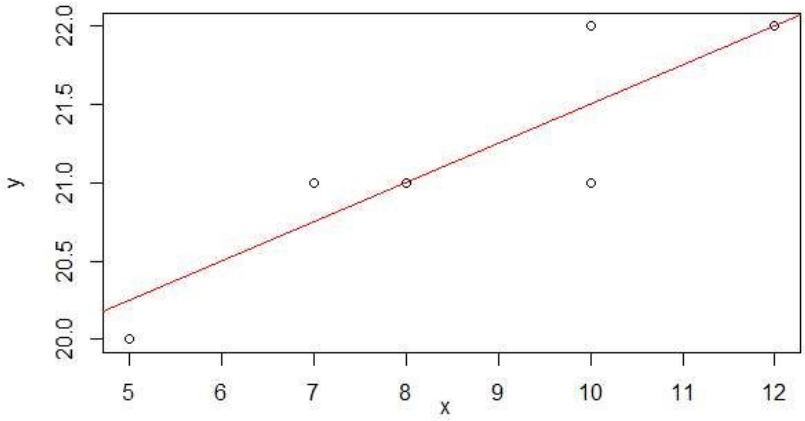
Topic 8a: Linear Regression and Correlation: Part 2

In the previous video we saw that we could compute the "sum of the squared differences between the observed and expected values" so that we could tell if one proposed line is a "better fit" than is a different proposed line. The lower that sum is the better the line is at approximting the observed points. Here is a computation for a set of points and a given line.

	x	5	7	8	10	10	12	
observed	y	20	21	21	22	21	22	
expected for y=0.25x+19		20.25	20.75	21.00	21.50	21.50	22.00	
observed - expected		-0.25	0.25	0.00	0.50	-0.50	0.00	
(observed-expected)^2		0.0625	0.0625	0.0000	0.2500	0.2500	0.0000	sum = 0.625

Here is a plot of that data and the line.

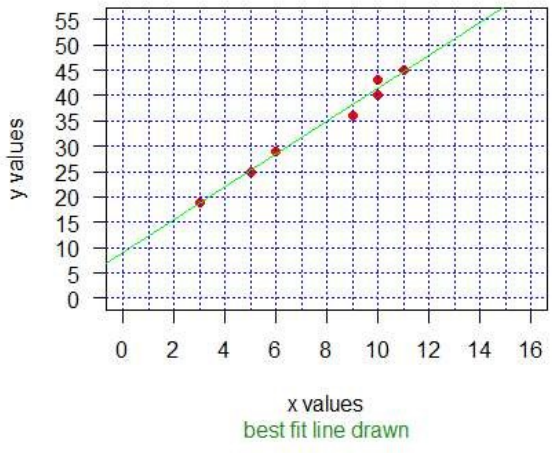


The value that we computed, 0.625 in this case, is good for comparing proposed lines. But that value is determined by the number of data points that we have and by the scale of our values. Furthermore, what we really want is a measure of how well the best fit line matches the data. That new measure is called the correlation coefficient. In the notes you can find how to compute the correlation coefficient but here we just want to see what it means.

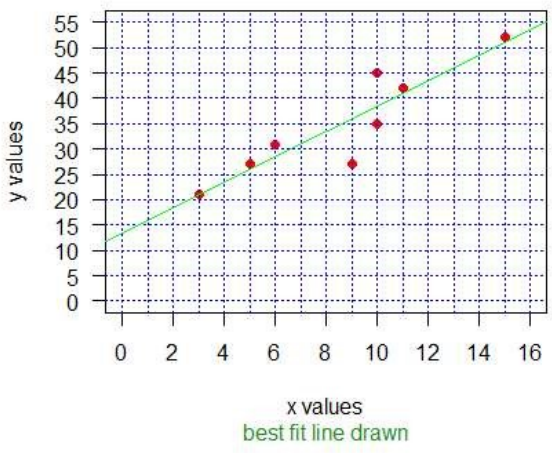
The correlation coefficient, usually represented by r is number that is always between -1 and 1. The closer the value r is to -1 or to 1 the more the points will be close to the line.

The correlation coefficient for the data above is r = 0.8844364. Let us see some other data and the related r values.

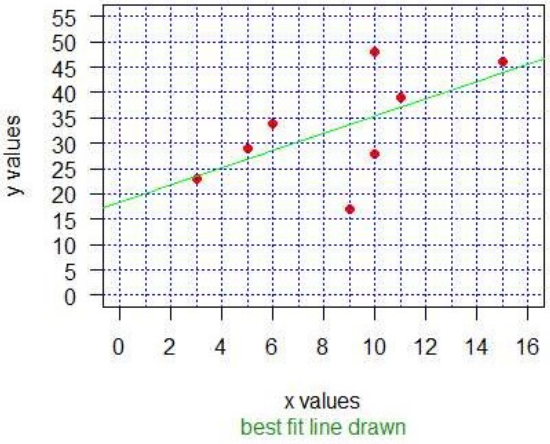
example with r= 0.9954



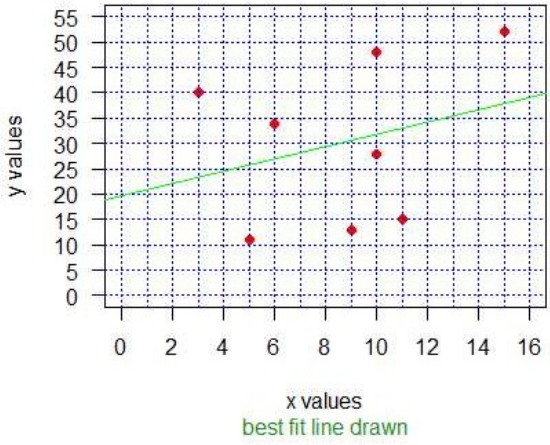
example with r= 0.9022



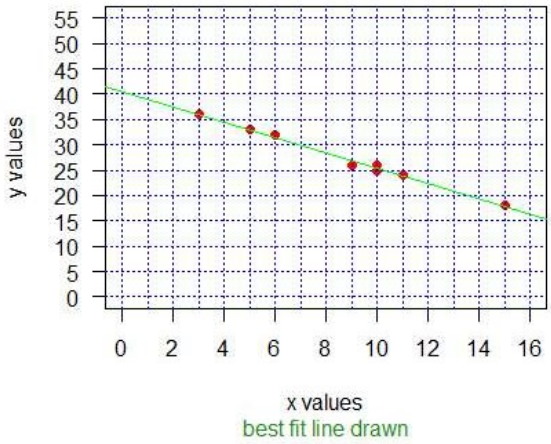
example with r= 0.5957



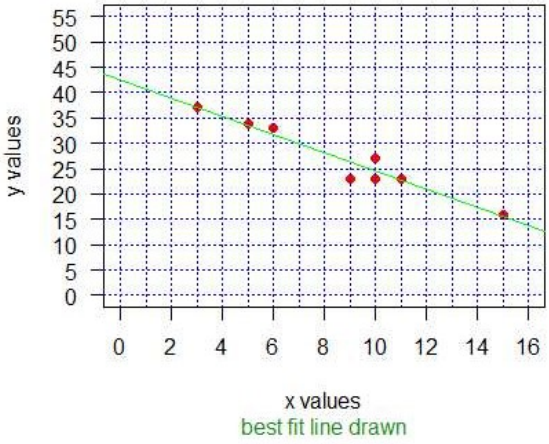
example with r= 0.2855



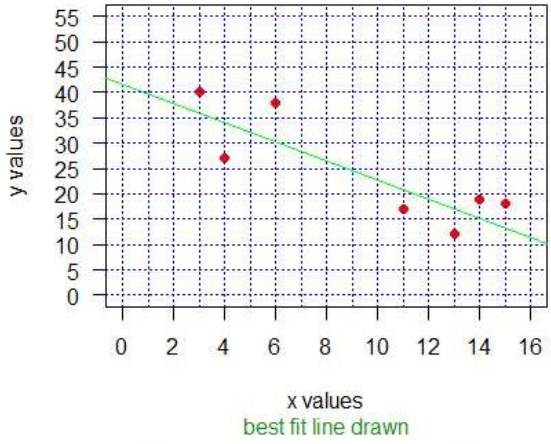
example with r= -0.9964



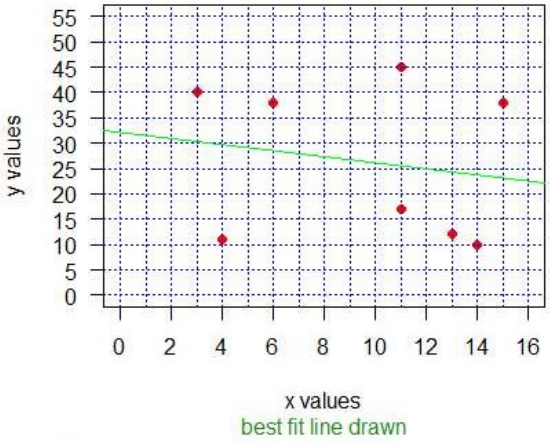
example with r= -0.9689



example with r= -0.8451



example with r= -0.1842



Please note that the square of the correlation coefficient, r^2, tells us the percent of the variability in the data that is explained by the linear model. Thus, a r value of -0.8451 means that the linear model explains about 71.4% of the variability of the data since r^2=0.71419401.