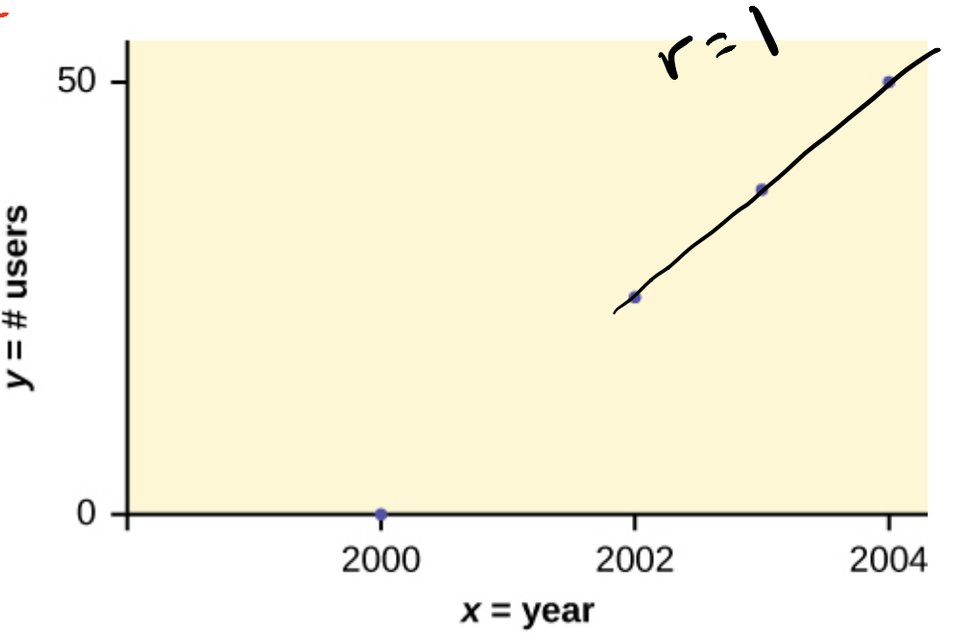
Example 12.5

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

Table 12.1

(a) Table showing the number of m-commerce users (in millions) by year.



(b) Scatter plot showing the number of m-commerce users (in millions) by year.

Figure 12.5

Using the TI-83, 83+, 84, 84+ Calculator

To create a scatter plot:

1. Enter your X data into list L1 and your Y data into list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Try It

12.5 Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	*Y* (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Table 12.2

Construct a scatter plot and state if what Amelia thinks appears to be true.

A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.

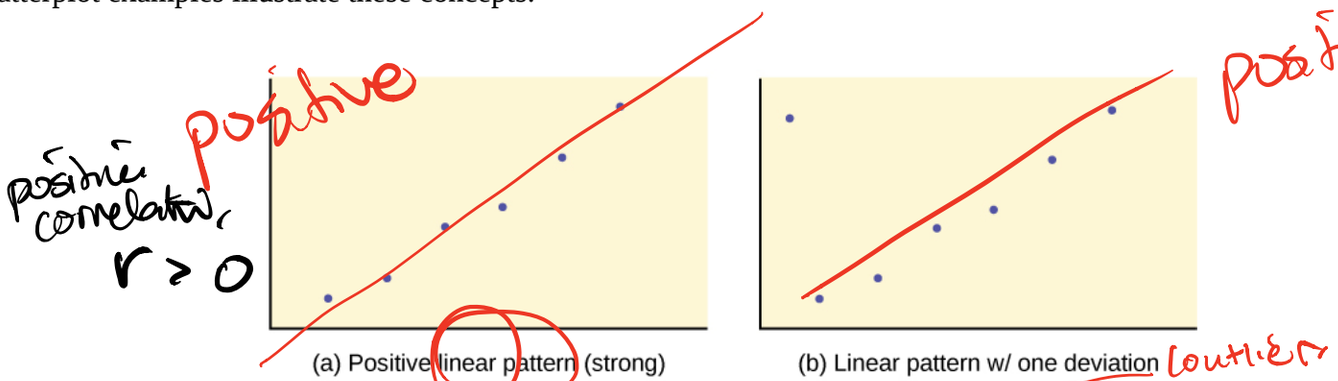


Figure 12.6

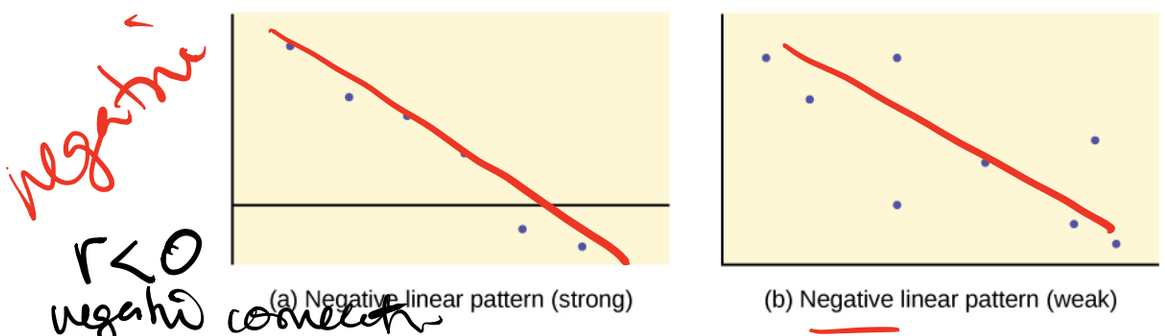
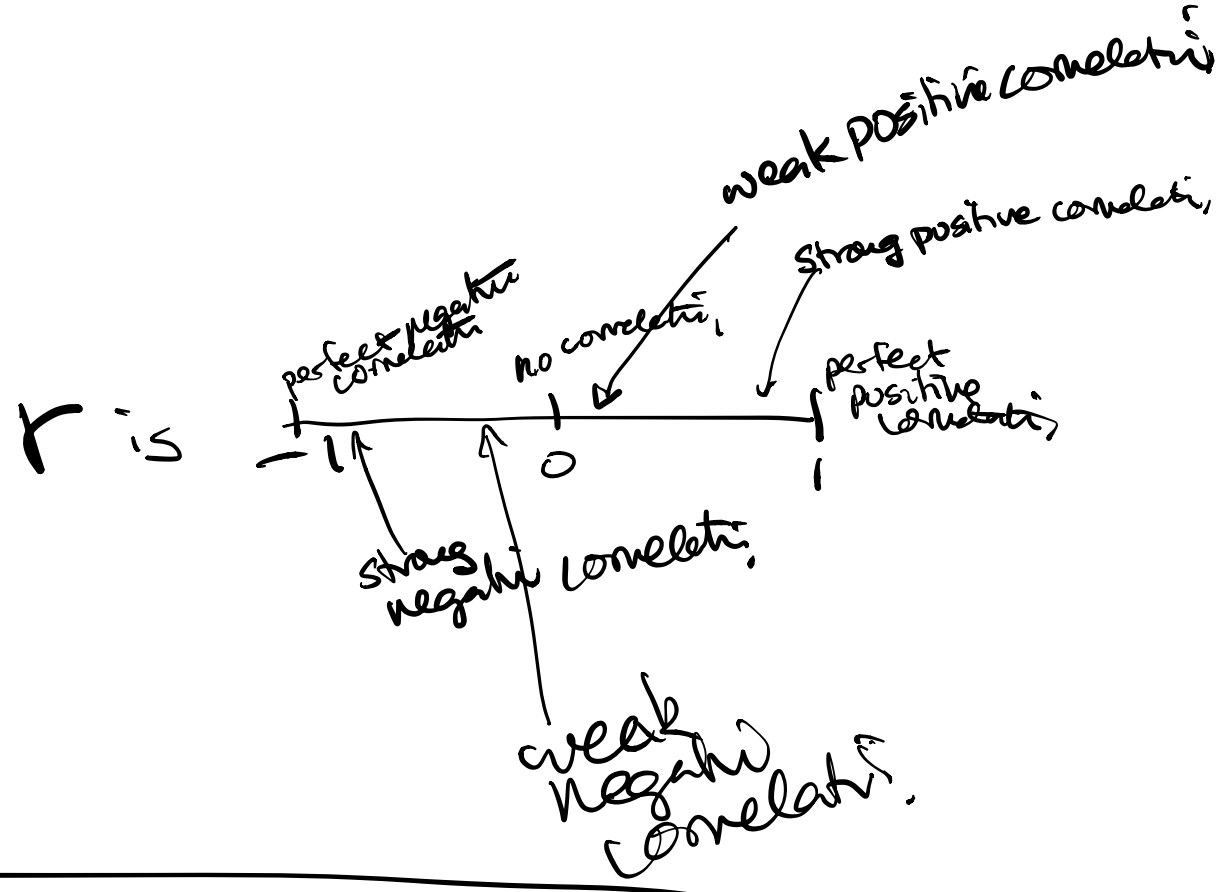


Figure 12.7



Fact:

r is always $-1 \leq r \leq 1$

Definition r^2 coefficient of determination and $0 \leq r^2 \leq 1$

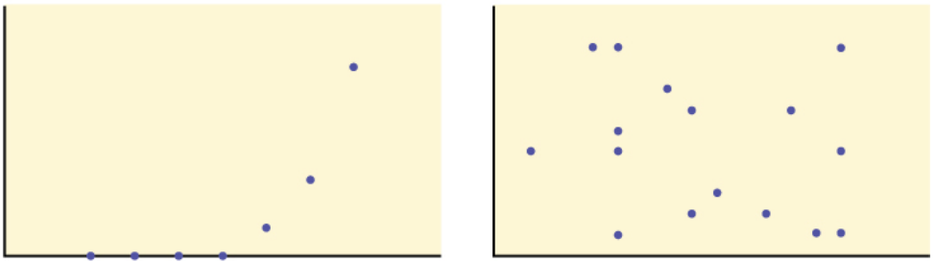


Figure 12.8

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

12.3 | The Regression Equation

Formal to get the Best Fit Line

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit** or **Least-Squares Line**.

Collaborative Exercise

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Example 12.6

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

NOTE

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

1. Make sure you have done the scatter plot. Check it on your screen.
2. Go to LinRegTTest and enter the lists.
3. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
4. Press Y = (you will see the regression equation).
5. Press GRAPH. The line will be drawn."

The Correlation Coefficient *r*

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between *x* and *y*.

The **correlation coefficient, *r*,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable *x* and the dependent variable *y*.

The correlation coefficient is calculated as

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where n = the number of data points.

If you suspect a linear relationship between *x* and *y*, then *r* can measure how strong the linear relationship is.

What the VALUE of *r* tells us:

- The value of *r* is always between –1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation *r* indicates the strength of the linear relationship between *x* and *y*. Values of *r* close to –1 or to +1 indicate a stronger linear relationship between *x* and *y*.
- If $r = 0$ there is absolutely no linear relationship between *x* and *y* **(no linear correlation)**.
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course,in the real world, this will not generally happen.

What the SIGN of *r* tells us

- A positive value of *r* means that when *x* increases, *y* tends to increase and when *x* decreases, *y* tends to decrease **(positive correlation)**.
- A negative value of *r* means that when *x* increases, *y* tends to decrease and when *x* decreases, *y* tends to increase **(negative correlation)**.
- The sign of *r* is the same as the sign of the slope, *b*, of the best-fit line.

NOTE

Strong correlation does not suggest that *x* causes *y* or *y* causes *x*. We say **"correlation does not imply causation."**

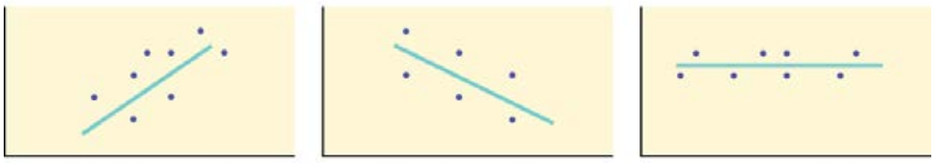


Figure 12.13 (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called the coefficient of determination and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the **third exam/final exam example** introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation (1 – 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

we will learn how to compute these numbers.

12.4 | Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
ρ = population correlation coefficient (unknown)
r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the

I. Regression

Given the 4 data below, find the regression line prediction $\hat{y}$ for $x = 80$.

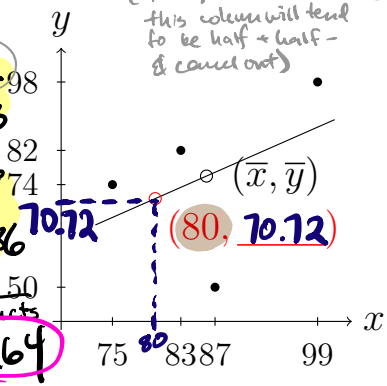
when signs of x & y deviations match they cancel out & deviation product is positive which is evidence for positive correlation

product deviations

(if it's just random noise this column will tend to be half + half − & cancel out)

x	y	(deviations) $x-\bar{x}$	(deviations) $y-\bar{y}$	(square deviating) $(x-\bar{x})^2$	(square deviating) $(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
75	74	$75-86=-11$	$74-76=-2$	$(-11)^2=121$	$(-2)^2=4$	$(-11)(-2)=22$
83	82	$83-86=-3$	$82-76=6$	$(-3)^2=9$	$(6)^2=36$	$(-3)(6)=-18$
87	50	$87-86=1$	$50-76=-26$	$(1)^2=1$	$(-26)^2=676$	$(1)(-26)=-26$
99	98	$99-86=13$	$98-76=22$	$(13)^2=169$	$(22)^2=484$	$(13)(22)=286$
Σx sum: 344	$\Sigma y = 304$	sum of deviations is always 0		sum of squares $SS_x = 300$	$SS_y = 1200$	sum of products $SP = 264$
sum / n: $\div 4 =$ 86	$\bar{y} = 76$			sum / $(n-1)$: $\frac{300}{3} = 100$	400	88 $= \frac{264}{3} = \frac{264}{n-1}$
$= \bar{x}$ mean	$= \bar{y}$			variance $= s_x^2$	$= s_y^2$	$= c$ = covariance
				stdev $= \sqrt{\text{variance}} = \sqrt{100} = 10$	$\sqrt{400} = 20$	$88 \div 10 \div 20 = 0.44 = r$ = correlation coefficient
				$= s_x$	$= s_y$	$= r$

There are 4 points so $n = 4$



Recall from lesson (F) (on Feb 10)

s = standard deviation

s_x = standard deviation for the x column

(simplified) equation of regression line.

$$\hat{y} = 0.32 + 0.88x$$

$$\hat{y} = b_0 + b_1 x$$

$0.88 = b_1$ (from below)

Final answer: $70.72 = \hat{y}(80)$

$\hat{y}(80) = 0.32 + 0.88(80) = 70.72$

$\hat{y} = b_1(x - \bar{x}) + \bar{y}$

$= 0.88(x - 86) + 76 = 0.88x - 75.68 + 76 = 0.88x + 0.32$

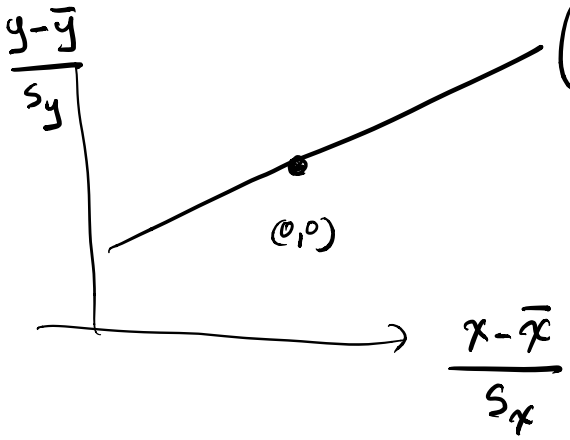
$b_1 = r\frac{s_y}{s_x} = \frac{c}{s_x^2} = \frac{SP}{SS_x}$

$\frac{264}{300} = 0.88$

$(0.44)\frac{(20)}{10} = 0.88$

$\frac{88}{100} = 0.88$

Explain the slope of the regression line b_1



$$\left(\frac{y-\bar{y}}{s_y}\right) = r\left(\frac{x-\bar{x}}{s_x}\right)$$

one form of the equation of the regression line.

Simplify $\left(\frac{y-\bar{y}}{s_y}\right) = r\left(\frac{x-\bar{x}}{s_x}\right)$

$$s_y \frac{y-\bar{y}}{s_y} = r\frac{x-\bar{x}}{s_x} s_y$$

multiply both sides by s_y

$$(y-\bar{y}) = r\frac{(x-\bar{x})}{s_x} s_y$$

$$y-\bar{y} = \frac{r s_y}{s_x}(x-\bar{x})$$

another popular form of the formula for regression line

$$y = (r s_y / s_x)(x-\bar{x}) + \bar{y}$$

$$b_1 = r s_y / s_x$$

the slope of the regression line

Explain the b_1 shortcut

$$b_1 = r\frac{s_y}{s_x} = \left(\frac{c}{s_x s_y}\right)\frac{s_y}{s_x} = \frac{c}{s_x^2} = b_1$$

shortcut

second shortcut: $b_1 = \frac{C}{s_x^2} = \frac{SP/(n-1)}{SS_x/(n-1)} = \boxed{\frac{SP}{SS_x} = b_1}$