of Howard, Blumstein & Schwartz (1986), there are columns of standardized regression coefficients under columns for the predictor variables and the rows are the predicted (Y) variables, corresponding to what influence tactics a person did to his/her partner. For the first row regarding "Manipulation" as a tactic, there are numbers very close to zero and with no asterisks for Sex of Actor (-0.07), Sexual Orientation (-.018), Relative Masculinity (-.030), or Relative Femininity (.089), meaning that none of these predictors were reliably related to using Manipulation. Sex of Target, though has a coefficient of almost .20 (.191***) with three asterisks that the Note of the table explains means $p < .001$. The note also explains that the direction of this effect is that "positive regression coefficients indicate that … partners of men are perceived to use this tactic more than partners of women." Using this "lesson" you should be able to infer that this means that men are perceived to have manipulation used against them more than women do, by about 20%. Further, because there is no reliable effect of sexual orientation ("-.018" is too close to zero to count) we can infer the former effect means that straight women are manipulating straight men more than would be expected by chance.

Reference

Howard, J. A., Blumstein, P, & Schwarts, P. (1986). Sex, Power, and Influence Tactics in Intimate Relationships. *Journal of Personality and Social Psychology, 51,* 102-109.