
Everything you wanted to know about (multiple) regression, but were afraid to ask.

1. whirlwind review of simple regression (one dependent variable [y], one independent variable [x])

- (a) “fit” a line to a scatterplot
- i. tells us the relationship between x and y
 - ii. allows us to predict the mean of y for unobserved x
 - iii. allows us to test statistically the strength of the association
- (b) this line has two parameters: the slope (β) and the intercept (α). The equation (model) we are fitting is:

$$E(y|x) = \alpha + \beta x + \epsilon$$

The $E(\cdot)$ is the expectation operator, and denotes that we are fitting the expected value, or average, of y . Since we are dealing with nonexperimental data, we write $y|x$ to note that the y values are conditional on the observed x . Don't worry about that at this point. Also don't worry too much about ϵ right now; just know that it is the error term with presumed zero average. This equation also takes some other variant forms; don't worry about that for now, either.

- (c) by “fit” we mean to minimize the sum of the squared vertical distance from each point to the regression line.
- (d) there is a unique combination of $\{\alpha, \beta\}$ that minimizes the sum of squared errors.
2. whirlwind preview of multiple regression
- (a) multiple regression is where we *simultaneously* consider two or more x variables
- (b) it is much-misunderstood

- suppose x_1 is something that a hypothesis predicts is correlated with y
- suppose further that x_2 is something we suspect could cause a spurious correlation between x_1 and y
- provided x_2 is observable, we would like to know the correlation between x_1 and y , net of x_2 . That is to say, also taking x_2 into account, we wish to know the correlation between x_1 and y . The language used for this is “controlling for x_2 ”, which is misleading because in nonexperimental data nothing is controlled at all. Sometimes one reads the phrase “conditioning on x_2 ”, which is more statistical and somewhat less misleading in my opinion.

- This time the equation (model) we are fitting is:

$$E(y|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Think of an aquarium (a three-dimensional space), with x_1 (education) along one axis and x_2 (age) along another. Everyone has an age and an education, and so lies at some point of that aquarium. Their income is represented by their height in the aquarium.
- Dealing with age and education simultaneously is like dealing with observations distributed throughout that 3D aquarium instead of in a 2D scatterplot. Yet-higher dimensions are conceptually no different but are hard to visualize.
- People always talk about the regression “line”, but the model we are fitting in the case of two independent variables is that of a plane. *In multiple regression there is no regression line, no matter what anyone tells you.* There is a regression plane (or, for higher dimensions, a hyperplane). We are finding some plane, described by $\{\alpha, \beta_1, \beta_2\}$ that describes the data better than any other plane. It may be that the age-education-pay relationship is not planar, but the whole point is that we are taking a complex data set and making a simpler model, from which to understand better. Later on, we will learn about ways to improve the fit of our models, but for right now, we assume that there is some cloud of points in the aquarium and we are running a plane through them, just as we ran a line through the 2D cloud of points.
- This sheds some more light on what we *really* mean by the relationship between x_1 and y “holding x_2 constant”. Imagine our data-aquarium. Salary goes up with age and up with education. So the plane is slanting upward in two dimensions. But it is still a flat plane, so certain relationships are fixed. Imagine an ant walking up the plane. The ant is a sociologist-ant so it always walks in a straight line. It walks up education but holds age fixed. Or it walks up age and holds education fixed. The ant will rise β_1 units in height (salary) up the aquarium for every unit rise in education if it walks holding age constant, and β_2 units in height (salary) for every unit rise in age if it walks along a line of constant salary. Because of the nature of the model we have fitted, *it does not matter which level of education the the ant holds constant while it walks up in age, or vice versa.* The mathematical property of the plane is that the ant may choose *any* age- or education-invariant line to walk up and the other category will change by β units. Thus we may speak of “holding constant”. The β s in the multiple regression model will not have the same numerical value as the β s from each simple regression model run separately, though in practice the differences may be modest.