Bivariate Statistics

When we have two variables, we are often interested in describing any relationships that may exist between them.

1. We use scatter diagrams or scatter plots as a visual aid to help determine if a relationship exists and what type of relationship there is.

2. This will help to determine equations which let us predict the value of one variable if we know the other variable. We will only be working with linear relationships, so our regression lines will be of the form y=mx + b

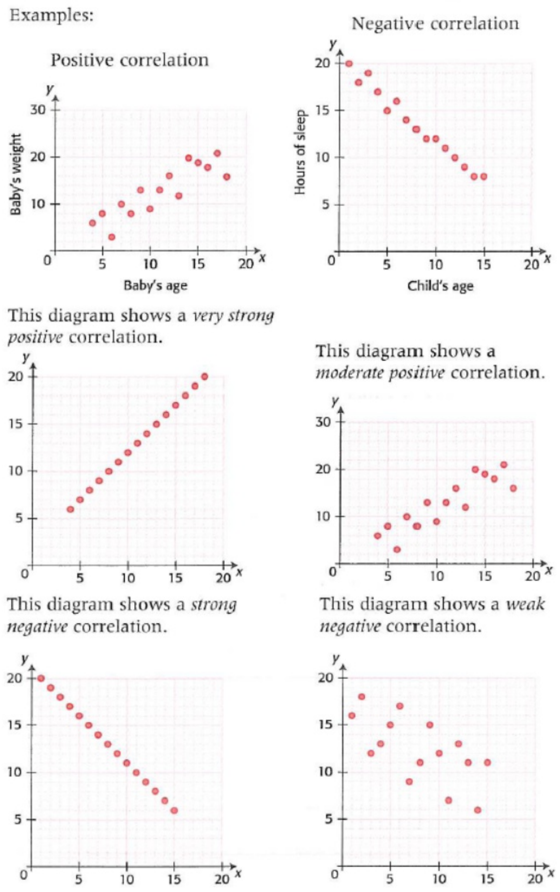
3. Then we will learn how to consider regression analysis to measure the strength of any relationship.

Correlation----

We describe the types of correlations in several ways...

For a general upward trend, we say the correlation is positive.
For a general downward trend, we say the correlation is negative.
For randomly scattered points, we say there is no correlation.

We also look at the pattern of points to make a judgement about the strength of the correlation. We classify correlations as strong, moderate or weak.



Examples:

Positive correlation

Negative correlation

No correlation

This diagram shows a *very strong positive* correlation.

This diagram shows a *moderate positive* correlation.

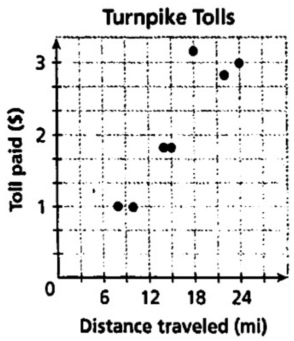
This diagram shows a *strong negative* correlation.

This diagram shows a *weak negative* correlation.

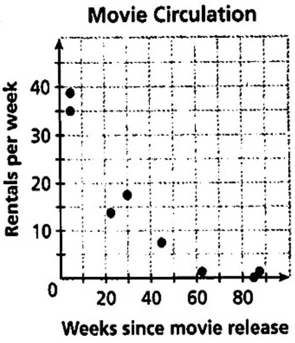
Practice/Homework:
Cirrito p. 570:1 to 5 all

Describe the correlation illustrated by each scatter plot.

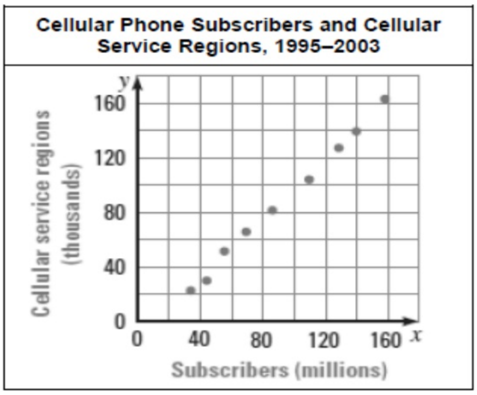
5.

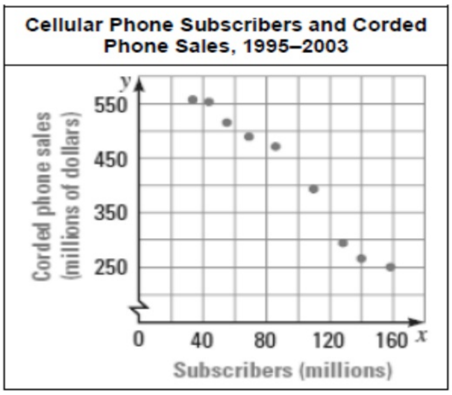

Turnpike Tolls

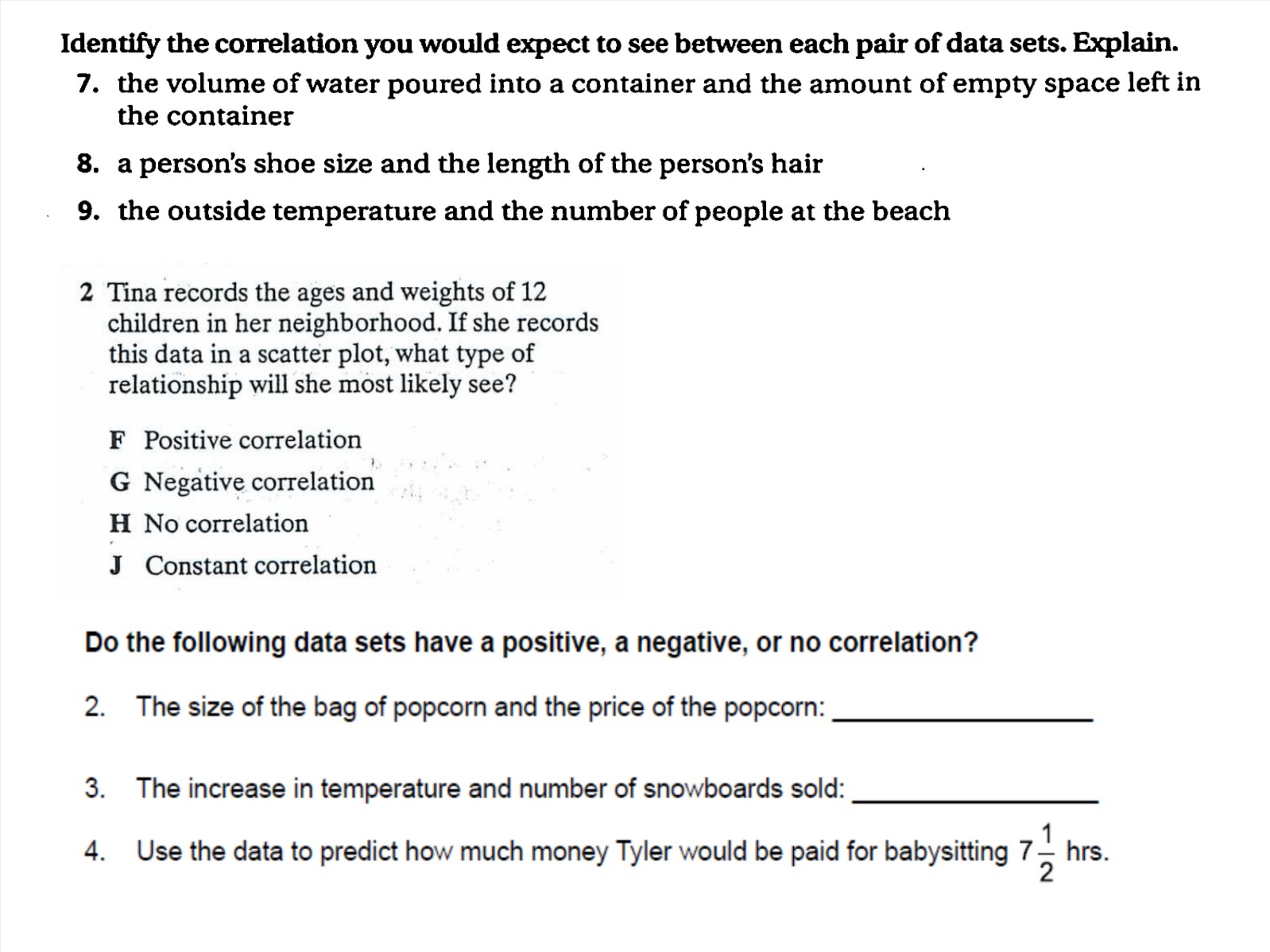
6.


Movie Circulation

TELEPHONES Describe the correlation shown by each scatter plot.


Cellular Phone Subscribers and Cellular Service Regions, 1995–2003


Cellular Phone Subscribers and Corded Phone Sales, 1995–2003



Identify the correlation you would expect to see between each pair of data sets. Explain.

7. the volume of water poured into a container and the amount of empty space left in the container

8. a person's shoe size and the length of the person's hair

9. the outside temperature and the number of people at the beach

2 Tina records the ages and weights of 12 children in her neighborhood. If she records this data in a scatter plot, what type of relationship will she most likely see?

F Positive correlation

G Negative correlation

H No correlation

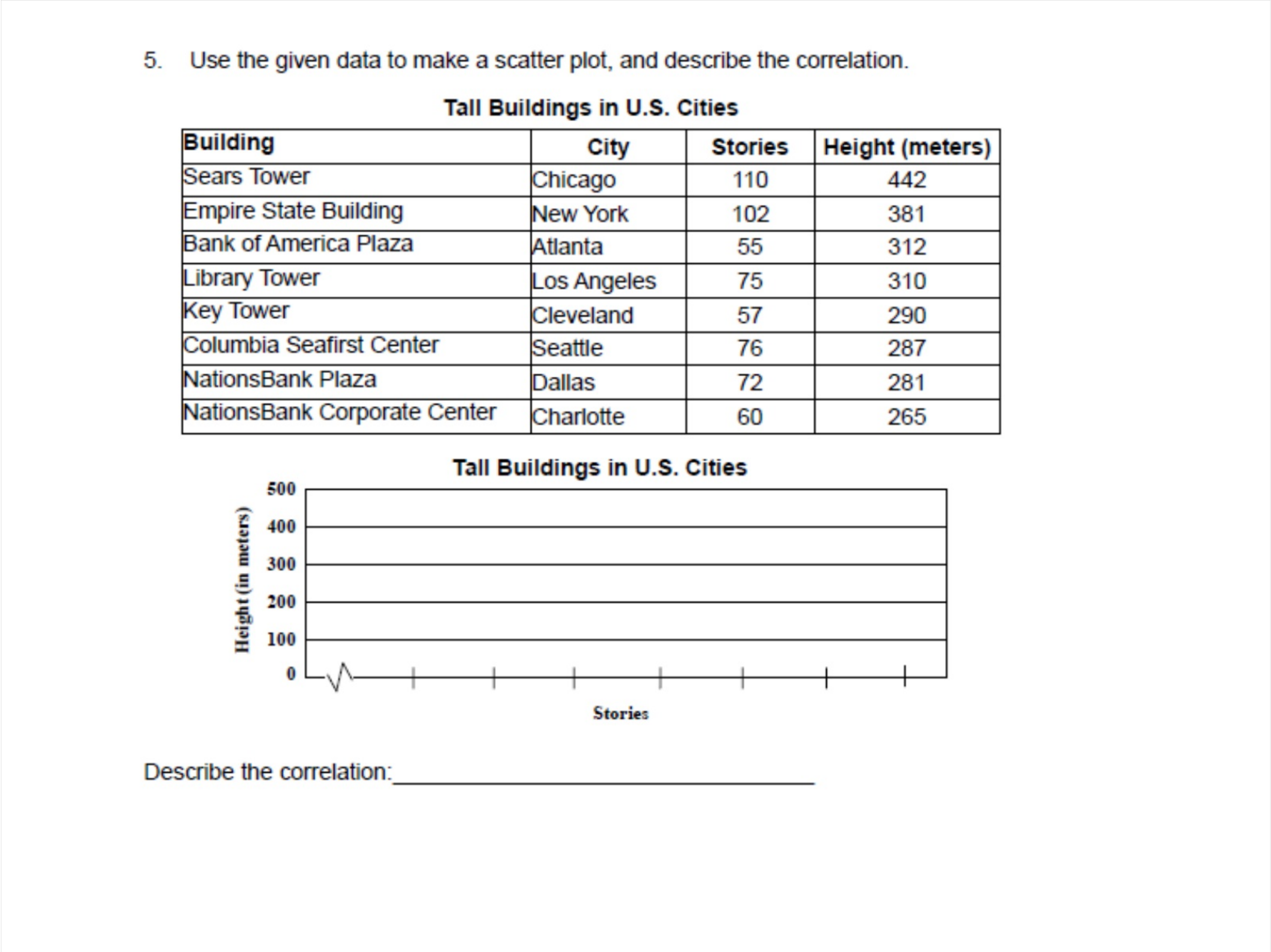
J Constant correlation

Do the following data sets have a positive, a negative, or no correlation?

2. The size of the bag of popcorn and the price of the popcorn: ____________________

3. The increase in temperature and number of snowboards sold: ____________________

4. Use the data to predict how much money Tyler would be paid for babysitting $7\frac{1}{2}$ hrs.



5. Use the given data to make a scatter plot, and describe the correlation.

Tall Buildings in U.S. Cities

Building	City	Stories	Height (meters)
Sears Tower	Chicago	110	442
Empire State Building	New York	102	381
Bank of America Plaza	Atlanta	55	312
Library Tower	Los Angeles	75	310
Key Tower	Cleveland	57	290
Columbia Seafirst Center	Seattle	76	287
NationsBank Plaza	Dallas	72	281
NationsBank Corporate Center	Charlotte	60	265

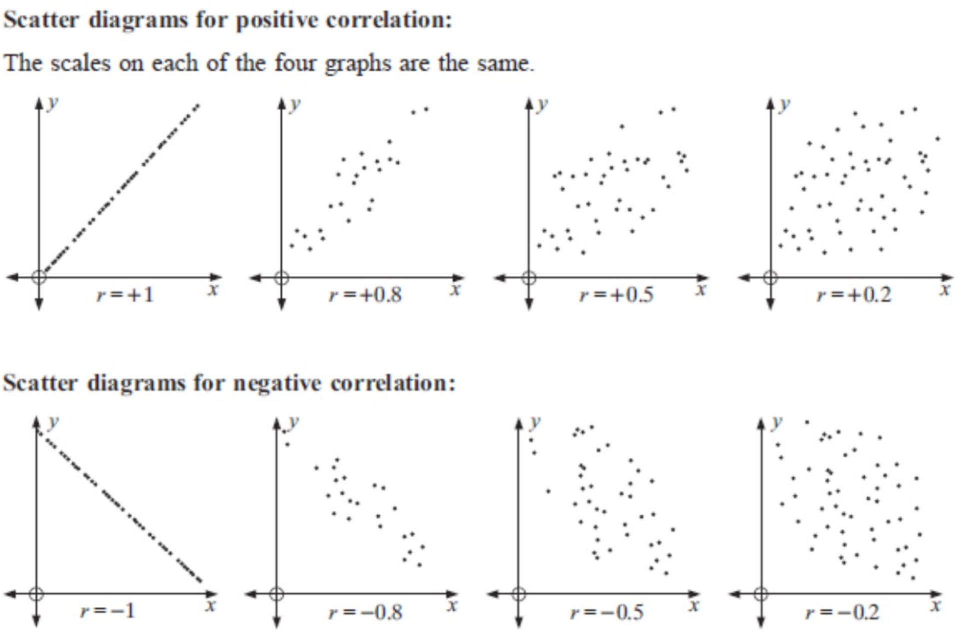
Tall Buildings in U.S. Cities

Describe the correlation:________________________________

Now let's look at how to **measure** the strength of a linear relationship.

We use something called Pearson's product-moment correlation coefficient (r) to find a numerical value that can be used to determine the strength of a linear correlation between two sets of data.

When dealing with a linear association, the correlation coefficient (r) is between 1 and -1.

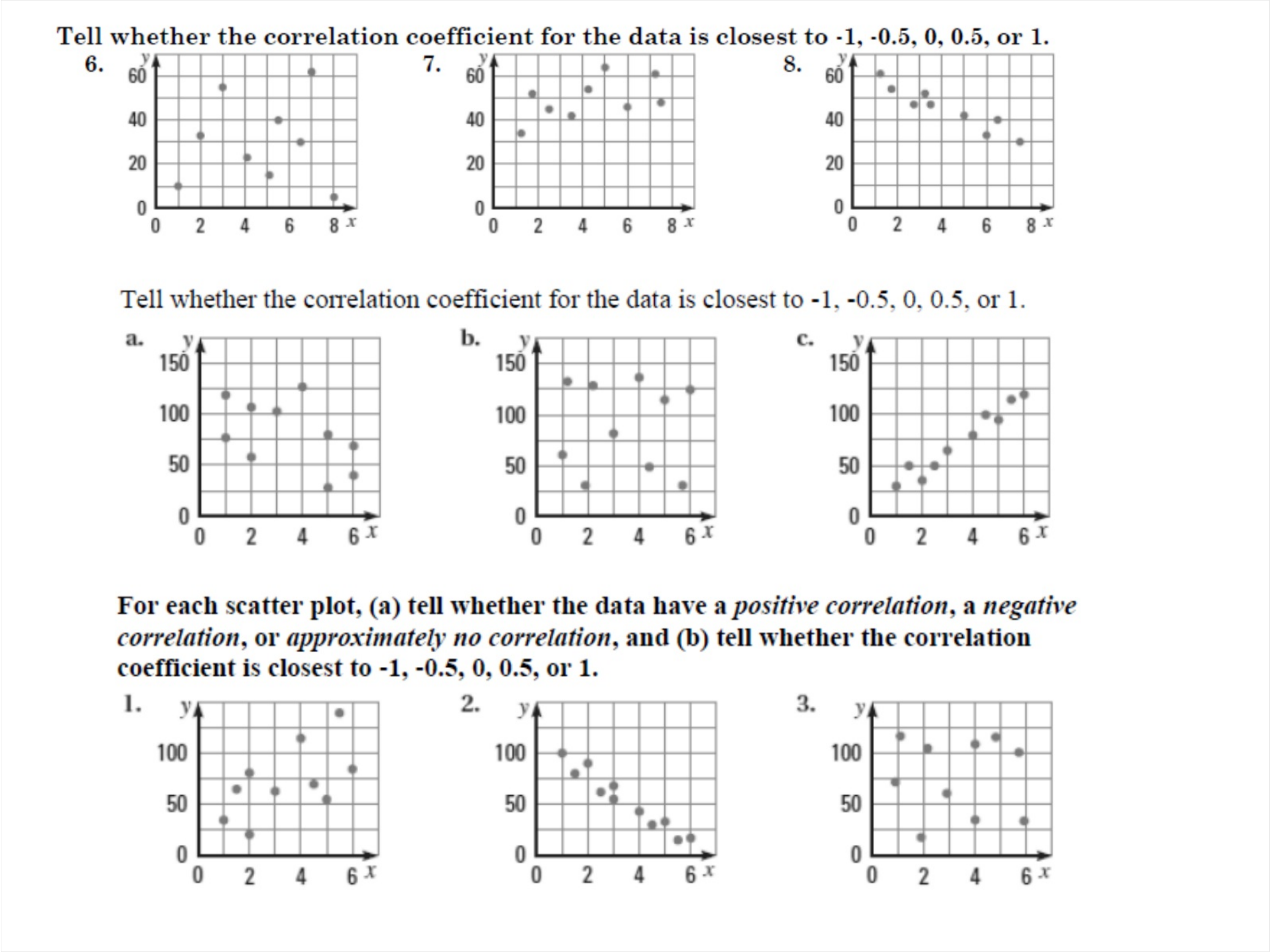


Scatter diagrams for positive correlation:

The scales on each of the four graphs are the same.

Scatter diagrams for negative correlation:

The closer r is to 1 or -1, the stronger the correlation.



Tell whether the correlation coefficient for the data is closest to -1, -0.5, 0, 0.5, or 1.

6.

7.

8.

Tell whether the correlation coefficient for the data is closest to -1, -0.5, 0, 0.5, or 1.

a.

b.

c.

For each scatter plot, (a) tell whether the data have a *positive correlation*, a *negative correlation*, or *approximately no correlation*, and (b) tell whether the correlation coefficient is closest to -1, -0.5, 0, 0.5, or 1.

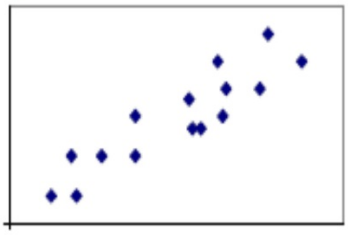
1.

2.

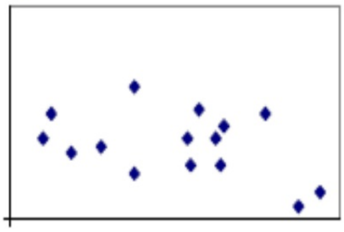
3.

6. The correlation coefficients for the six scatter plots shown below are -0.85, -0.40, 0, 0.50, 0.90 and 0.99. Match each scatter plot with the correct correlation coefficient.

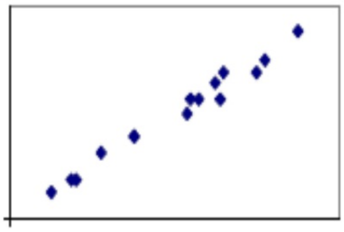
a.



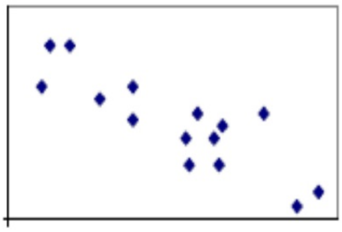
b.



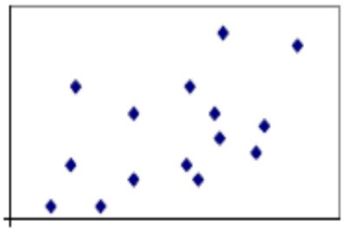
c.



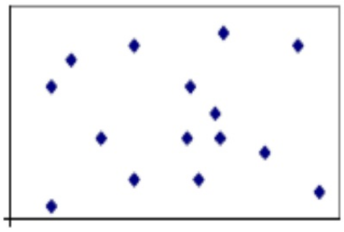
d.

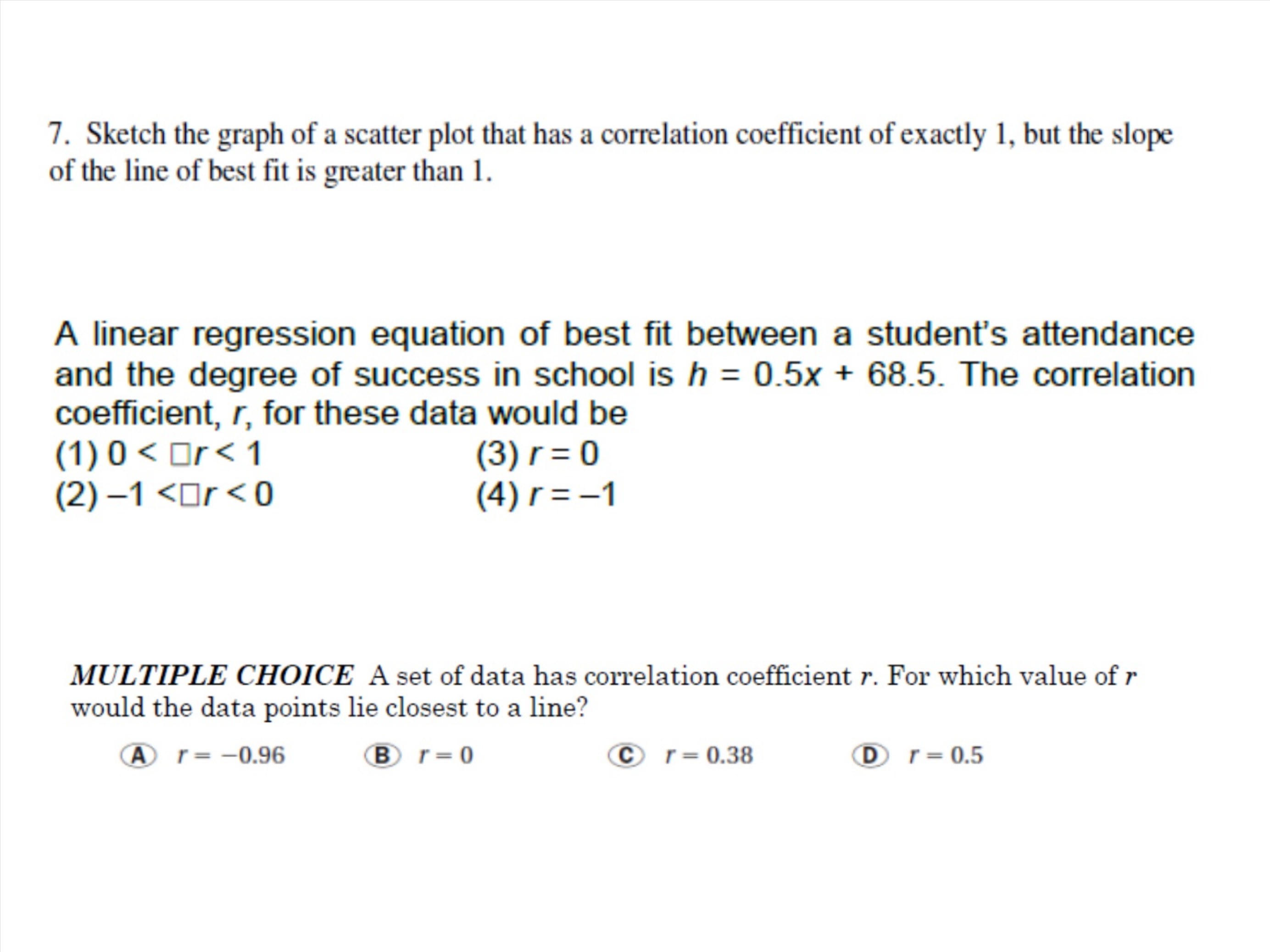


e.



f.





7. Sketch the graph of a scatter plot that has a correlation coefficient of exactly 1, but the slope of the line of best fit is greater than 1.

A linear regression equation of best fit between a student's attendance and the degree of success in school is $h = 0.5x + 68.5$. The correlation coefficient, r, for these data would be

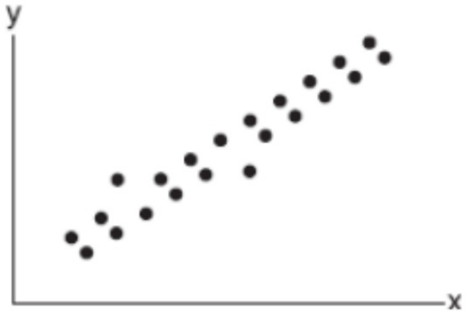
(1) $0 < \square r < 1$ (3) $r = 0$

(2) $-1 < \square r < 0$ (4) $r = -1$

MULTIPLE CHOICE A set of data has correlation coefficient r. For which value of r would the data points lie closest to a line?

Ⓐ $r = -0.96$ Ⓑ $r = 0$ Ⓒ $r = 0.38$ Ⓓ $r = 0.5$

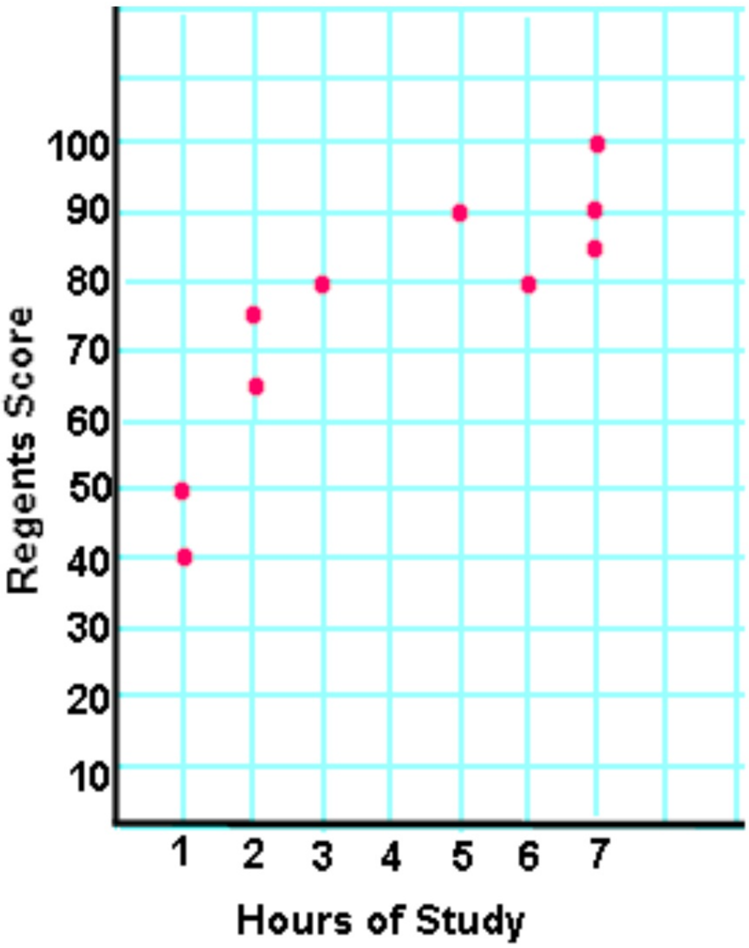
What could be the approximate value of the correlation coefficient for the accompanying scatter plot?



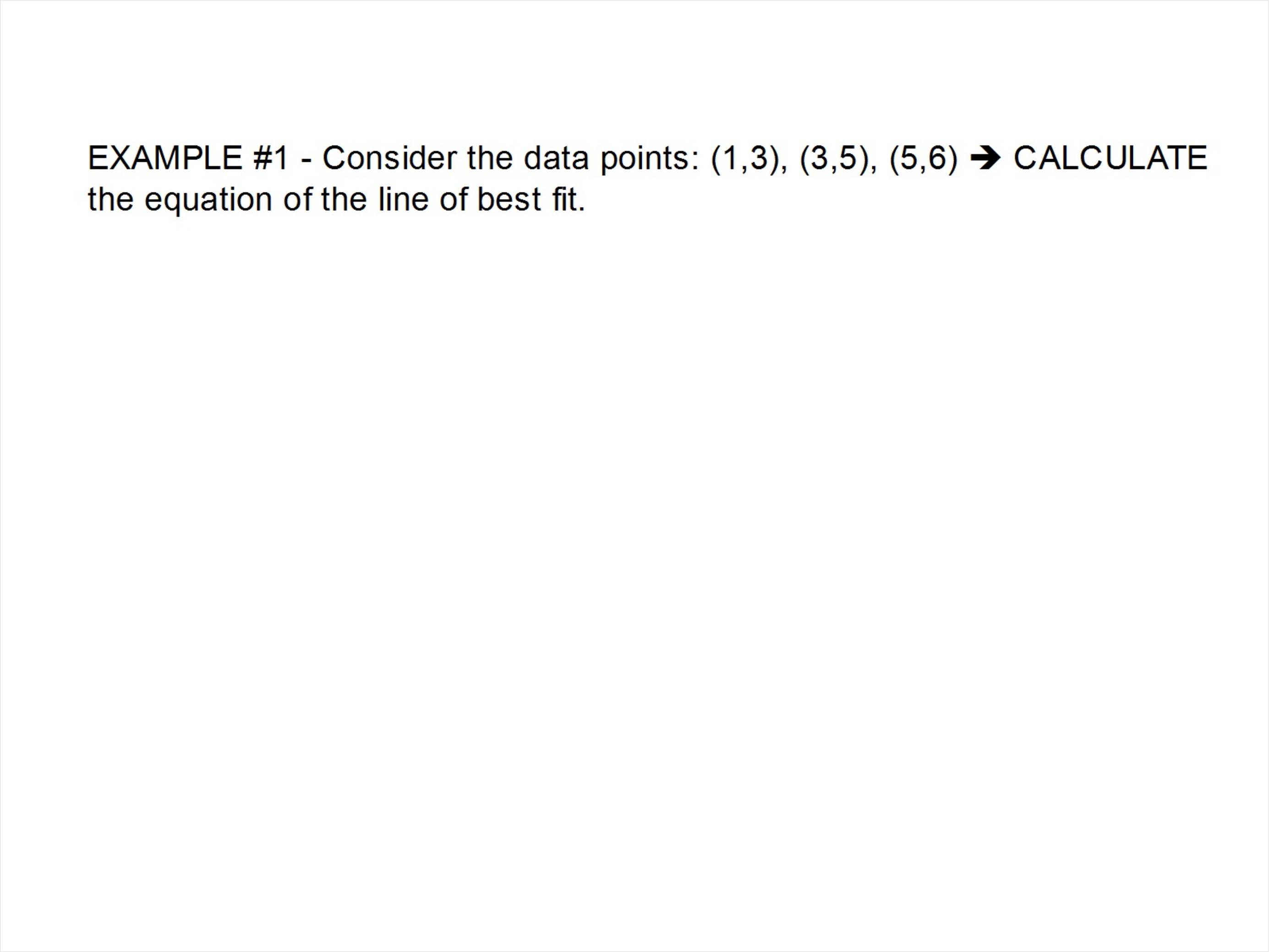
(1) -0.85
(2) -0.16

(3) 0.21
(4) 0.90

Study Hours	Test Score
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100

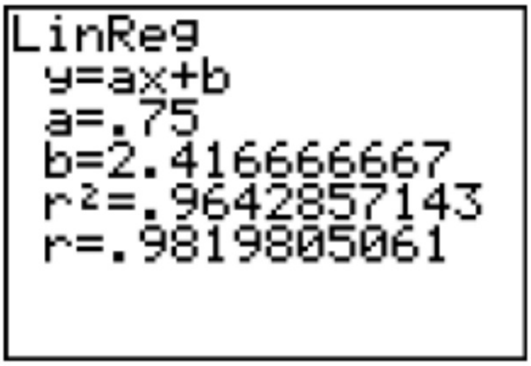


Lesson 67 - Pearson's Product-Moment Correlation Coefficient

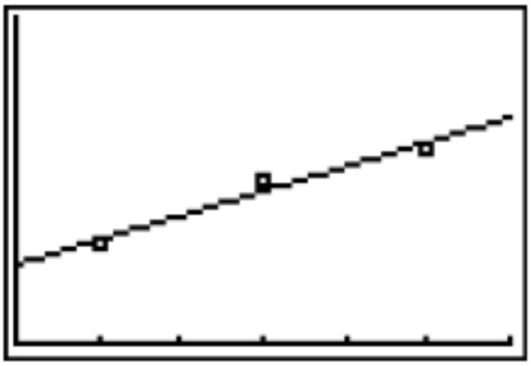


EXAMPLE #1 - Consider the data points: (1,3), (3,5), (5,6) ➔ CALCULATE the equation of the line of best fit.

EXAMPLE #1 - Consider the data points: (1,3), (3,5), (5,6) ➔
CALCULATE the equation of the line of best fit.

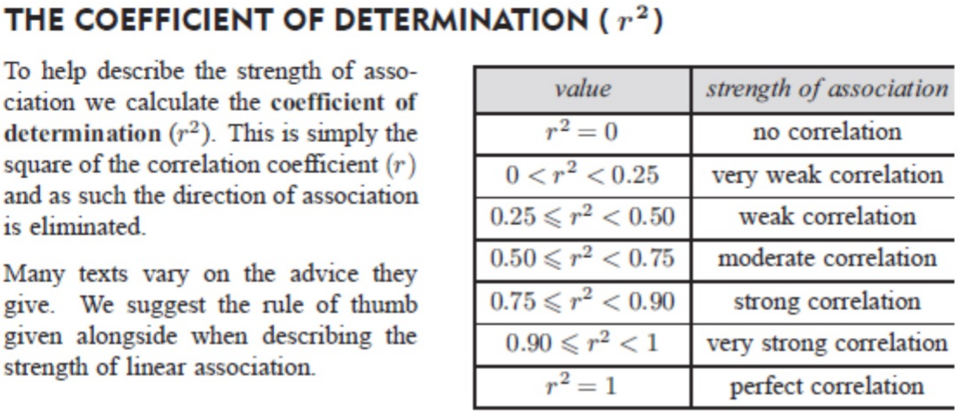


```
LinReg
 y=ax+b
 a=.75
 b=2.416666667
 r²=.9642857143
 r=.9819805061
```



The coefficient of determination (r^2)

What Haese and Harris says:



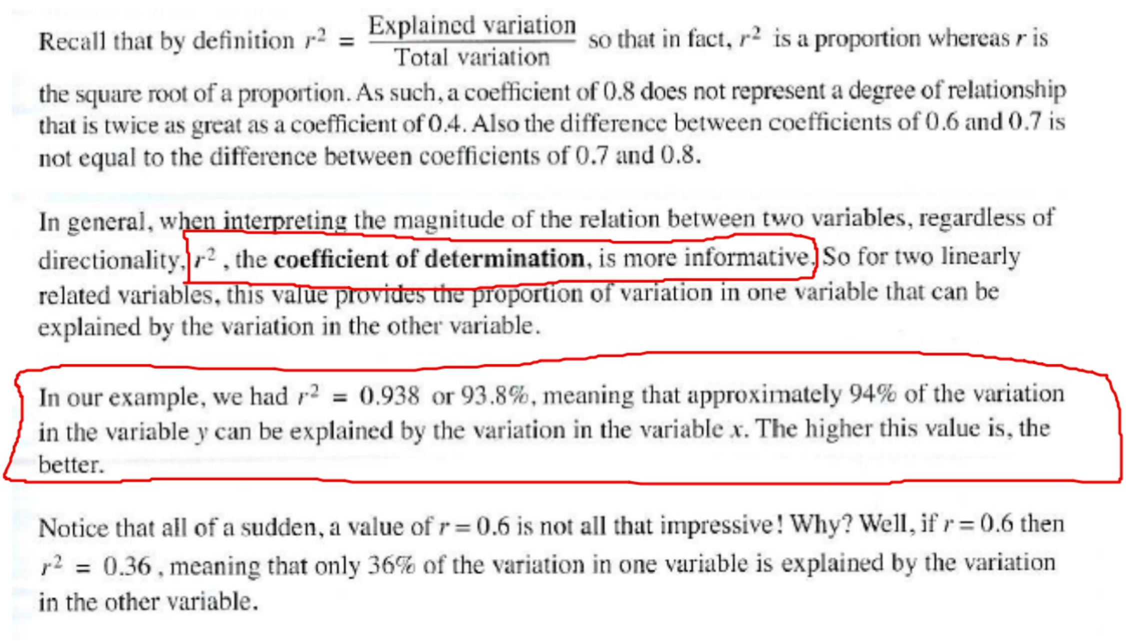
THE COEFFICIENT OF DETERMINATION (r^2)

To help describe the strength of association we calculate the **coefficient of determination** (r^2). This is simply the square of the correlation coefficient (r) and as such the direction of association is eliminated.

Many texts vary on the advice they give. We suggest the rule of thumb given alongside when describing the strength of linear association.

value	*strength of association*
$r^2 = 0$	no correlation
$0 < r^2 < 0.25$	very weak correlation
$0.25 \leqslant r^2 < 0.50$	weak correlation
$0.50 \leqslant r^2 < 0.75$	moderate correlation
$0.75 \leqslant r^2 < 0.90$	strong correlation
$0.90 \leqslant r^2 < 1$	very strong correlation
$r^2 = 1$	perfect correlation

What Cirrito says:



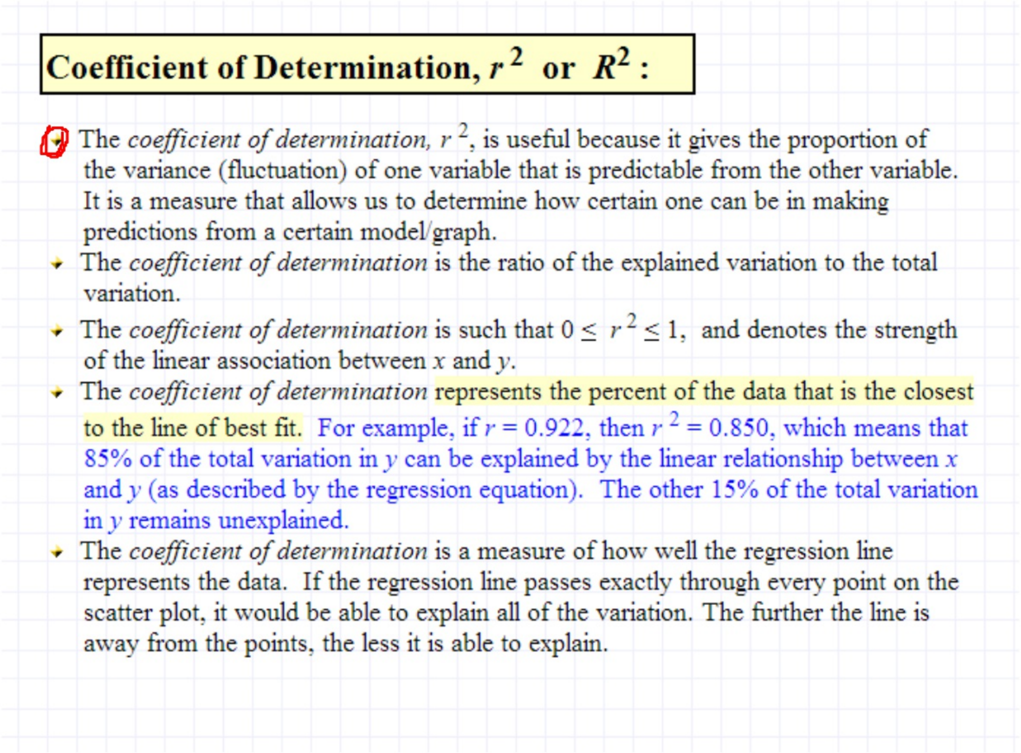
Recall that by definition $r^2 = \dfrac{\text{Explained variation}}{\text{Total variation}}$ so that in fact, r^2 is a proportion whereas r is the square root of a proportion. As such, a coefficient of 0.8 does not represent a degree of relationship that is twice as great as a coefficient of 0.4. Also the difference between coefficients of 0.6 and 0.7 is not equal to the difference between coefficients of 0.7 and 0.8.

In general, when interpreting the magnitude of the relation between two variables, regardless of directionality, r^2, the **coefficient of determination**, is more informative. So for two linearly related variables, this value provides the proportion of variation in one variable that can be explained by the variation in the other variable.

In our example, we had $r^2 = 0.938$ or 93.8%, meaning that approximately 94% of the variation in the variable y can be explained by the variation in the variable x. The higher this value is, the better.

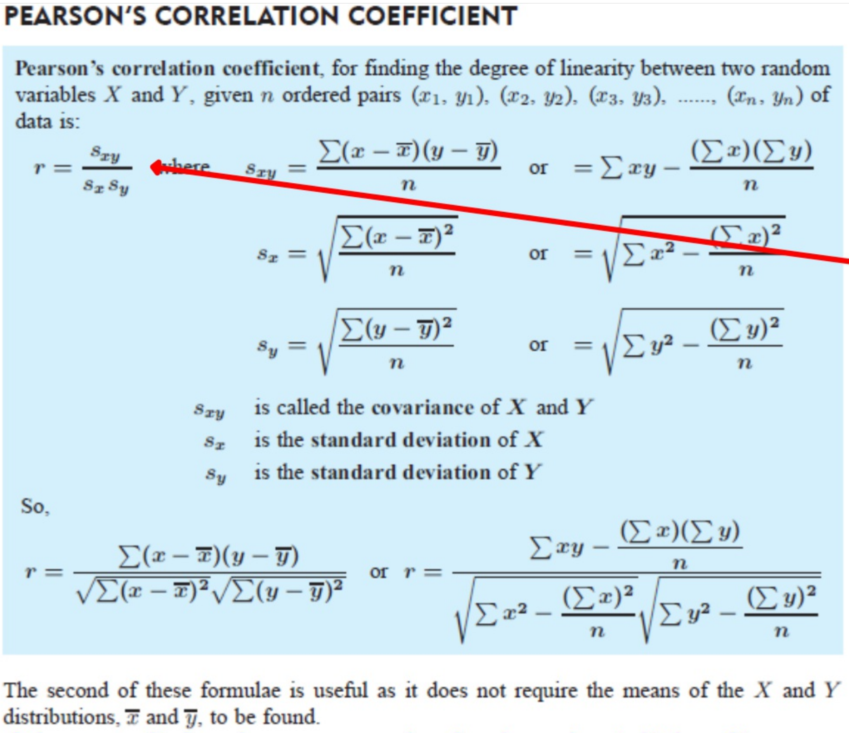
Notice that all of a sudden, a value of $r = 0.6$ is not all that impressive! Why? Well, if $r = 0.6$ then $r^2 = 0.36$, meaning that only 36% of the variation in one variable is explained by the variation in the other variable.

From MathBits:



Coefficient of Determination, r^2 or R^2 :

- The *coefficient of determination,* r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.
- The *coefficient of determination* is the ratio of the explained variation to the total variation.
- The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between *x* and *y*.
- The *coefficient of determination* represents the percent of the data that is the closest to the line of best fit. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in *y* can be explained by the linear relationship between *x* and *y* (as described by the regression equation). The other 15% of the total variation in *y* remains unexplained.
- The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.



PEARSON'S CORRELATION COEFFICIENT

Pearson's correlation coefficient, for finding the degree of linearity between two random variables X and Y, given n ordered pairs (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n) of data is:

$$r = \frac{s_{xy}}{s_x s_y} \quad \text{where} \quad s_{xy} = \frac{\sum(x-\overline{x})(y-\overline{y})}{n} \quad \text{or} \quad = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$s_x = \sqrt{\frac{\sum(x-\overline{x})^2}{n}} \quad \text{or} \quad = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$s_y = \sqrt{\frac{\sum(y-\overline{y})^2}{n}} \quad \text{or} \quad = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$$

s_{xy} **is called the covariance of** X **and** Y

s_x **is the standard deviation of** X

s_y **is the standard deviation of** Y

So,

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2}\sqrt{\sum(y-\overline{y})^2}} \quad \text{or} \quad r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

The second of these formulae is useful as it does not require the means of the X and Y distributions, $\overline{x}$ and $\overline{y}$, to be found.

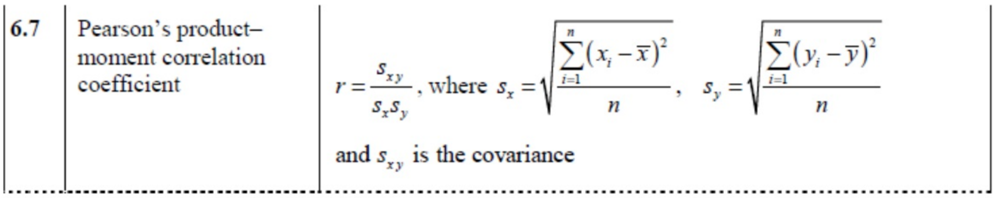
How to calculate r.

Note: You will use this formula in exams.

I know it looks overwhelming but it isn't.

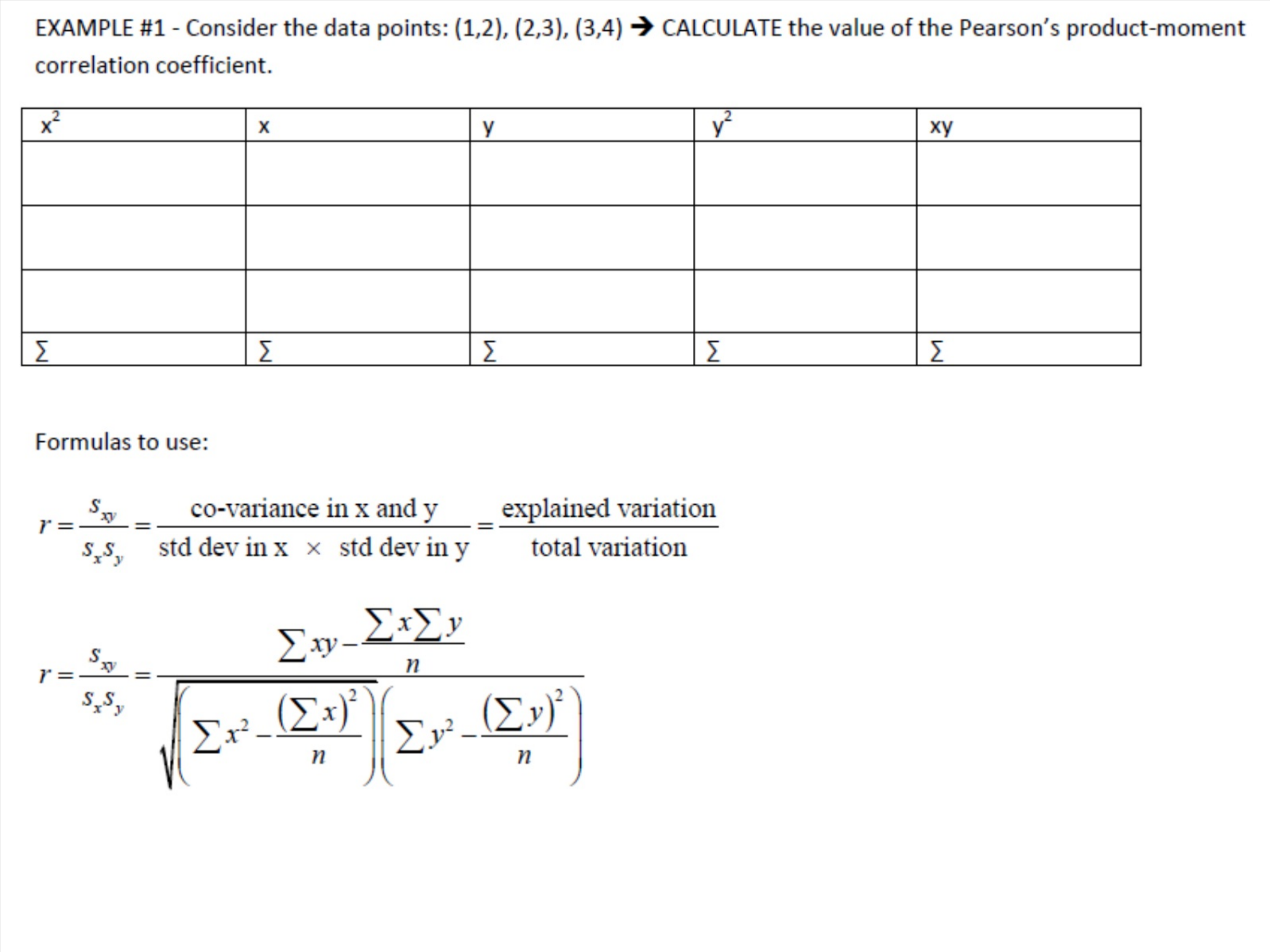
1. Your calculator can find it.
2. Even if you have to calculate it by hand, you can get most of the information from your calculator.
3. In exams, they will always give you S_{xy} and S_x and S_y are standard deviations from your calculator.

Here is the formula as written in your formula packet.



6.7	Pearson's product–moment correlation coefficient	$r = \frac{s_{xy}}{s_x s_y}$, where $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$, $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$ and s_{xy} is the covariance

Calculator notes:
1. Turn Diagnostics On.....
2. Show the various features of the two-variable stats display



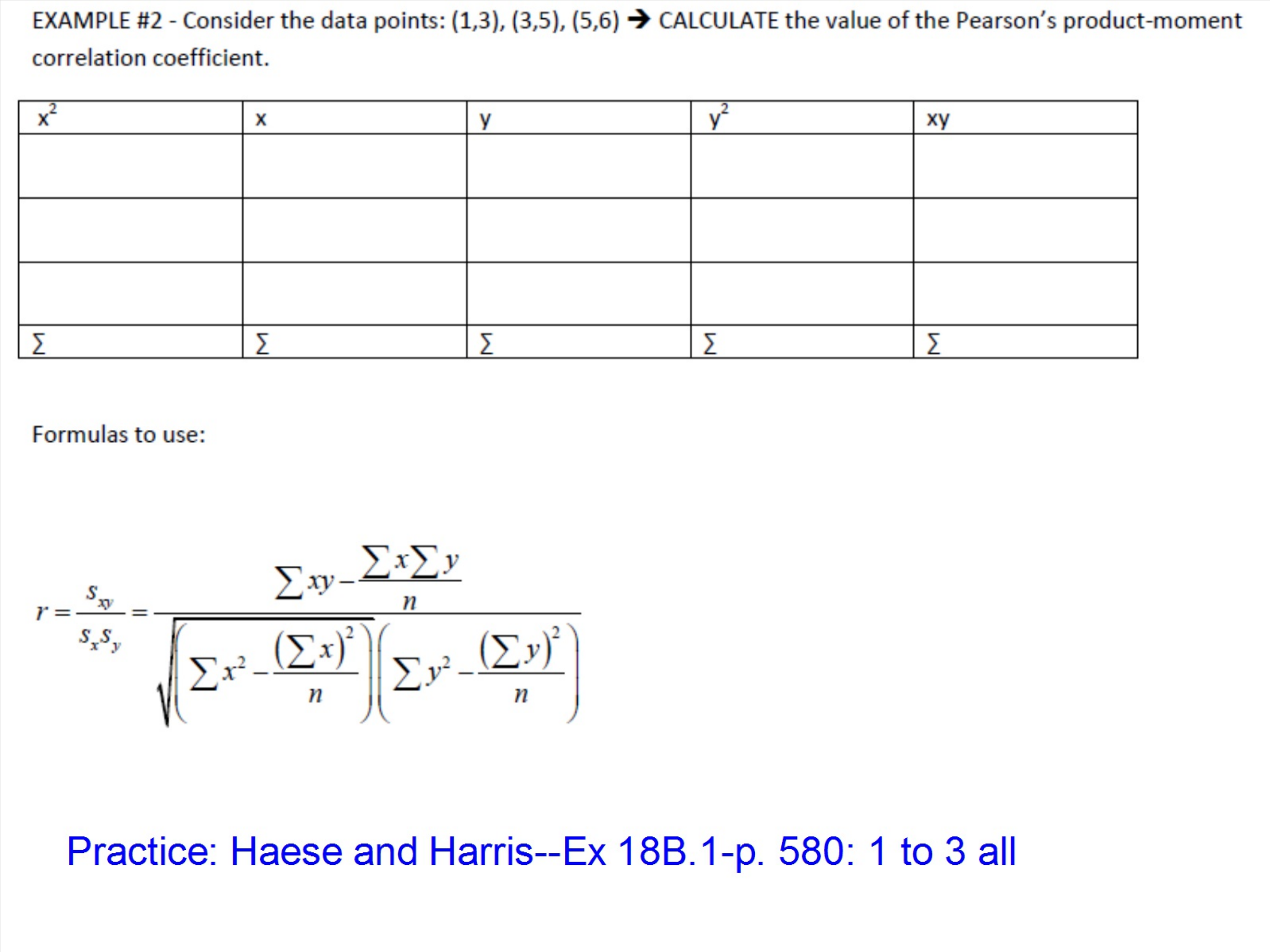
EXAMPLE #1 - Consider the data points: (1,2), (2,3), (3,4) ➔ CALCULATE the value of the Pearson's product-moment correlation coefficient.

x^2	x	y	y^2	xy
∑	∑	∑	∑	∑

Formulas to use:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\text{co-variance in x and y}}{\text{std dev in x} \times \text{std dev in y}} = \frac{\text{explained variation}}{\text{total variation}}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}$$



EXAMPLE #2 - Consider the data points: (1,3), (3,5), (5,6) ➔ CALCULATE the value of the Pearson's product-moment correlation coefficient.

x^2	x	y	y^2	xy
∑	∑	∑	∑	∑

Formulas to use:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}$$

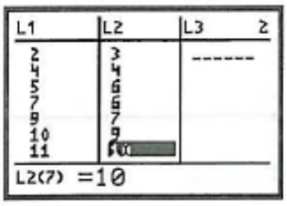
Practice: Haese and Harris--Ex 18B.1-p. 580: 1 to 3 all

Using technology to find r.
This is explained well in Cirrito: p. 576 and 577 or here:

We now make use of the TI–83:

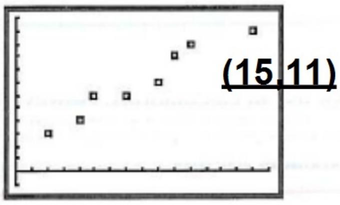
1. Enter data as two lists, $x \leftrightarrow L_1$ and $y \leftrightarrow L_2$.
2. Check to see if in fact there is a linear relationship.
3. Press, **STAT** then **CALC** and then choose **2:2–Var Stats** and enter L_1, L_2 .

1.

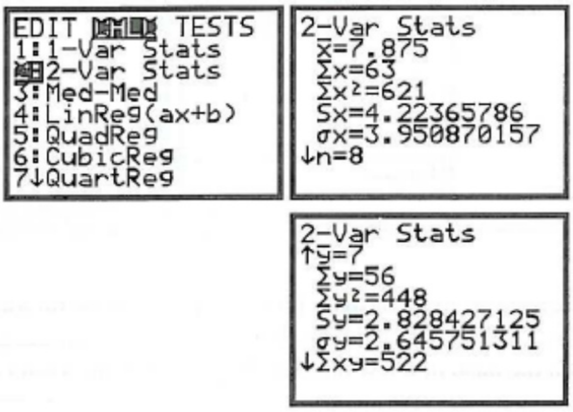


```
L1      L2      L3     2
2       3       -------
4       4
5       6
7       6
9       7
10      9
11      10
L2(7) =10
```

2.



3.



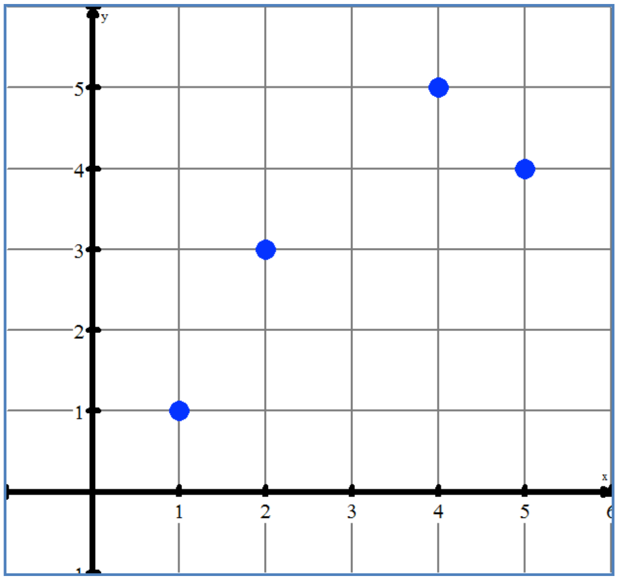
```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```

```
2-Var Stats
 x̄=7.875
 Σx=63
 Σx²=621
 Sx=4.22365786
 σx=3.950870157
↓n=8
```

```
2-Var Stats
↑ȳ=7
 Σy=56
 Σy²=448
 Sy=2.828427125
 σy=2.645751311
↓Σxy=522
```

Now these results can be used to calculate r.

(A) **EXPLORATION #1** → Data Set #1 → (1,1); (2,3); (4,5); (5,4)

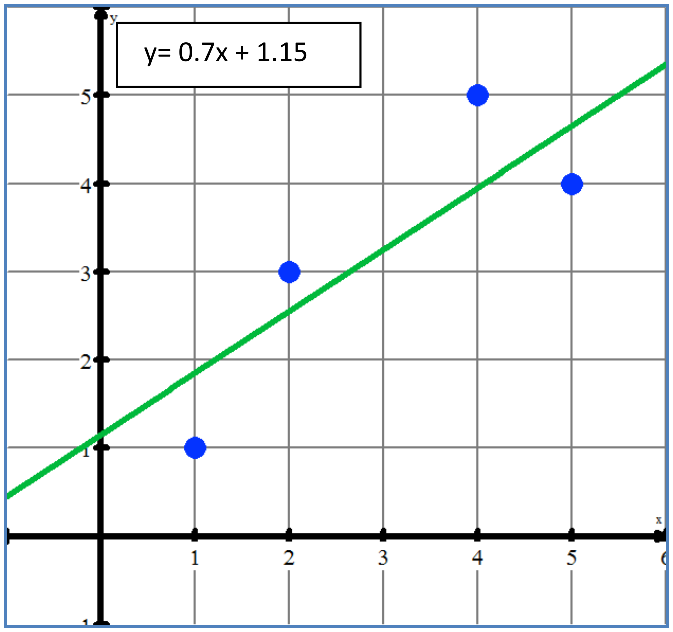


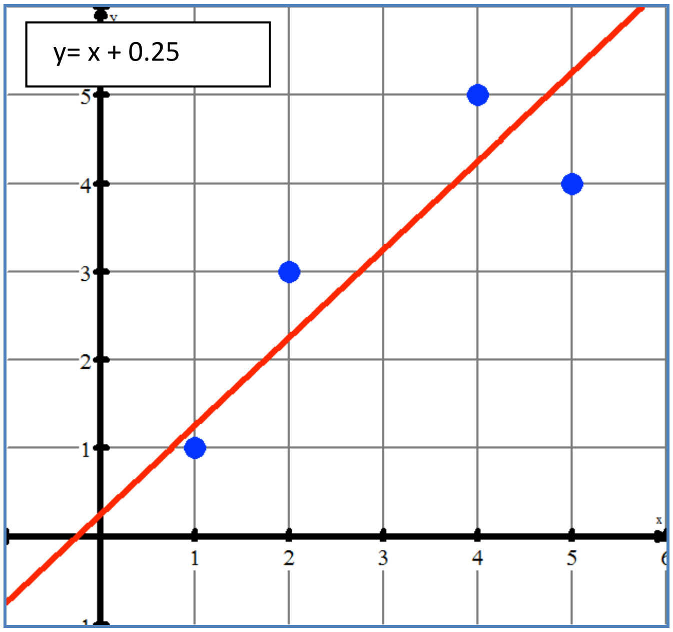
a. Calculate the mean point

b. Draw the line of best fit AS YOU PERCEIVE IT!!!

c. Determine the equation of the line you drew.

d. Student #1 drew the LoBF & decides that the line of best fit has an equation of y = 0.7x + 1.15. But Student #2 drew the LoBF & decides that the equation of the the line should be y = x + ¼. (see graph #2 & graph #3)





e. Q → Is there some way ***descriptive way*** we can determine whose line is a "better" fit for the data?

f. Q → Is there some ***analytical/numeric way*** we can determine whose line is a "better" fit for the data?

(B) Calculating “Residuals”

a. Working With Student #______. and the equation f(x) = ____________________________.

x	y	f(x) =	Residual = y – f(x)	Square of residual
1	1			
2	3			
4	5			
5	4			
$\bar{x}$ =	$\bar{y}$ =			$\sum$

b. Working With Student #______. and the equation f(x) = ____________________________.

x	y	f(x) =	Residual = y – f(x)	Square of residual
1	1			
2	3			
4	5			
5	4			
$\bar{x}$ =	$\bar{y}$ =			$\sum$

c. CONCLUSION → Which line “fits” better? Why?

d. Now let’s go to the following animation and “play” with a line of best fit on a data set: at http://hspm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html

Closing Question → How do I determine which line minimizes the squares of the residuals???

(C) Determining the Eqn of the Least Squares Regression Line & Regression Coefficient

Recall some of our formulas from Lesson 67:

$$s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = \frac{\sum xy}{n} - \bar{x}\bar{y} \quad \text{and} \quad s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} \quad \text{and}$$

$$s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}} = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} \quad \text{and} \quad r = \frac{s_{xy}}{s_x s_y}$$ as well as other formulas for *r* ➔

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}} = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}},$$

so let's apply these calculations here:

x^2	X	y	y^2	xy
	1	1		
	2	3		
	4	5		
	5	4		
$\sum$	$\sum$ so $\bar{x}$ =	$\sum$ so $\bar{y}$ =	$\sum$	$\sum$

So then s_x = ________, s_y = ________, s_{xy} = ________, and r = __________.

Then we add one more formula that will help us write an equation of the least square regression line ➔

$$y - \bar{y} = \frac{s_{xy}}{(s_x)^2}(x - \bar{x})$$ ➔ so our equation in our example should be ➔

Now test it out on the TI-84 ➔

(D) Further Examples → HH Textbook 18C, p589, Q3

Data Set:

Spray Concentration (mL/L)	3	5	6	8	9	11
Yield of Tomatoes (per bush)	67	90	103	120	124	150

x^2	x	y	y^2	xy
	3	67		
	5	90		
	6	103		
	8	120		
	9	124		
	11	150		
$\sum$	$\sum$ so $\bar{x}$ =	$\sum$ so $\bar{y}$ =	$\sum$	$\sum$

So then s_x = ________, s_y = ________, s_{xy} = ________, and r = __________.

the least square regression line → $y-\bar{y}=\frac{s_{xy}}{(s_x)^2}(x-\bar{x})$ → so our equation in our example should be →

(E) Further Examples → HH Textbook 18C, p589, Q5

Data Set:

Frost Free days	75	100	125	150	175	200
Rate of Reaction	44.6	42.1	39.4	37.0	34.1	31.2

x^2	x	y	y^2	xy
	75	44.6		
	100	42.1		
	125	39.4		
	150	37		
	175	34.1		
	200	31.2		
$\sum$	$\sum$ so $\bar{x}$ =	$\sum$ so $\bar{y}$ =	$\sum$	$\sum$

So then s_x = ________, s_y = ________, s_{xy} = ________, and r = __________.

the least square regression line → $y-\bar{y}=\frac{s_{xy}}{(s_x)^2}(x-\bar{x})$ → so our equation in our example should be →

Now test it out on the TI-84 →