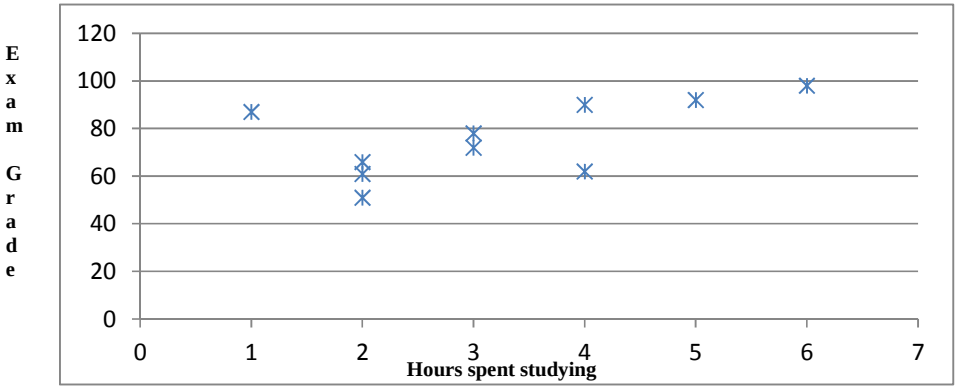
16 –Correlation and Regression

Two-dimensional data is data with two associated values for each data point.

Example 1: In a recent survey, students were asked how many hours they spent preparing for an exam. The table shows the results, along with the exam grades. Each row in this table corresponds to a specific student.

Hours (x)	Grade (y)
5	92
6	98
3	72
2	51
1	87
4	90
2	66
4	62
3	78
2	61

A scatterplot shows two variables on a 2-dimensional grid, with emphasis on their association. Each data point is shown as a point on the scatterplot. The scatterplot for the data to the right appears below.

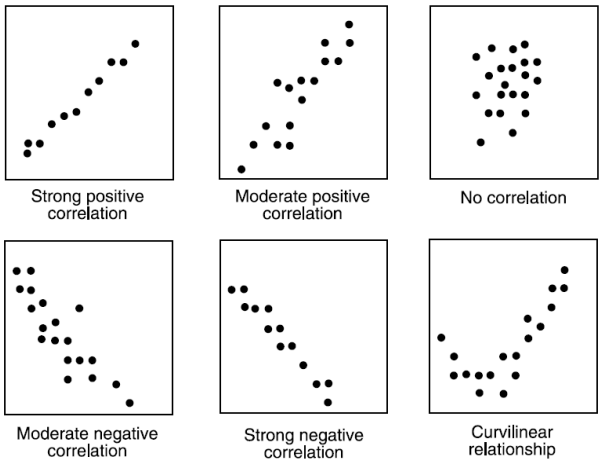


Correlation is the quantification of the strength of the relationship between the variables in 2-dimensional data.

In general, when larger values of one variable correspond to larger values of the other variable (and smaller correspond to smaller), it is a positive correlation. When larger values of one variable correspond to smaller values of the other variable (and smaller correspond to larger), it is a negative correlation.

For example, an adult's strength naturally decreases with age, and so the age and strength of adults will generally have a negative correlation.

The scatterplots to the right illustrate the differences between positive and negative relationships, between strong, moderate, and weak relationships, and between linear and non-linear (aka, curvilinear) relationships. In this course, we only study linear relationships.



The most commonly used value for correlation is called *Pearson's Correlation Coefficient* (r) and its computation is not worth doing by hand (though it is $r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n(\sum x^2)-(\sum x)^2][n(\sum y^2)-(\sum y)^2]}}$ in case you want to try). The value of r will always lie between -1.0 and 1.0. Values of r near -1.0 indicate a strong negative correlation, near +1.0 indicate a strong positive correlation, and near zero indicate either no correlation or a weak one.

The determination of a sufficient value to suggest significantly strong correlations is simple with the help of table I (textbook page 655). This table eliminates the need for a hypothesis testing protocol for correlation. The table gives threshold values of r to indicate significant correlation at $\alpha = 0.05$ and $\alpha = 0.01$ (95% and 99% confidence levels, respectively). The instructions at the top of Table I are very clear, and for 2-dimensional data, $df = n - 2$.

Example 1, continued: In the data above, $r = 0.5779$. Since there are 10 data points, $df = 8$. In the $df = 8$ row of table I, the threshold values for $\alpha = 0.05$ is 0.632 and for $\alpha = 0.01$ it is 0.765. Since the value of r is smaller than both of these thresholds, we cannot reject the hypothesis that the data are not correlated. (in other words, we cannot suggest with a minimum of 95% confidence that the hours spent studying is related to exam score).

Regression is a technique for developing a formula for a graph that comes as close as possible to all of the points in a scatterplot. We will only consider straight line regression with the "*least-squares regression line.*"

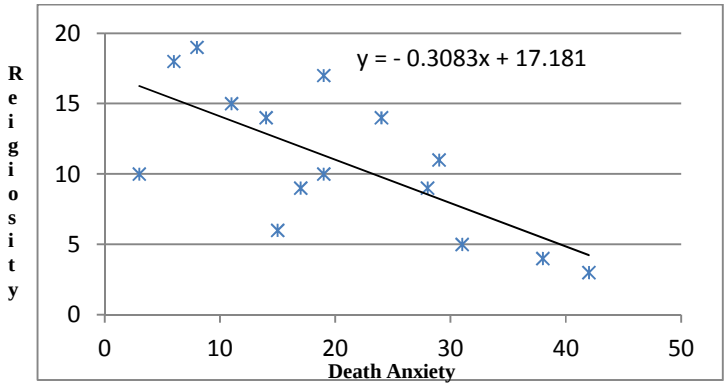
The computation of the formula for the least-squares regression line is also not worth doing by hand (the formula in slope-intercept form is $= \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2}\ x + \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2}$), but the important information that comes out of the formula is the slope and intercept of that line. These can be used to estimate values of either variable, but it is only appropriate to do so when the data are significantly correlated (using table I).

Example 2: Fifteen people were surveyed for a sociology study. Their responses to several questions in the survey were used to rate each person's anxiety about dying as well as their level of religiosity (how religious they are). Pearson's correlation coefficient is computed for these data: $r = -0.696$.

Since $df = n - 2$, we use the $df = 13$ row of table I, and it is clear that these data are correlated at a significance of $\alpha = 0.01$. So, we can be 99% confident that the data are negatively correlated, and can use the least-squares regression formula to make estimations.

The least-squares regression line for the data is computed and the result is

y = -0.3083x + 17.181. The scatterplot, along with this line, is shown below.



Death Anxiety	Religiosity
3	10
6	18
8	19
11	15
14	14
15	6
17	9
19	10
19	17
24	14
28	9
29	11
31	5
38	4
42	3

Using this formula, we can estimate the religiosity level of a person with a death anxiety of 50. It would be *y = -0.3083(50) + 17.181= 1.76.*

Similarly, a person with death anxiety level 25 would be expected to have religiosity level *y = -0.3083(25) + 17.181= 9.47.*

It is also possible to estimate the death anxiety level of a person with religiosity level of 0 by solving the equation *0 = -0.3083x + 17.181.* The solution to that equation is 55.7.