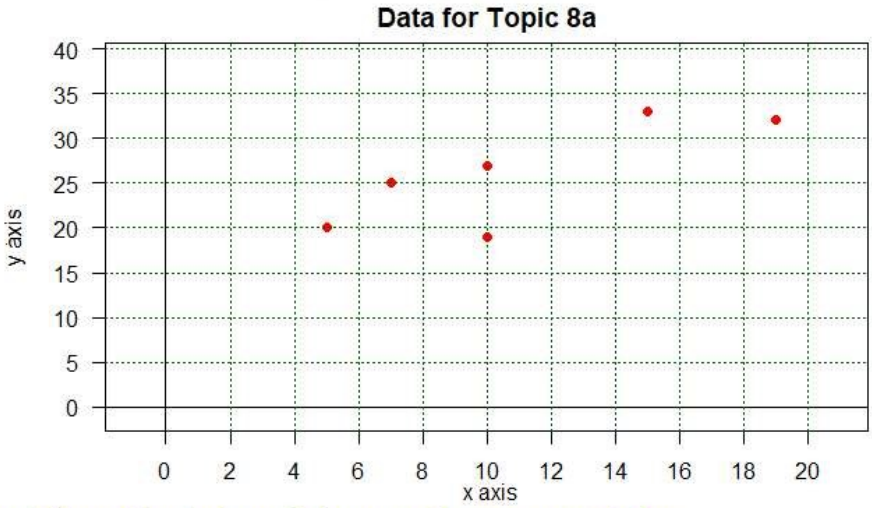
Topic 8a: Linear Regression and Correlation

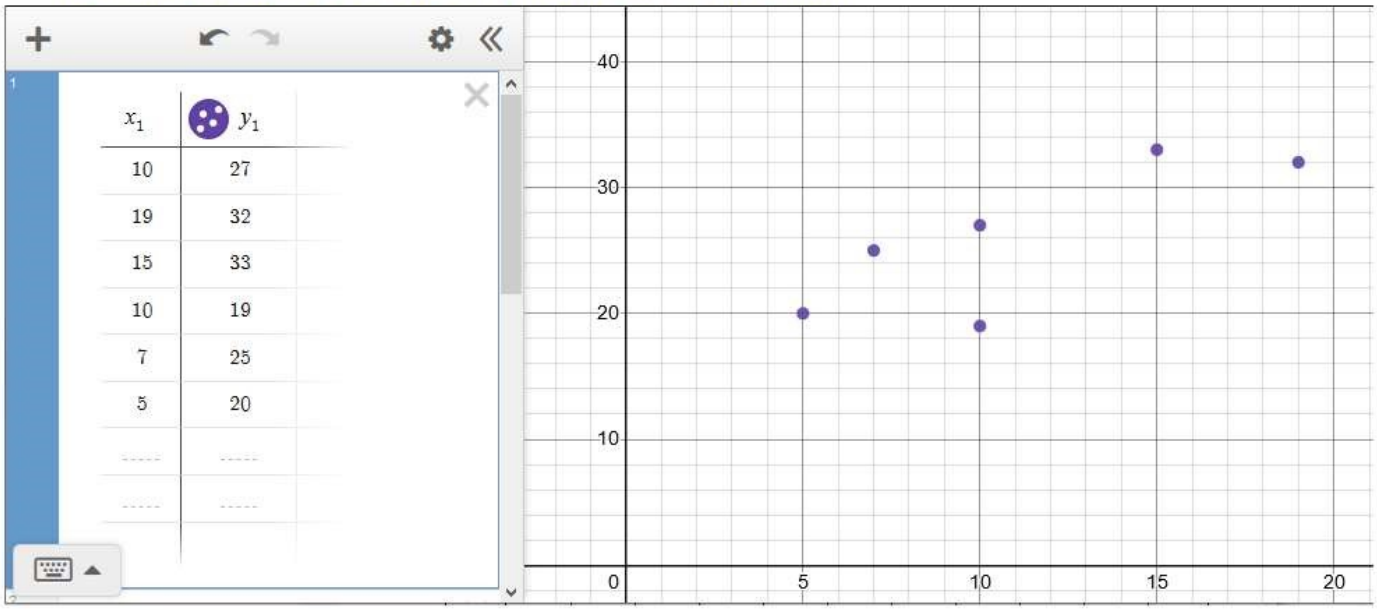
We start with ordered pairs of values. In algebra we called these coordinates of points on a graph.

x	10	19	15	10	7	5
y	27	32	33	19	25	20

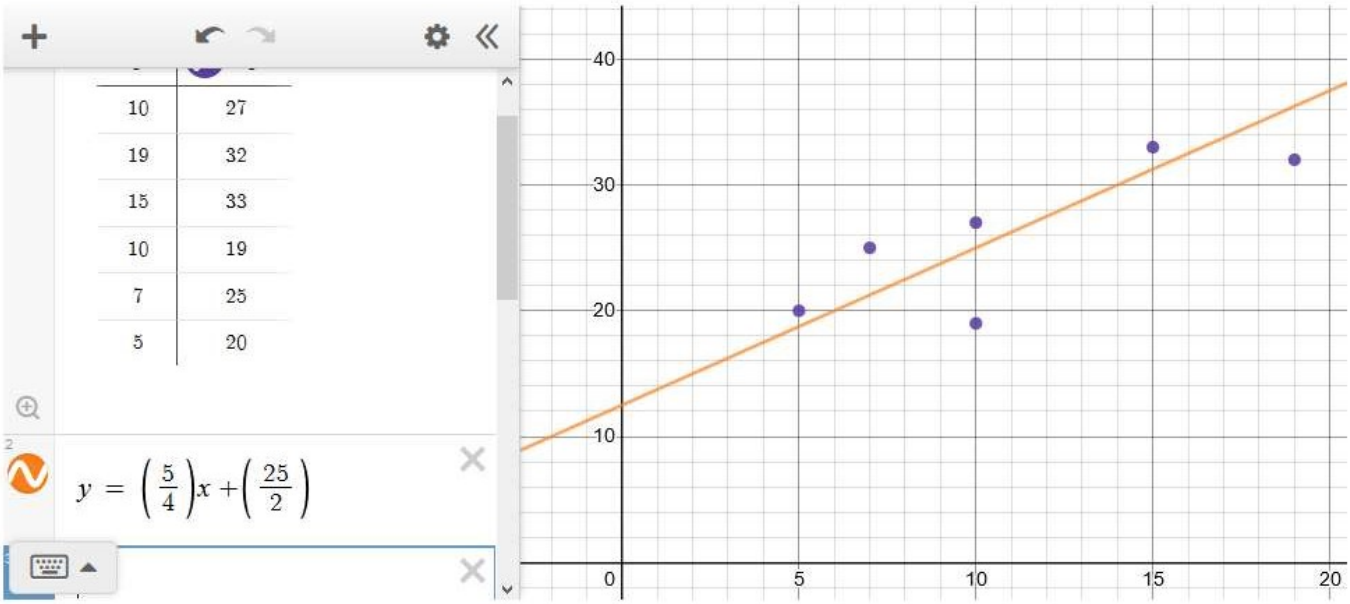
We could generate a plot of these pairs as



That same information is shown below on a desmos.com window.



Either way, there seems to be a linear relation between the x and the y values. That is, it looks like we could find values for m and c such that the line y = mx + c would be close to all of the points that we graphed. Clearly, we cannot hit all of the points with one straight line, but we might be able to come close. Just to try one line, consider the line through the points (6,20) and (18,35). That line has slope m=(35-20)/(18-6)=15/12=5/4, and y-intercept c=25/2=12.5.



How do we measure the "goodness" of this line? That is, what can we use to compare the "goodness" of this line to the "goodness" of a different line. To do this we compute a single value from our data. First, we call all of the y values in our table the "observed" values. Then, we find the y values that our current equation predicts. That is, we find the y value that is given by the equation when we substitute each of the x values. We call those the "expected" values. Now we have a new table.

	x	10	19.000	15	10	7	5
observed	y	27	32	33	19	25	20
expected for y=1.25x+12.5		25.00	36.25	31.25	25.00	21.25	18.75

Then we find the difference between the observed values and the expected values. In texts this is often shown as O - E.

	x	10	19.000	15	10	7	5
observed	y	27	32	33	19	25	20
expected for y=1.25x+12.5		25.00	36.25	31.25	25.00	21.25	18.75
observed - expected		2.00	-4.25	1.75	-6.00	3.75	1.25

Then, to get rid of the negative values and to give more importance to the observed values that are far from the expected values, we square the observed - expected values.

	x	10	19.000	15	10	7	5
observed	y	27	32	33	19	25	20
expected for y=1.25x+12.5		25.00	36.25	31.25	25.00	21.25	18.75
observed - expected		2.00	-4.25	1.75	-6.00	3.75	1.25
(observed-expected)^2		4.0000	18.0625	3.0625	36.0000	14.0625	1.5625

Finally, we get the sum of those squared values.

sum = 76.75

That becomes the measure of how "good" this line fits the data. If we look at a different line, perhaps y=2x+3, then we get different expected values, and thus a different sum of the (observed-expected)^2 values.

	x	10	19.000	15	10	7	5	
observed	y	27	32	33	19	25	20	
expected for y=2x+3		23.00	41.00	33.00	23.00	17.00	13.00	
observed - expected		4.00	-9.00	0.00	-4.00	8.00	7.00	
(observed-expected)^2		16.0000	81.0000	0.0000	16.0000	64.0000	49.0000	sum = 226

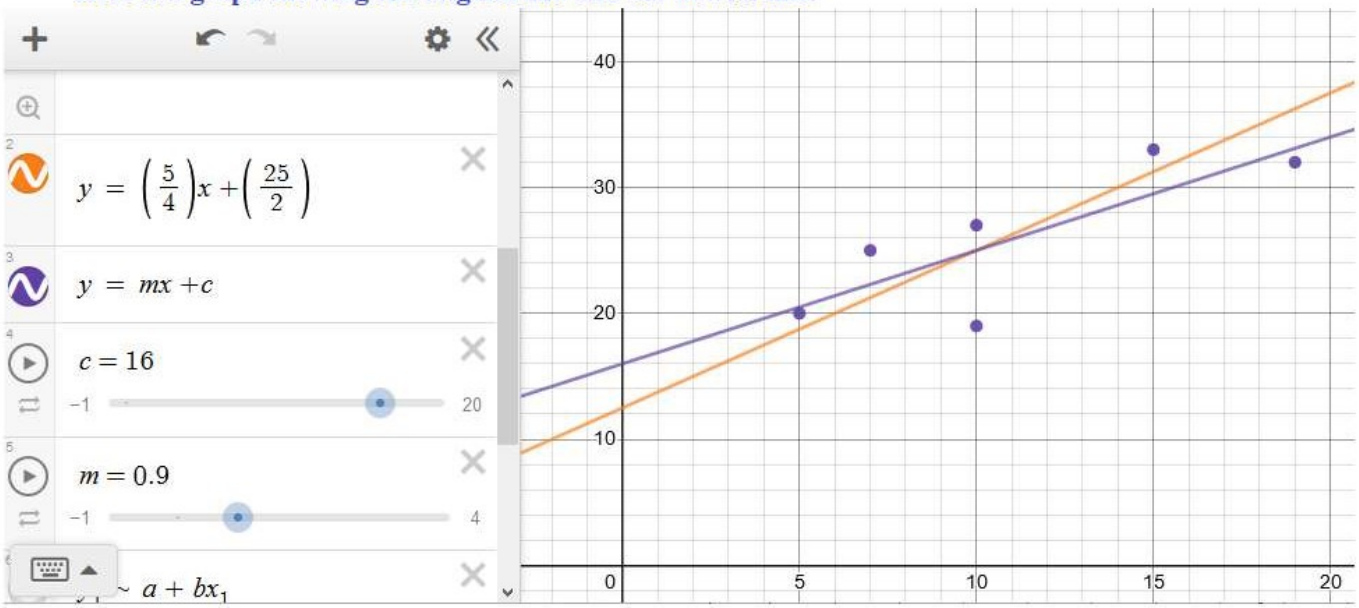
The value for this line, 226, is higher than was the value of our first line, 76.75. Therefore, the first line, the one with the lower sum of the (observed-expected)^2, is the better line.

Let us look at another line, y = 0.9x + 16. This gives us a new table.

	x	10	19	15	10	7	5	
observed	y	27	32	33	19	25	20	
expected for y=0.9x+16		25.00	33.10	29.50	25.00	22.30	20.50	
observed - expected		2.00	-1.10	3.50	-6.00	2.70	-0.50	
(observed-expected)^2		4.0000	1.2100	12.2500	36.0000	7.2900	0.2500	sum = 61

This is the lowest sum of the (observed-expected)^2 so far, so this is now the best line we have tried.

Here is a graph showing the original line and our newest line.



We now have a method for comparing the "goodness of fit" for proposed lines given our original values. The question becomes how do we find the line that has the very best fit. The sort of good news is that we can use the original data to compute the best values for the slope, m, and the y-intercept, c. The bad news is that we would have to get through the material of Calculus III to prove that our computations produce the best line. The far better news is that there is a single statement in R that does all of this work and more.