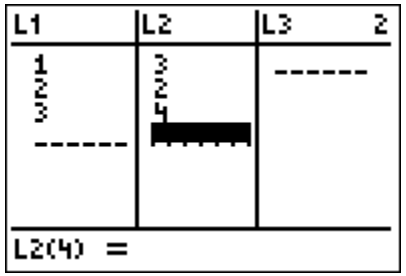
LINEAR REGRESSION AND CORRELATION

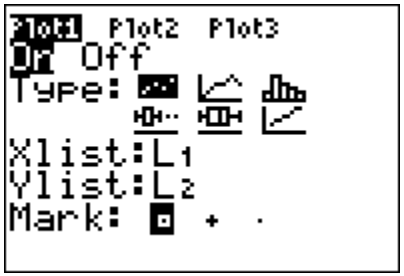
Suppose we have the following paired data, also known as *bivariate data*.

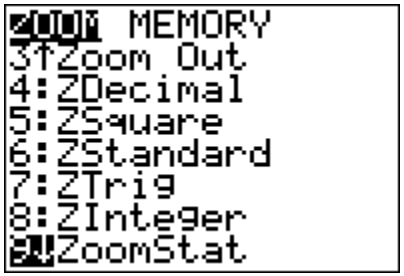
$$(1,3),\ (2,2),\ (3,4)$$

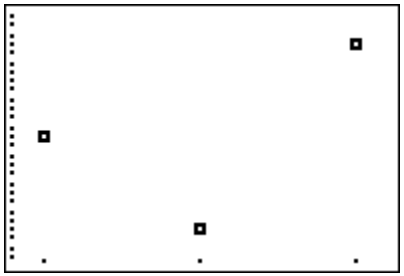
We can enter this data into lists in our calculator and create a scatterplot.

$$(1,3),\ (2,2),\ (3,4)$$



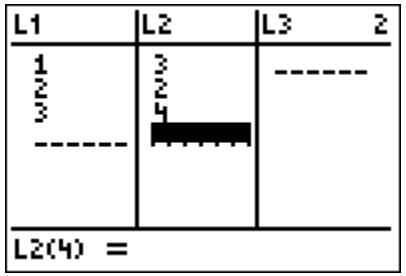


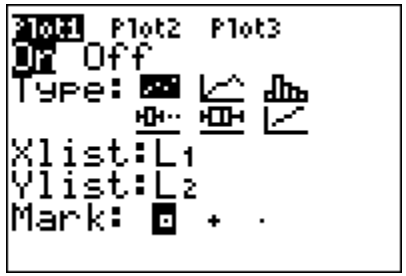


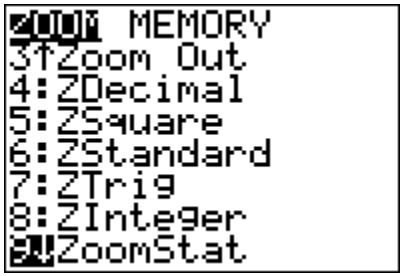


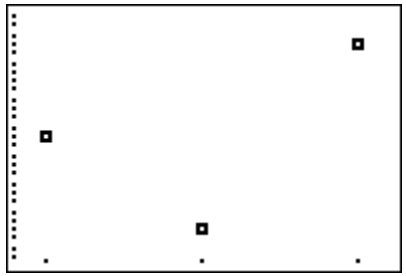
We can also use our calculator to find the best fitting straight line for this data.

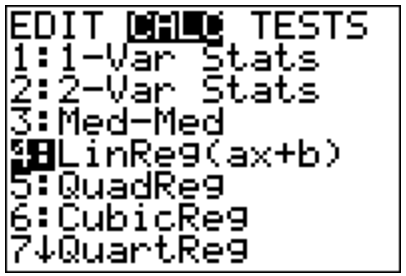
$$(1,3),\ (2,2),\ (3,4)$$

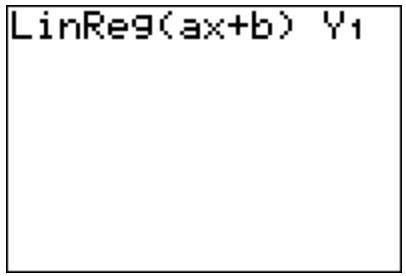


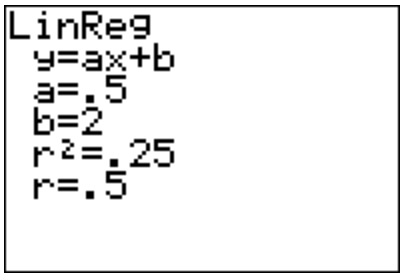


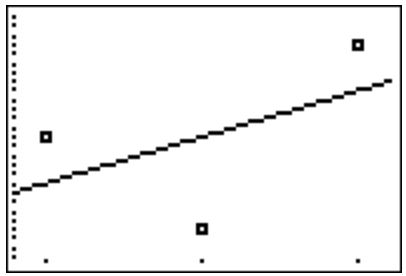










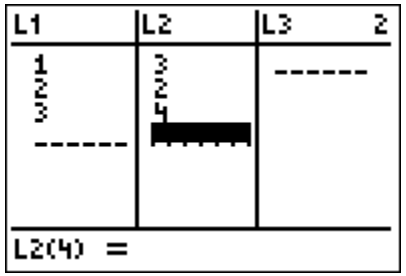


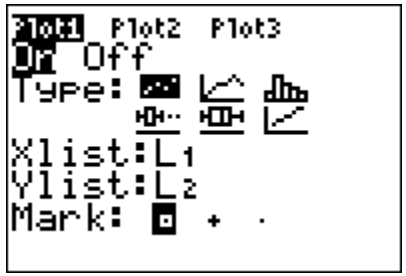
$VARS \rightarrow Y-VARS$

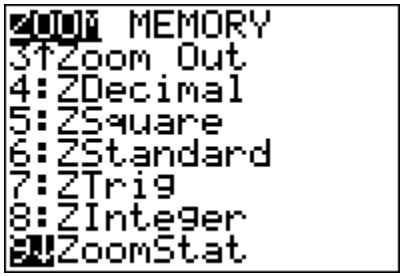
$\rightarrow FUNCTION \rightarrow Y_1$

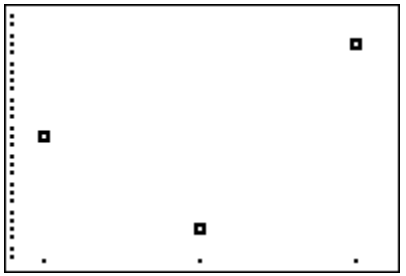
And now we have much to discuss.

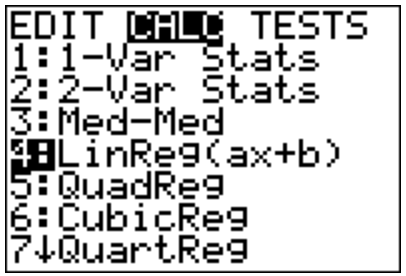
$$(1,3),\ (2,2),\ (3,4)$$

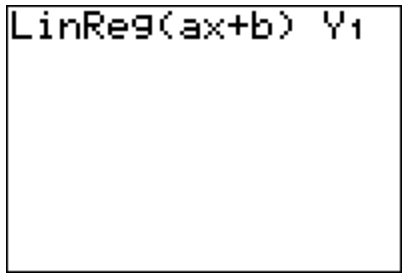


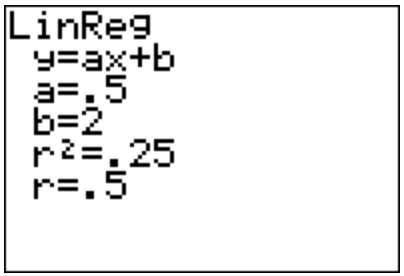


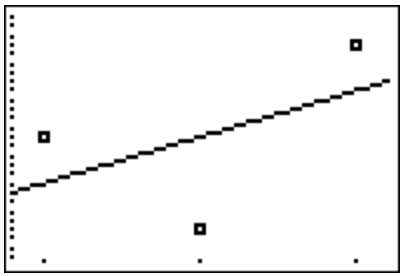








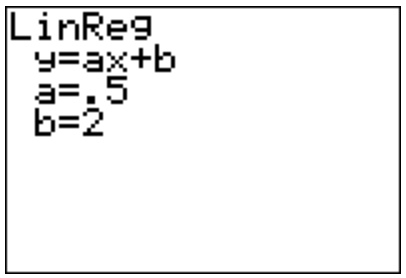




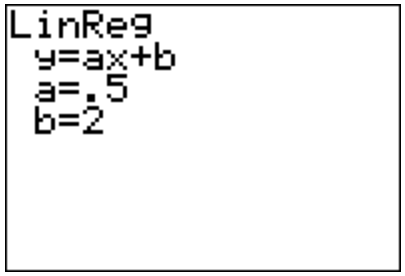
$VARS \rightarrow Y-VARS$

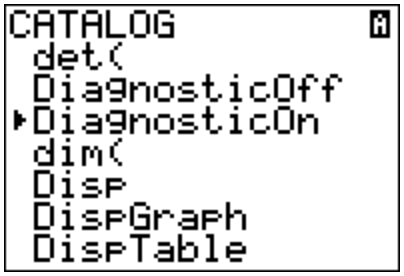
$\rightarrow FUNCTION \rightarrow Y_1$

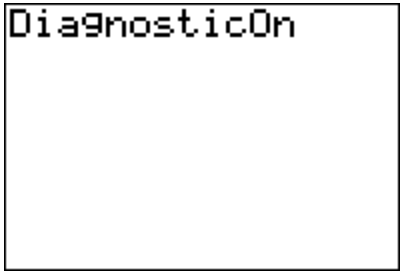
If you don’t see the values for r and r^2 on your screen, then do this:

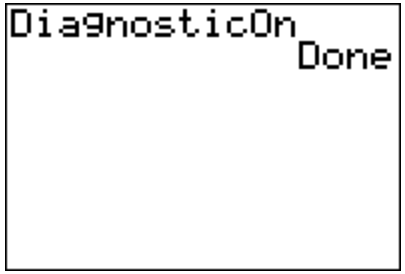


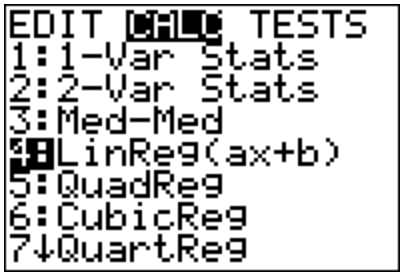
If you don't see the values for r and r^2 on your screen, then do this:

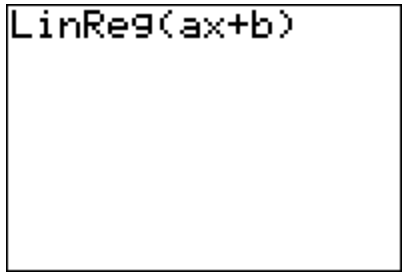


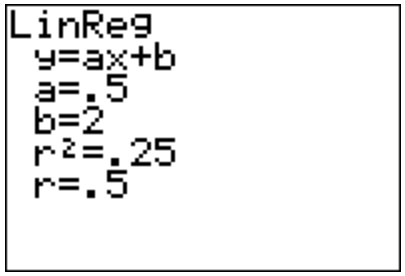




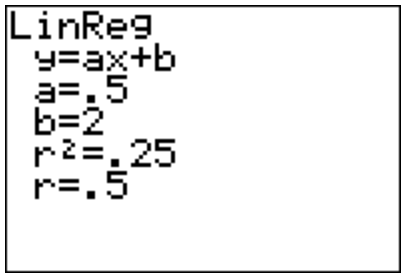




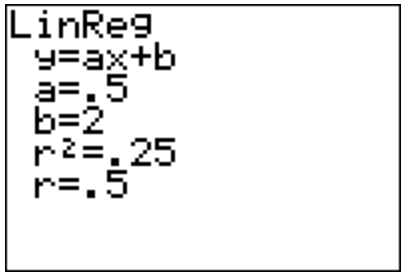




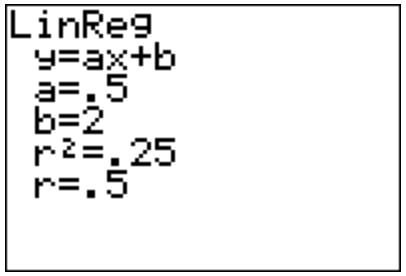
The number *r* is called the *coefficient of linear correlation*, or the *pearson product moment correlation coefficient* (among other things).



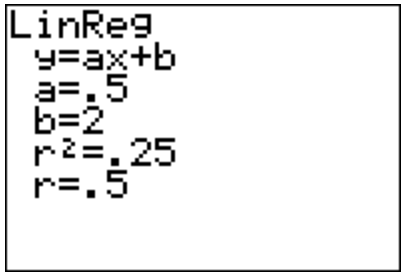
This number measures the strength of the linear or straight line relationship between the two variables.



We always have that $-1 \leq r \leq 1$.



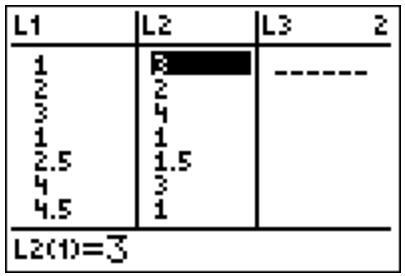
The closer |*r* | is to 1, the stronger the linear relationship, and the closer |*r* | is to 0, the weaker the linear relationship.



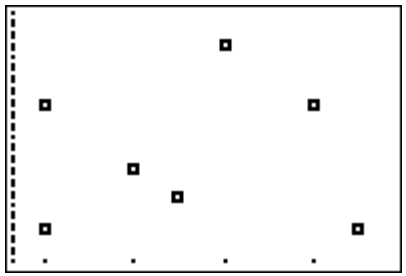
Some statisticians use the following classification scheme:

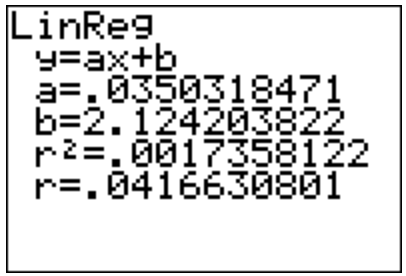
$$\begin{aligned} \text{Weak:} \quad & 0 \le |r| < .4 \\ \text{Moderate:} \quad & .4 \le |r| < .7 \\ \text{Strong:} \quad & .7 \le |r| \le 1 \end{aligned}$$

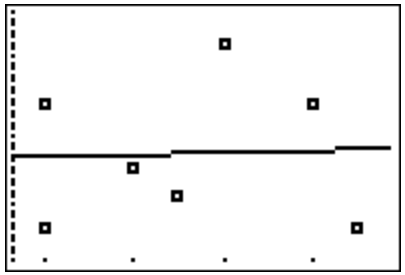
Example 1:


L1 L2 L3 2
1 3
2 2
3 4
1 1
2.5 1.5
4 3
4.5 1

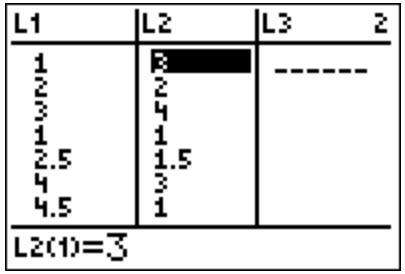
L2(1)=3

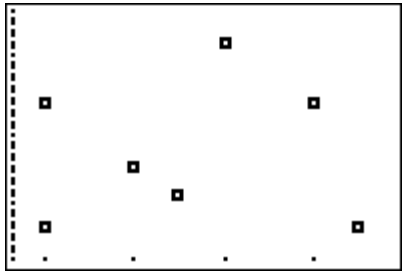


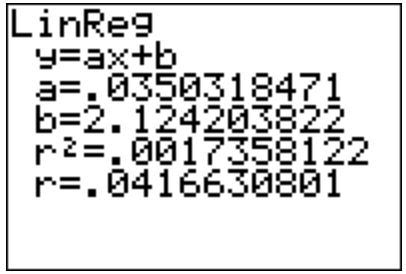

LinReg
y=ax+b
a=.0350318471
b=2.124203822
r²=.0017358122
r=.0416630801

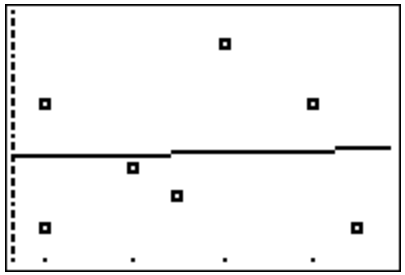


Example 1:



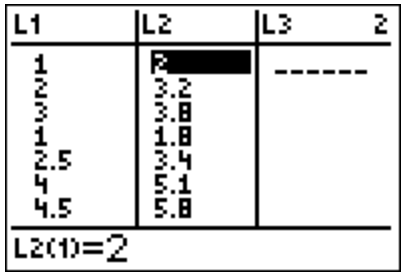




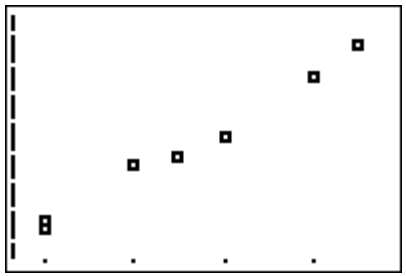


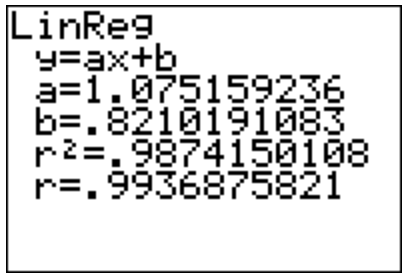
The linear correlation is very weak.

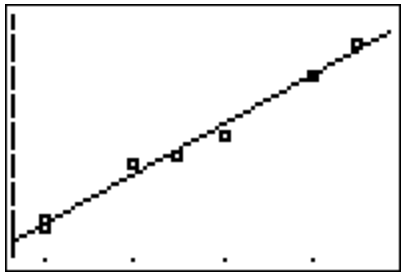
Example 2:


L1 L2 L3 2
1 2
2 3.2
3 3.8
1 1.8
2.5 3.4
4 5.1
4.5 5.8

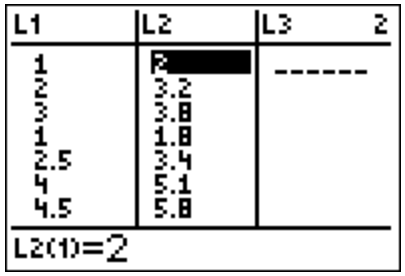
L2(1)=2

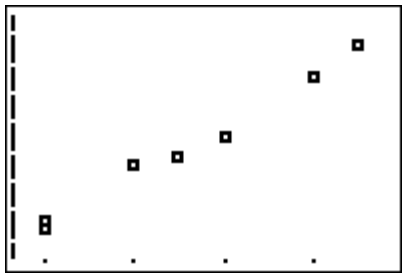


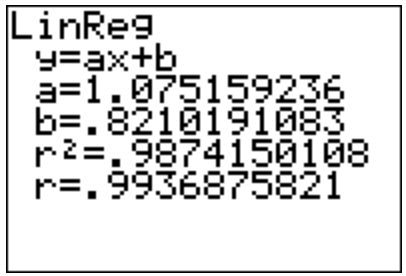

LinReg
y=ax+b
a=1.075159236
b=.8210191083
r²=.9874150108
r=.9936875821

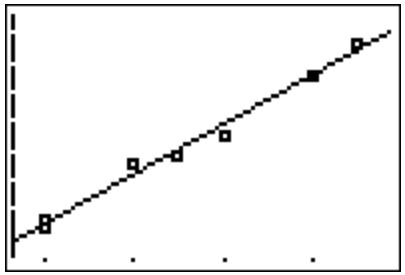


Example 2:





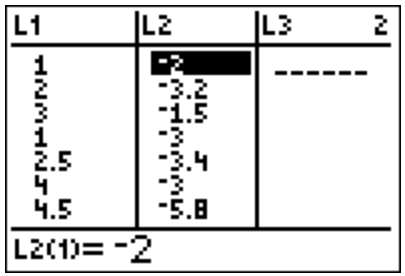


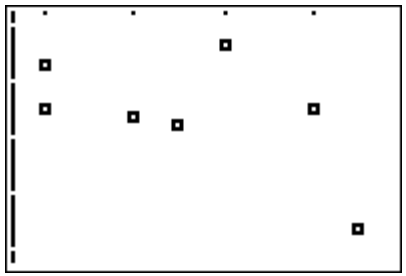


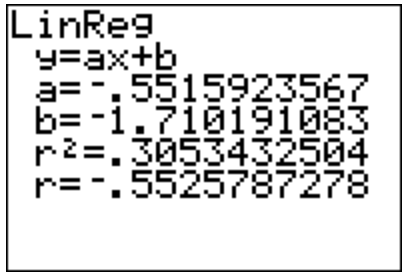
This is a strong, positive linear correlation.

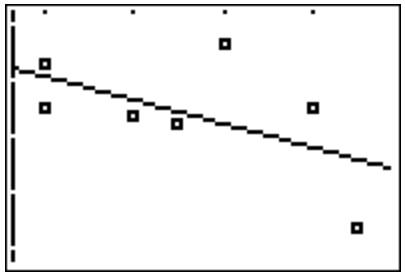
As the input variable increases, the output variable also increases.

Example 3:

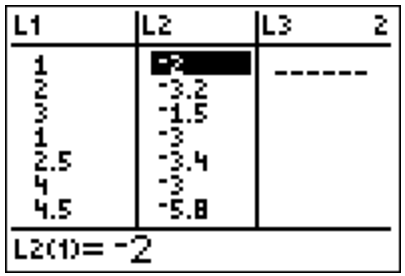

L1 L2 L3 2
1 -2 ------
2 -3.2
3 -1.5
1 -3
2.5 -3.4
4 -3
4.5 -5.8
L2(1)= -2

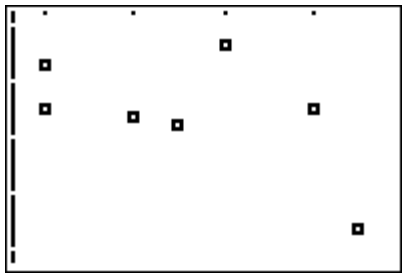


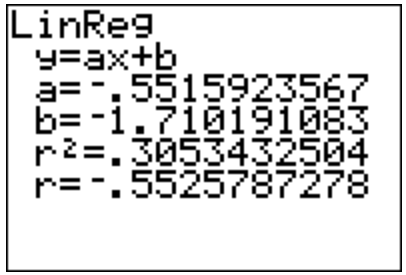

LinReg
y=ax+b
a=-.5515923567
b=-1.710191083
r²=.3053432504
r=-.5525787278

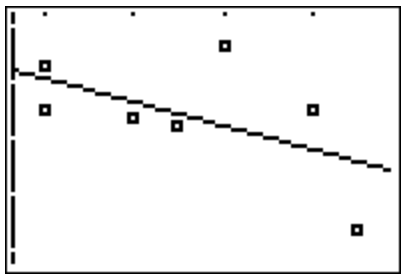


Example 3:



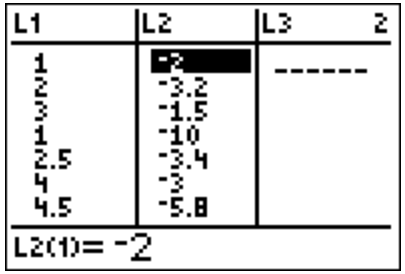


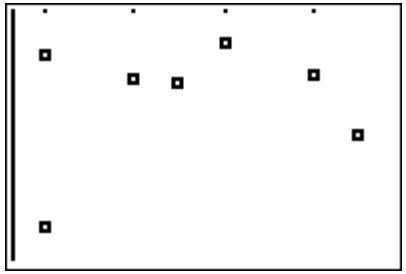


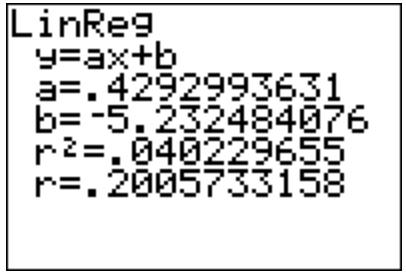


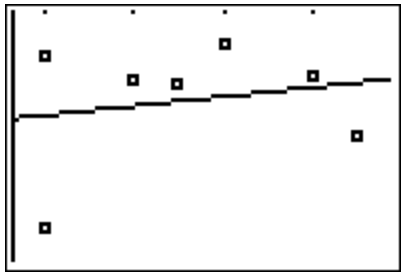
This is a moderate, negative linear correlation. As the input variable increases, the output variable decreases.

Example 4:



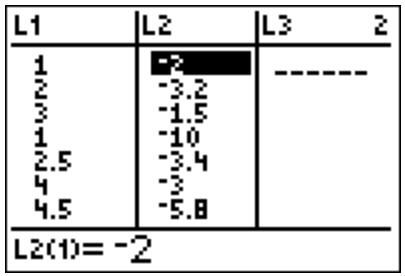


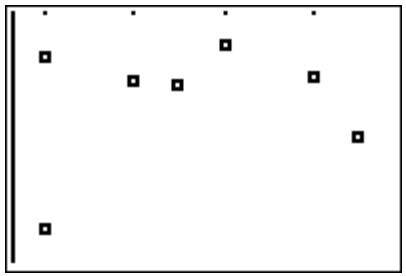


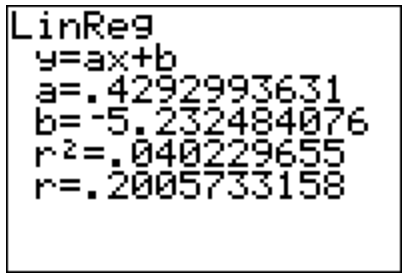


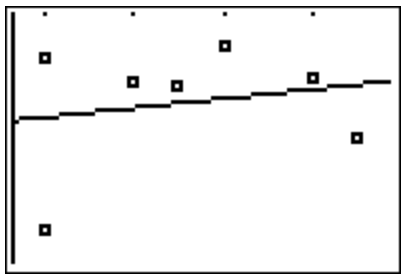
In this data set we changed just a single point, and it changes the correlation from moderate & negative to weak & positive.

Example 4:





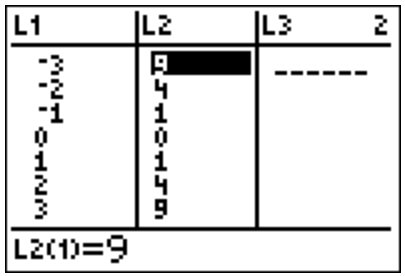


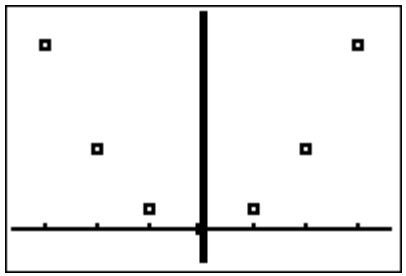


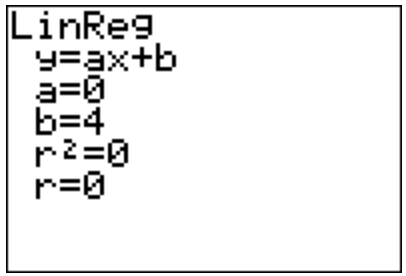
In this data set we changed just a single point, and it changes the correlation from moderate & negative to weak & positive.

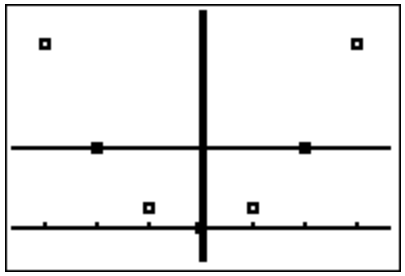
A single data point can strongly affect the degree of linear correlation.

Example 5:



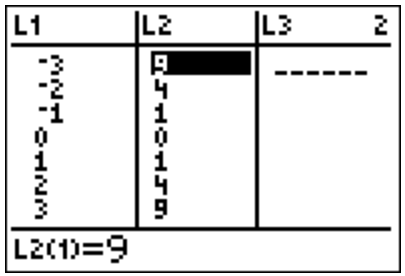


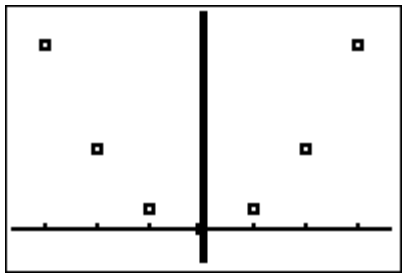


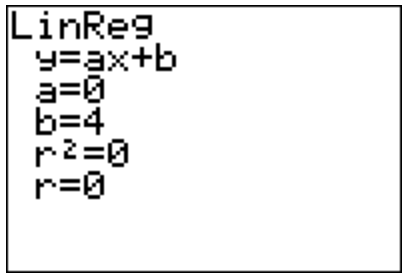


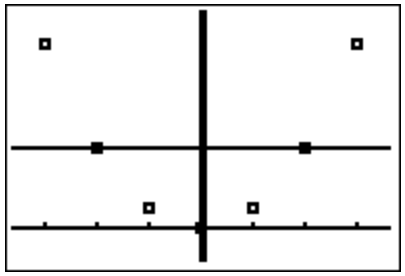
There is clearly a mathematical relationship between the variables, but it is not a linear one. Hence, $r = 0$.

Example 5:









There is clearly a mathematical relationship between the variables, but it is not a linear one. Hence, $r = 0$.

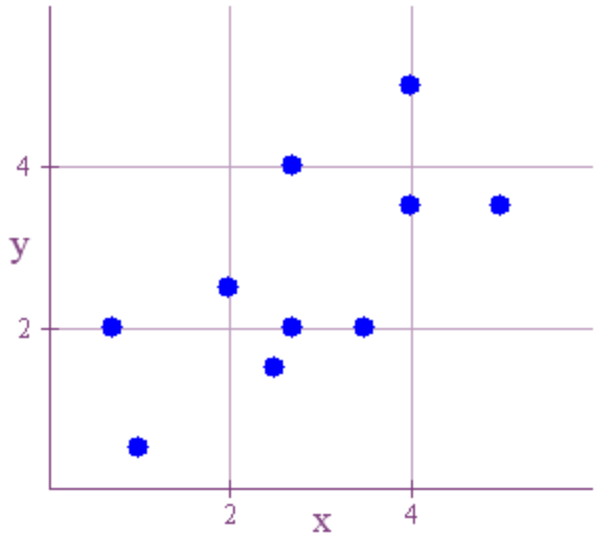
NOTE: If we are dealing with a population instead of a sample, then the linear correlation coefficient is denoted by *rho* = ρ.

How do we compute the coefficient of linear correlation? We basically convert each *x* and *y* coordinate to a *z* score, multiply them together, add them up, and then divide by the number of ordered pairs in order to adjust the size of our sum. Of course, when we are dealing with sample data, we divide by *n-1* instead of *n*.

$$r = \frac{\sum z_x z_y}{n-1}$$

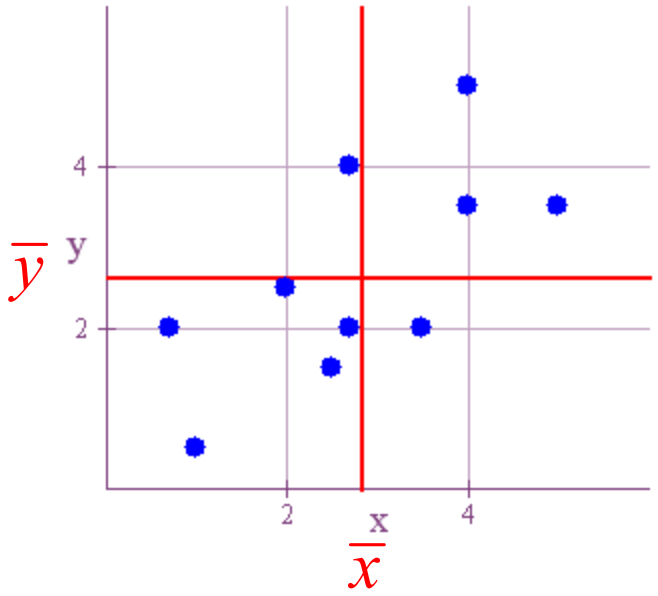
Here's how it works in practice. The scatterplot below clearly represents a positive linear correlation. Also, recall the formula for converting a raw score to a *z* score.

$$z = \frac{x - \mu}{\sigma}$$



If we draw some lines to represent the mean of the *x* coordinates and the mean of the *y* coordinates, then most of our coordinates are either both above their means or both below their respective means.

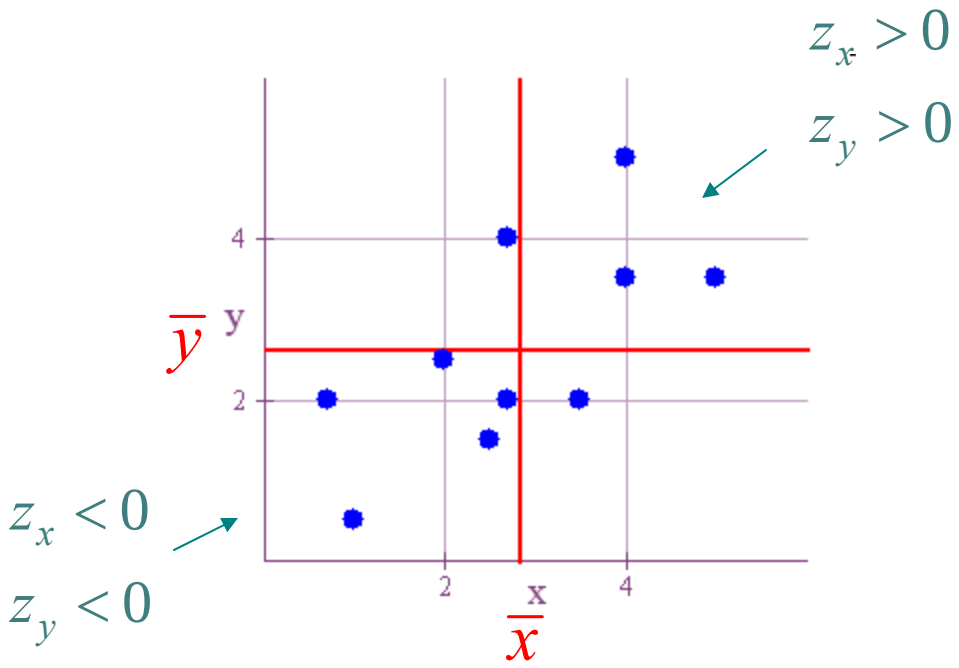
$$z = \frac{x - \mu}{\sigma}$$



This means that most of the products $z_x z_y$ will be positive, and, hence, r will be positive.

$$r = \frac{\sum z_x z_y}{n-1} > 0$$

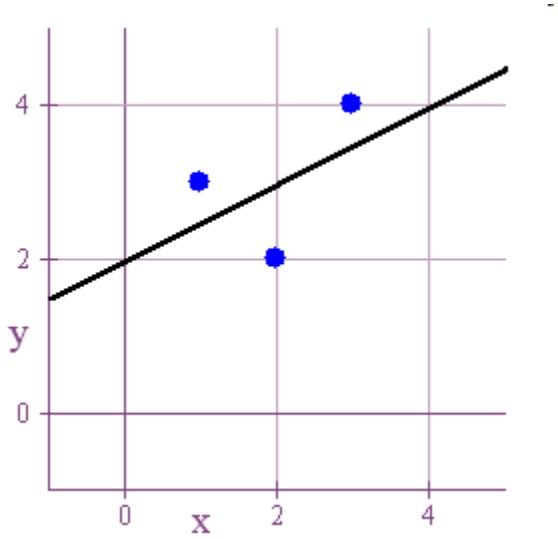
$$z = \frac{x - \mu}{\sigma}$$



Let’s now go back to our first example.

```
LinReg
 y=ax+b
 a=.5
 b=2
 r²=.25
 r=.5
```

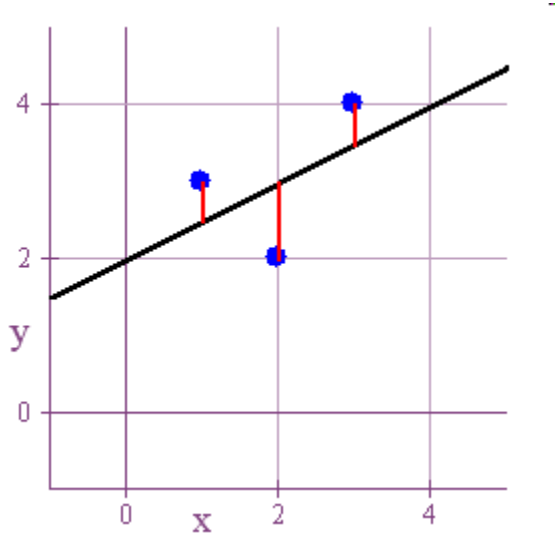
$(1,3),\ (2,2),\ (3,4)$



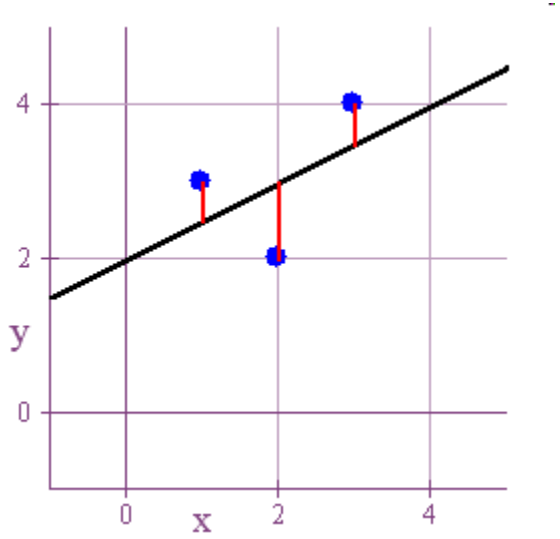
The line which best fits out data is called the *least squares regression line*.

```
LinReg
 y=ax+b
 a=.5
 b=2
 r²=.25
 r=.5
```

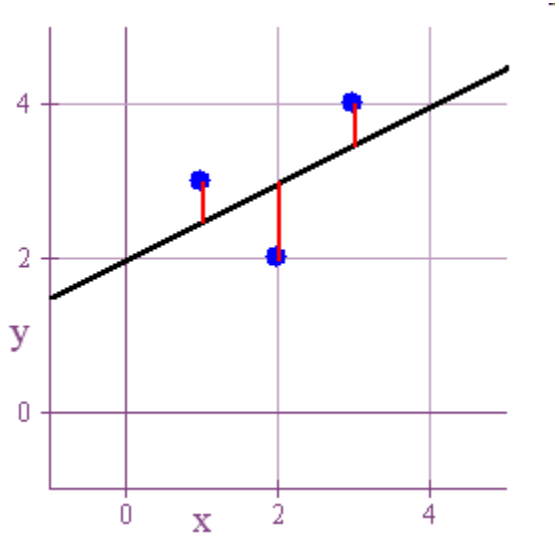
$(1,3),\ (2,2),\ (3,4)$



What this means is that if you look at the distance from the line to each data point, square that distance, and add everything up, then the sum of the squared distances will be minimized if we are using the *least squares regression line.*



Another way to look at it is this. Think of our data points as nails on a board, and think of the red lines as rubber bands connecting the nails to a broom stick. Then eventually the broom stick will settle down to some equilibrium position. That final position is the *least squares regression line*.

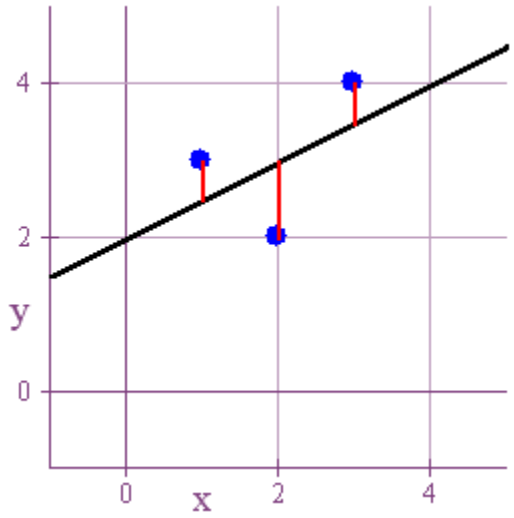


We often write the regression equation as *y-hat* to distinguish it from other *y-values.*

```
LinReg
 y=ax+b
 a=.5
 b=2
 r²=.25
 r=.5
```

$(1,3),\ (2,2),\ (3,4)$

$\hat{y} = .5x + 2$

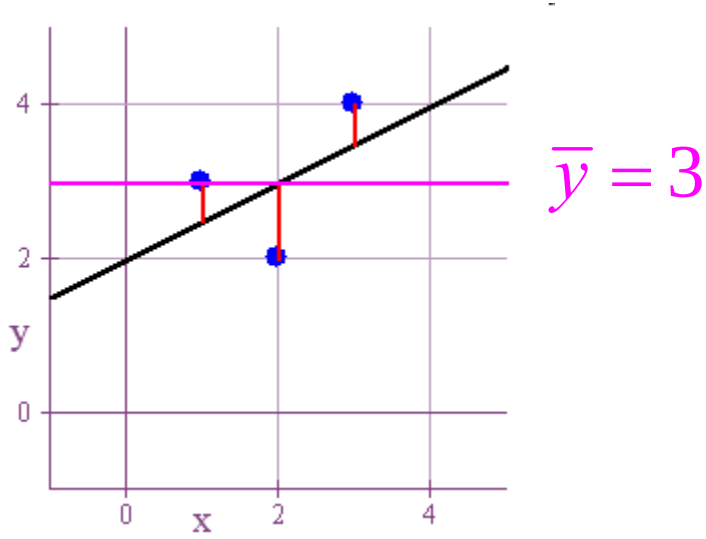


Now let's draw in a horizontal line at the position of the mean of the *y-values.*

```
LinReg
 y=ax+b
 a=.5
 b=2
 r²=.25
 r=.5
```

$(1,3),\ (2,2),\ (3,4)$

$\hat{y} = .5x + 2$

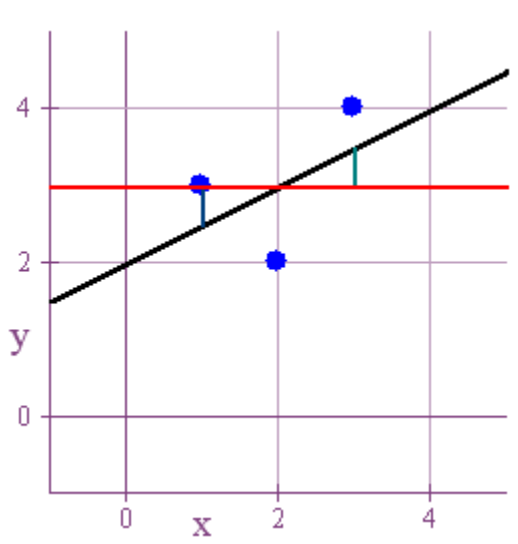


$\bar{y} = 3$

For each *x* value in our data points, the explained variance is the square of the distance between the line *y-bar* and the regression line.

$(1,3),\ (2,2),\ (3,4)$

$\hat{y} = .5x + 2$



$\bar{y} = 3$

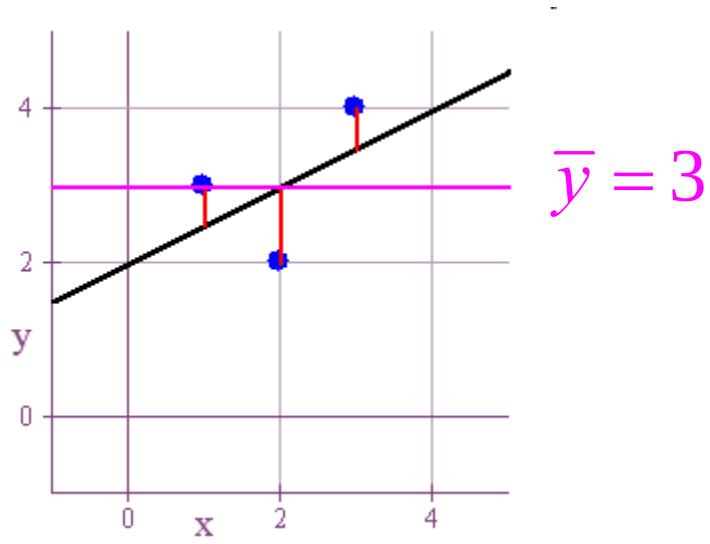
$$\text{Explained Variance} = \sum (\hat{y} - \bar{y})^2$$

$$= (2.5-3)^2 + (3-3)^2 + (3.5-3)^2 = 0.5$$

The unexplained variance is the square of the distance between the regression line and the *y* values of the data points.

$(1,3),\ (2,2),\ (3,4)$

$\hat{y} = .5x + 2$



$$\text{Explained Variance} = \sum (\hat{y} - \overline{y})^2$$

$$= (2.5-3)^2 + (3-3)^2 + (3.5-3)^2 = 0.5$$

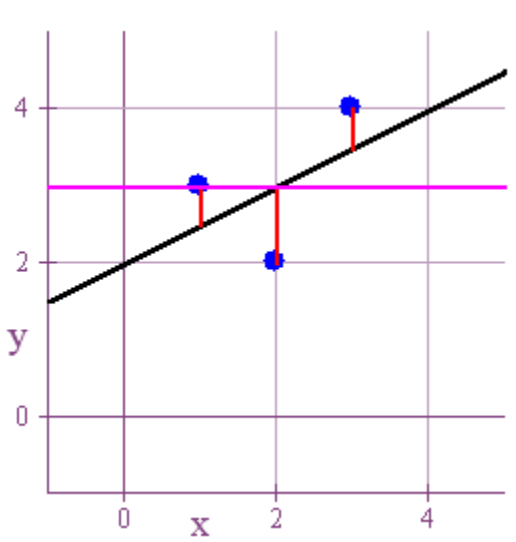
$$\text{Unexplained Variance} = \sum (y - \hat{y})^2$$

$$= (3-2.5)^2 + (2-3)^2 + (4-3.5)^2 = 1.5$$

The *coefficient of determination* is the ratio of *explained variance* to *total variance.*

$(1,3),\ (2,2),\ (3,4)$

$\hat{y} = .5x + 2$



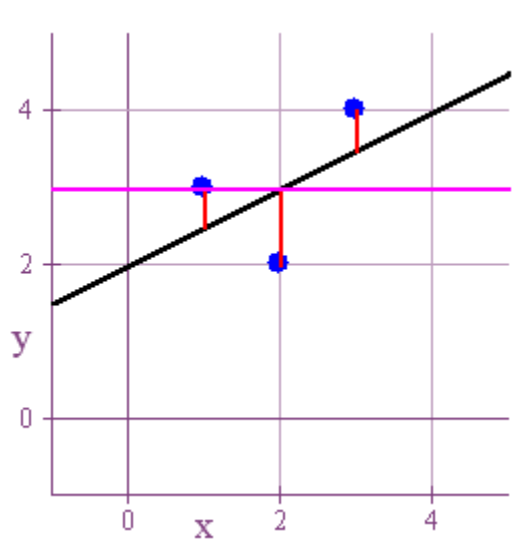
$\bar{y} = 3$

$$\text{Coefficient of Determination} = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{0.5}{0.5+1.5} = \frac{0.5}{2} = 0.25$$

The *coefficient of determination* is equal to r^2.

$(1,3),\ (2,2),\ (3,4)$

$\hat{y} = .5x + 2$



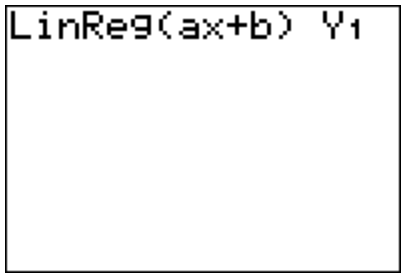
$\bar{y} = 3$

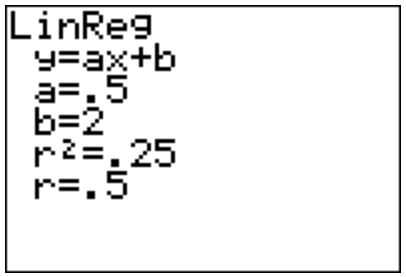
$$\text{Coefficient of Determination} = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{0.5}{0.5+1.5} = \frac{0.5}{2} = 0.25$$

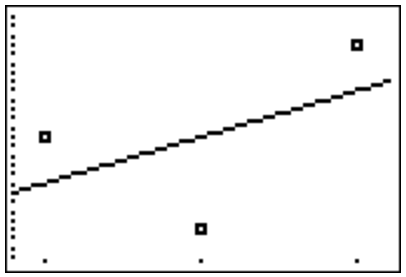
$$r^2 = 0.25$$

One of the more common uses of the linear regression equation is to make predictions.

$(1,3),\ (2,2),\ (3,4)$





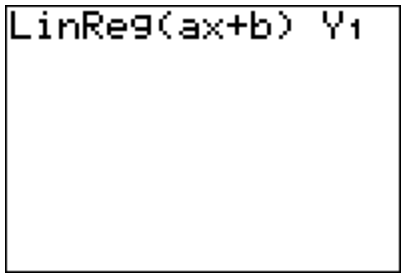


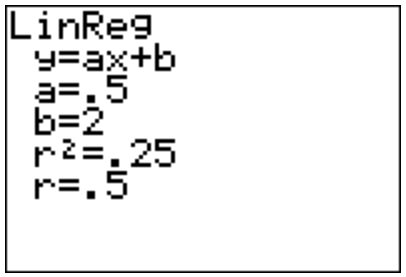
$VARS \rightarrow Y - VARS$

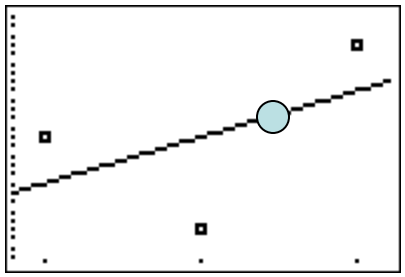
$\rightarrow FUNCTION \rightarrow Y_1$

If we use the equation to “fill in the gaps” within the range of our given *x* values, we call this *interpolation*.

$$(1,3),\ (2,2),\ (3,4)$$

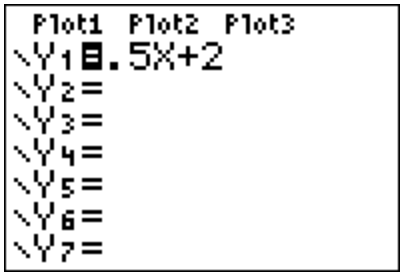


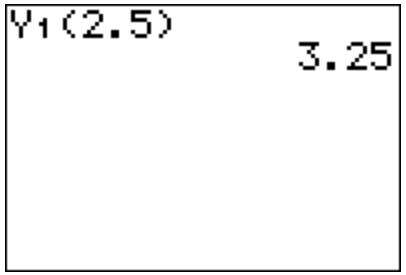




$VARS \rightarrow Y-VARS$

$\rightarrow FUNCTION \rightarrow Y_1$

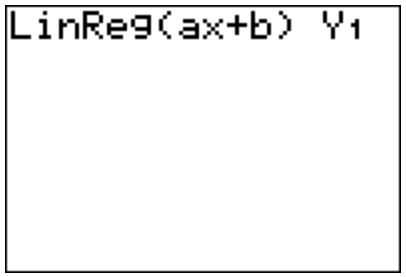


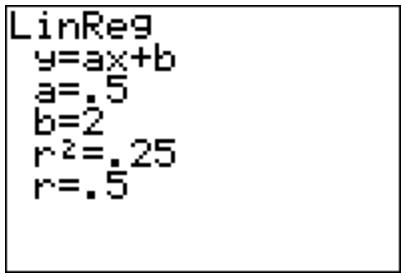


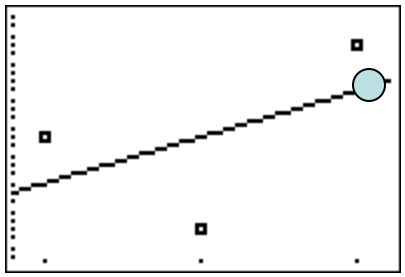
$$(2.5,3.25)$$

If we use the equation to predict values outside the range of our given *x* values, we call this *extrapolation*.

$$(1,3),\ (2,2),\ (3,4)$$

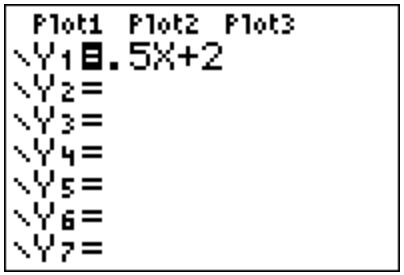


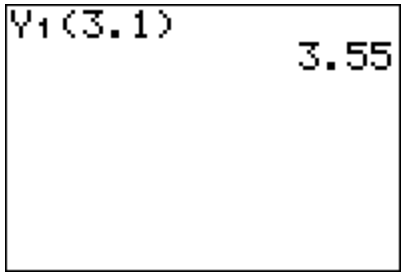




$VARS \rightarrow Y-VARS$

$\rightarrow FUNCTION \rightarrow Y_1$

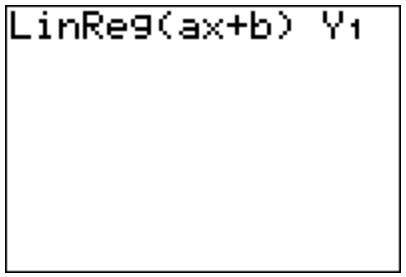


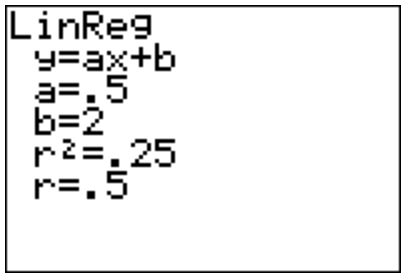


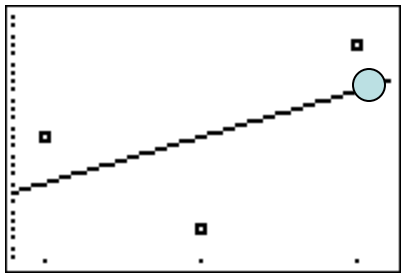
$$(3.1,3.55)$$

As a general rule, we should be careful about extrapolating too far into either the future or the past.

$$(1,3),\ (2,2),\ (3,4)$$

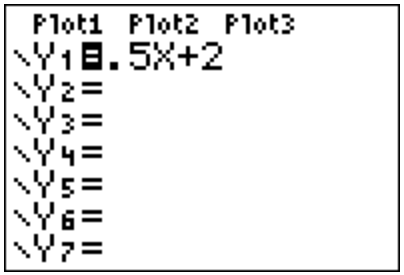


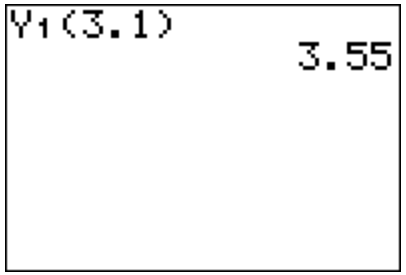




$VARS \rightarrow Y-VARS$

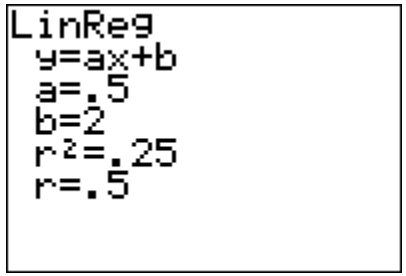
$\rightarrow FUNCTION \rightarrow Y_1$

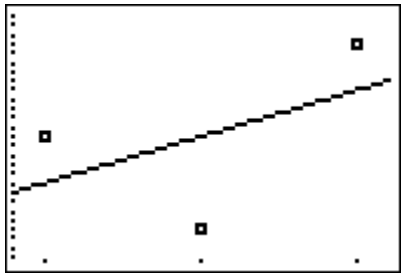




$$(3.1,3.55)$$

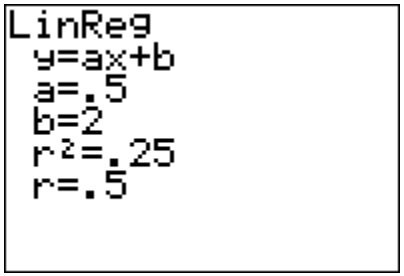
Finally, there's one more question we should address. Below we found a linear correlation coefficient of 0.5. How do we know that this is significantly different from zero?

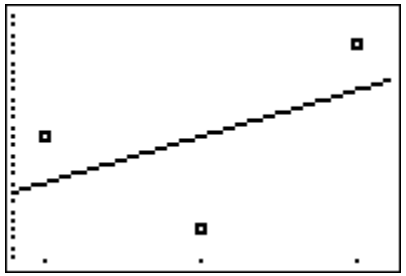




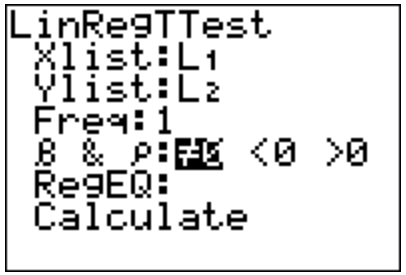
$(1,3),\ (2,2),\ (3,4)$

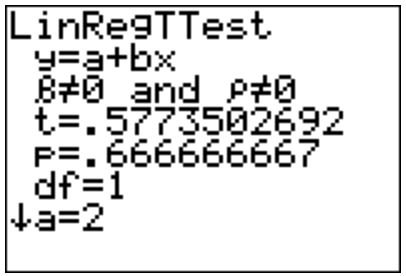
Fortunately, we have a test for that! Let's do it at the .05 level of significance. Our null hypothesis is that $\rho=0$.

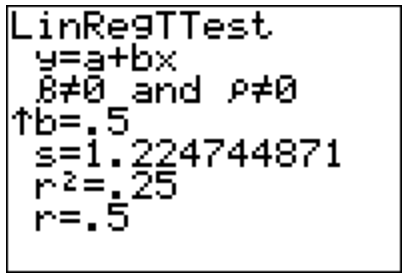




$(1,3),\ (2,2),\ (3,4)$







Also, here is a template to follow when doing a hypothesis test to see if a linear correlation is significantly different from zero.

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Level of Significance: $\alpha = .05$

P-Value: $P = 0.6667$

Decision: Accept H_0 (Fail to reject H_0)

Unfortunately, in this case we cannot reject the null hypothesis, and that also means we shouldn't use the regression equation for predictions. Instead, we use the mean of the *y* values as a point estimate for predictions.

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$(1,3),\ (2,2),\ (3,4)$

Level of Significance: $\alpha = .05$

P-Value: $P = 0.6667$

Decision: Accept H_0 (Fail to reject H_0)

$$\overline{y} = \frac{3+2+4}{3} = \frac{9}{3} = 3$$