Since the relationship is known in advance, students would be able to control their grade by choosing their study time appropriately. In fact, if the exact relationship were known in advance, taking the test would be necessary only as an intellectual exercise, since the grade would already be determined by the time studied and the model. Admittedly, there is no model that can precisely predict a test score solely on the basis of time studied; since there are many other variables that affect test scores. But suppose a model were available which, although imperfect, fairly reliably predicted test scores based on hours studied.

Test Score=45+3.8(Hours of Study Time)+Error

The new model introduces the error term. Now, if someone studies 10 hours, the model would predict

Test Score=45+3.8(10)+Error=83+Error.

Oct 18-10:14 AM

The predicted test score would still be 83, but there is an unknown random error associated with the prediction. If the error is reasonably small (say, at most 5 points), then the prediction will still be useful for planning purposes. But if the error is too large, then it will be difficult to rely on the model's predictions. If a model admits the possibility of an error, then gauging the expected magnitude of the error is essential in determining the model's usefulness. Estimating the mean and variance of the errors will be an important part of determining model utility. A model with a mean error of zero and small variation in the error terms would be desirable and should yield useful predictions. The model we have been using is simple. If two variables appear to be related in a straight line manner, we can use a simple linear regression model to describe their relationship.

Oct 18-10:16 AM

The model-building process begins with a desire to find a relationship between two or more variables. Since there are a great number of possible models that can be selected, choosing the type of model to represent the relationship is a complex problem.

A straight line is the simplest relationship between two variables. This straight line relationship is modeled by the simple linear regression model given by the following linear equation.

Definition

A **linear relationship** is graphically described as a line. Mathematically, a line is a set of points that satisfy the functional relationship

$$y = mx + b$$

where m is the **slope** of the line and b is the point where the function crosses the y-axis, which is called the **y-intercept**.

Oct 18-3:39 PM

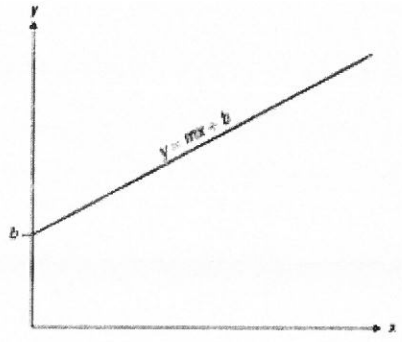


Figure 5.13

The slope determines if the line slopes upward (positive slope) or if the line slopes downward (negative slope). The relationship specified in Figure 5.13 is the linear equation

$$y = 5x + 3$$

Slope Intercept

In this case $m = 5$ and $b = 3$.

Once the relationship is specified, y is completely determined by the value of x. If $x = 10$, then

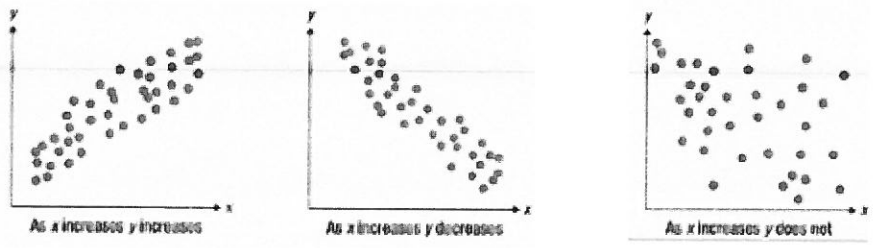
$$y = 5 \cdot 10 + 3 = 53$$

Oct 18-3:40 PM

Together, the slope and the intercept are called the parameters of the linear equation. That is, they completely define the equation of the line.

Developing a model to fit real world measurements is not trivial. Nature doesn't cooperate by requiring all relationships to be straight lines. Seldom, in fact, do pairs of measurements fall on perfectly straight lines.

If a linear relationship exists, the data will have some general tendency to move together or in opposite directions, as in Figure 5.15. In later sections we will look at ways of measuring the degree of linear relationship between two variables as well as defining the exact parameters (slope and intercept) of a line for a specific set of data.



Oct 18-3:41 PM

Measuring the Degree of Linear Relationship: The Correlation Coefficient

A scatter diagram is a useful exploratory tool for detecting relationships between two variables. Eventually, however, a researcher will want to know the strength of the relationship between the two variables. Karl Pearson developed a measure in 1896 called the correlation coefficient, r, to measure the degree of linear relationship

Formula: Correlation Coefficient

The correlation coefficient is an index number used to summarize the strength of a linear relationship.

$$r=\frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)\right\} \qquad -1\le r\le 1$$

Oct 18-3:43 PM

Formula: Correlation Coefficient

The correlation coefficient is an index number used to summarize the strength of a linear relationship.

$$r = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)\right\} \qquad -1 \le r \le 1$$

Within the parentheses, there are two familiar expressions:

$$\frac{y_i - \bar{y}}{s_y}$$

which is a z-score that shows how far y deviates from its mean measured in standard deviation units (sy is the standard deviation of y)

$$\frac{x_i - \bar{x}}{s_x}$$

which is a z-score that shows how far x deviates from its mean measured in standard deviation units (sx is the standard deviation of x)

Summing the products of these deviation measures for each data pair determines the sign of the correlation coefficient.

Oct 18-3:46 PM

Formula: Computational Formula for the Correlation Coefficient

The computational formula for the correlation coefficient is as follows:

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

MUST MEMORIZE

NOT!!

Oct 18-3:50 PM

To turn the DiagnosticOn:

Press [2nd] [0] [x-1] to access the CATALOG and jump to the "D" section of the commands.

Press the [Down Arrow] until you reach DiagnosticOn.

Enter.

Compute a regression, for example LinReg(a+bx), and the values for r and r^2 will be displayed.

Oct 18-4:01 PM

Example 5.1

Table 5.3 contains the age and the annual maintenance costs of a certain model compact vehicle. Calculate the correlation coefficient to determine if a relationship exists between the age and annual maintenance cost of that model.

Table 5.3 – Vehicle Age and Maintenance Cost

Observation	Age	Annual Maintenance Cost (Dollars)
1	2	225
2	4	400
3	5	475
4	7	650
5	9	800
6	12	1175

$$r = \frac{6(30275)-(39)(3725)}{\sqrt{6(319)-(39)^2}\sqrt{6(2879375)-(3725)^2}} \approx 0.9950$$

Given the large magnitude of r, one can say that there is a strong, positive, linear relationship between the age of the vehicle and the annual maintenance cost. That is, as the vehicle gets older, one can expect to spend more money on annual maintenance.

Oct 18-3:52 PM

Properties of the Correlation Coefficient

This leads to the following properties of the correlation coefficient.

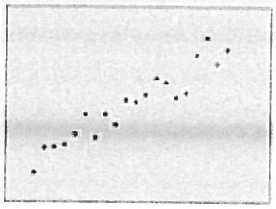
Properties

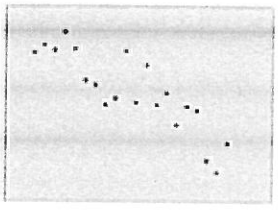
The correlation coefficient, r, measures the degree of linear relationship; i.e., how well the data cluster around a line.

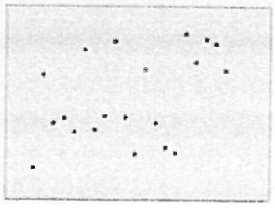
- The value of r is always between -1 and $+1$.
- A value of r near -1 or $+1$ means the data are tightly bundled around a line.
- A value of r near -1 or $+1$ means that we have a very strong negative relationship or positive relationship, respectively, such that predictions within the scope of the data are very reliable.
- Positive association is indicated by $r>0$ and an upward sloping relationship.
- Negative association is indicated by $r<0$ and a downward sloping relationship.
- A value of r near zero means there is no linear relationship between x and y.
- It does not matter whether you correlate y with x or x with y, you will still get the same value for r.

Oct 18-4:04 PM

-.86 -.05 .95







Oct 18-4:04 PM

Avoiding Some Correlation Pitfalls

A high correlation does not imply causation. Suppose that a high correlation has been observed between the weekly sales of ice cream and the number of snake bites each week. It seems unlikely that ice cream sales would cause snakes to bite people or that more snake bites would cause higher ice cream sales. Yet when the data are analyzed, you may find an unexpectedly high correlation. If the two variables aren't actually related, what could explain such an observed relationship?

The apparent relationship is an illusion caused by a phenomenon called common response. That is, both variables are related to a third variable. In this case the high temperatures in the summer cause increases in both ice cream sales and reptile activity.

Oct 18-4:06 PM

Suppose there is good reason to believe that a causal relationship exists between two variables, but when a correlation is performed the value of the correlation is near zero, indicating no association. Does the lack of correlation between the two variables prove no relationship exists? There are several reasons two related variables might not have a high correlation.

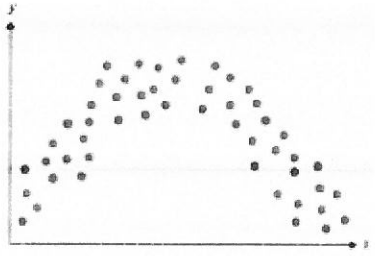


Figure 5.18

A low correlation could mean that no linear relationship exists. In Figure 5.18 the relationship between x and y is not a straight line. The correlation measure for these points is going to be very close to zero. Yet, there does appear to be a very strong relationship between x and y.

The kind of relationship exhibited by this data is called a quadratic relationship

Oct 18-4:08 PM

Another problem that can produce low correlations is **confounding**. Confounding occurs when more than one variable affects the dependent variable, and the effects of the variables cannot be distinguished from each other. Suppose that the **variable y is dependent on x. Thus, as x changes, it produces changes in y. Such a relationship should produce a** significant correlation measure between the two variables. But suppose there is another variable z, which also affects y. As z changes so does y. It is certainly possible that changes in z will mask the changes caused by x.

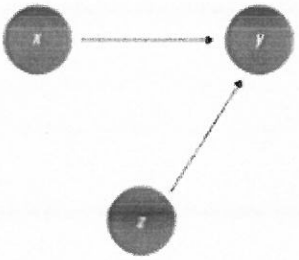


Figure 5.19

Oct 18-4:11 PM

The range of x-values selected for correlation can also significantly affect the value of the correlation coefficient. If the range of the x data is large, the correlation will usually be greater than if the range of the x-values is small, as shown in Figure 5.20.

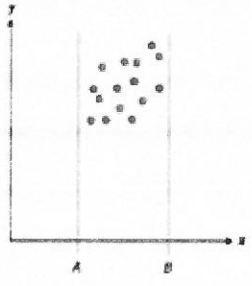


Figure 5.20

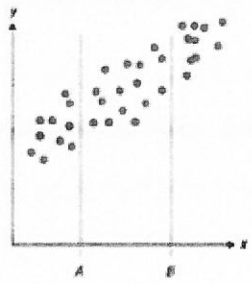


Figure 5.21

Oct 18-4:12 PM

Sometimes unrelated variables are highly correlated. When this occurs the variables are said to have a spurious correlation. For example, over a short period of time daily car sales and the number of penguins in Antarctica might be related. However, it is doubtful that a significant change in the penguin population will cause a change in car sales, or vice versa.

Oct 18-4:13 PM