

Computing and Using the Linear Regression Formula

In a prediction situation, you are trying to use a known value of one variable as a basis for estimating (predicting) the unknown (but desired) value of a second variable. In many prediction situations you want to use some piece of information you know **now** to predict the value of a variable that hasn't happen yet. The variable you know is called the **predictor**, and the variable you want to estimate the value of is called the **criterion**. You will now learn how to use linear regression to estimate (predict) a person's value for the criterion variable from their value on the predictor.

The term "linear regression" is used because the "model" used to predict the criterion from the predictor is a straight line (the same straight line referred to in the linear relationship examined using Pearson's correlation). As you may remember from algebra, you need two pieces of information to define a straight line:

- 1) REGRESSION COEFFICIENT -- which way the line is "leaning" and how far
or SLOPE -- usually represented as "b"
- 2) Y-INTERCEPT -- the value of the criterion when the predictor has a value of 0.00
-- usually represented as "a"

So, when doing linear regression, your first task is to determine the SLOPE and Y-INTERCEPT of the regression line. The second task is to apply the regression line to predict the value of the CRITERION.

Here are some other common examples of prediction situations: Graduate schools would like to admit those students who "will do well in graduate school". Unfortunately, admissions committees cannot tell the future any better than the rest of us. So, they might select some predictor and, if it has a strong correlation with the criterion, use it to help select students. (To do this properly, they would have to admit everyone that wanted into their program (say for one year), so that they would have a sample for which they would have both the predictor and the criterion.) As another example, a developmental psychologist might want to be able to predict "social competency" of high-school-age students from a measure taken during 6th grade. To do this the researcher would have to collect data from some 6th graders (the predictor), and wait the 4-6 years until they were in high school to collect the criterion, and then perform the analysis. As you can see, prediction research often requires a great deal of time and effort, and there are no promises that researchers will be able to select good predictors.

ONE VERY IMPORTANT THING, which is often overlooked: In order to perform a linear regression you need a sample of subjects for which you know the value of BOTH the predictor and the criterion. Then, after you have calculated the regression line, you can apply it to other subjects for which you know only the value of the predictor. Using our example, you would need to collect two pieces of information from each subject: 1) how many hours they worked (the predictor) and 2) how many items they sold (the criterion). Then you could compute the regression formula to predict the criterion for other workers.

A SECOND VERY IMPORTANT THING, which is also often overlooked: In order for the regression line to be useful for prediction there must be a substantial relationship between the predictor and the criterion. Take the extreme case -- if there were no linear relationship between the predictor and criterion (a nonsignificant Pearson's correlation), it would be "silly" to try to predict one from the other. Thus, the MINIMUM requirement for a useful regression is a significant Pearson's r . Ideally you would like there to be a strong linear relationship, with r 's of .85 to .95 (or -.85 to -.95, remember that a strong linear negative relationship is just as good a basis for prediction as a strong positive linear relationship).

Usually you will want to calculate the Pearson's correlation between the predictor and the criterion before calculating the elements of the regression line (remember, if there is no linear relationship, linear regression is senseless). So, the steps below follow the steps given earlier to compute Pearson's Correlation (and obtain the necessary values from those calculations).

Research Question: The researcher wants to be able to predict (estimate) the quality of fish (fishgood is the criterion) from the number of fish a store has on display (fishnum is the predictor).

Steps to Compute the Linear Regression Equation

Step 1. Compute the slope (b)

$$\begin{aligned} b &= \text{index of covariation} / (\text{variation of } X)^2 \\ &= -2332 / 110.36^2 = -2332 / 12179.33 = -.19 \end{aligned}$$

Step 2 Compute the mean of the CRITERION

$$\bar{Y} = \Sigma Y / n = 80 / 12 = 6.67$$

Step 3 Compute the mean of the PREDICTOR

$$\bar{X} = \Sigma X / n = 287 / 12 = 23.92$$

Step 4 Compute the Y-intercept (a)

$$a = \bar{Y} - (b * \bar{X}) = 6.67 - (-.19 * 23.92) = 6.67 - (-4.54) = 11.21$$

Steps to Use the Regression Equation for Prediction

Once you have the elements of the regression line (b and a), you can use them to predict the criterion variable score (symbolized as Y') for any given predictor score (X).

Step 5 Compute predicted score using the formula

$$Y' = (b * X) + a$$

If a store had 20 fish the estimated fishgood rating for that store would be

$$Y' = (-.19 * 20) + 11.21 = -3.8 + 11.21 = 7.41$$

If a store had 40 fish, the estimated fishgood rating for that store would be

$$Y' = (-.19 * 40) + 11.21 = -7.6 + 11.21 = 3.61$$