

# Unit 10

## Using Correlation to Describe a Linear Relationship

Objectives:

- To obtain and interpret the Pearson Product Moment correlation between two quantitative variables
- To interpret the Pearson Product Moment correlation as a measure of the strength of the linear relationship between two quantitative variables

We have previously considered the scatter plots of Figures 9-1a to 9-1j to illustrate different types of relationships which can exist between two quantitative variables. Recall that an initial goal when studying a possible relationship between two variables is to decide whether the relationship is linear or nonlinear. The scatter plots of Figures 9-1a to 9-1f show linear relationships of varying degrees of strength, while the scatter plots of Figures 9-1g and 9-1h each show no real relationship, and the scatter plots of Figures 9-1i and 9-1j each show a nonlinear relationship. There are statistical techniques for deciding whether or not a relationship is linear, but since we are not yet in a position to use any of these techniques, we shall use our best judgment to make these decisions after looking at a scatter plot.

We are going to focus primarily on linear relationships between two quantitative variables; we shall often refer to one variable as  $X$  and the other as  $Y$ . Recall that we represent a list of  $n$  observations of a variable  $X$  by  $x_1, x_2, \dots, x_n$ . Similarly, we represent a list of  $n$  pairs of observations of the variables  $X$  and  $Y$  by

$$(x_1, y_1) (x_2, y_2) \dots (x_n, y_n)$$

Such data is called *bivariate data*. One example of bivariate data would be the observations on the two variables “age” and “yearly income” in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1.

One measure of dispersion for a quantitative variable, which we have defined previously, is the variance. The variance, denoted by  $s^2$ , was defined to be the sum of the squared deviations from the mean divided by one less than the number of observations. Recall that the deviations from the mean for the  $X$  observations are

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x},$$

and that the deviations from the mean for the  $Y$  observations are

$$y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}.$$

Also, recall that the variance for the  $X$  observations and the variance for the  $Y$  observations are respectively

$$s_x^2 = \frac{\sum (x - \bar{x})^2}{n-1} \quad \text{and} \quad s_y^2 = \frac{\sum (y - \bar{y})^2}{n-1}.$$

Let us calculate the variance for the variable “age” and the variance for the variable “yearly income” in the SURVEY DATA. We shall designate the variable “age” as  $X$  and the variable “yearly income” as  $Y$ . To find the deviations from the mean for each variable, we first find the means. Table 10-1 displays calculations useful in obtaining the means and variances. The first two columns of the table represent lists of the ages and the yearly incomes respectively. The sum of the ages (as you may verify) is  $\Sigma x = 1323$ , and the sum of the yearly incomes (as you may verify) is  $\Sigma y = 1362$ . Since there  $n = 30$  observations, the mean age is  $\bar{x} = 1323/30 = 44.1$  years, and the mean yearly income is  $\bar{y} = 1362/30 = 45.4$  thousand dollars. The third and fifth columns of the table represent respectively the deviations of age from the mean and the deviations of

yearly income from the mean; as indicated in the table, the sum of the deviations from the mean is always zero. The fourth and sixth columns of the table represent respectively the squared deviations from the mean for age and yearly income. (The seventh column of the table we shall discuss shortly.) At the bottom of the table, the variances are obtained by dividing the sum of the squared deviations from the mean divided by one less than the number of observations. Specifically, the variance of the ages is  $s_x^2 = 3814.70/29 = 131.5414$ , and the variance of the yearly incomes is  $s_y^2 = 7327.20/29 = 252.6621$ .

Recall that the square root of the variance is the standard deviation; variance and standard deviation are measures of dispersion in the distribution of a quantitative variable. Specifically, the standard deviation of the ages is  $s_x = 11.4691$  years, and the standard deviation of the yearly incomes is  $s_y = 15.8953$  thousand dollars. Obtaining the variance for observations of a quantitative variable is nothing new; we have done this previously. The new concept that we now want to introduce is numerical measures of the relationship between two quantitative variables.

The *covariance* between two quantitative variables  $X$  and  $Y$  can be used as a measure of whether the relationship between the two variables tends to be positive or negative. Using  $c$  to denote the covariance, we may write a formula to calculate the covariance as follows:

$$c = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

From this formula, we see that the covariance is the sum of the product of deviations from the mean divided by one less than the number of observations. The seventh column of Table 10-1 represents the product of the deviations from the mean for age and yearly income; specifically, each entry of the seventh column is the product of the corresponding entries in the third column and the fifth column. To find the covariance between age and yearly income in the SURVEY DATA, the sum of the seventh column  $\Sigma(x - \bar{x})(y - \bar{y}) = +2522.80$  is divided by one less than the number of observations  $n - 1 = 29$ . At the bottom of Table 10-1 is where we find this covariance to be  $c = +86.9931$ .

A variance  $s^2$  can never be negative, since it is calculated from a sum of squares. Unlike the variance, however, a covariance  $c$  can be negative or positive. If the relationship between  $X$  and  $Y$  tends to be positive, then  $X$  and  $Y$  will tend to increase together and decrease together, which implies that  $(x - \bar{x})$  and  $(y - \bar{y})$  will often either both be positive or both be negative, which in turn implies  $(x - \bar{x})(y - \bar{y})$  will often be positive (since the product of positive numbers is positive, and the product of two negative numbers is positive). Consequently, when the relationship between  $X$  and  $Y$  tends to be positive, then the covariance  $c$  will be positive. On the other hand, if the relationship between  $X$  and  $Y$  tends to be negative, then  $Y$  will tend to decrease as  $X$  increases and vice versa, which implies that  $(x - \bar{x})$  and  $(y - \bar{y})$  will often both have opposite signs from one another, which in turn implies  $(x - \bar{x})(y - \bar{y})$  will often be negative (since the product of a positive number and a negative number is negative). Consequently, when the relationship between  $X$  and  $Y$  tends to be negative, then the covariance  $c$  will be negative.

**Table 10-1**  
**Calculation of the Means, Variances, and Covariance**  
**for  $X = \text{"age"}$  and  $Y = \text{"yearly income"}$**   
**in the SURVEY DATA**

$x$	$y$	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
35	34	- 9.1	82.81	-11.4	129.96	+ 103.74
20	28	-24.1	580.81	-17.4	302.76	+ 419.34
35	71	- 9.1	82.81	+25.6	655.36	- 232.96
.	.	.	.	.	.	.
.	.	.	.	.	.	.
53	64	+ 8.9	79.21	+18.6	345.96	+ 165.54
<u>59</u>	<u>39</u>	<u>+14.9</u>	<u>222.01</u>	<u>- 6.4</u>	<u>40.96</u>	<u>- 95.36</u>
1323	1362	0	3814.70	0	7327.20	+2522.80
		$\bar{x} = 1323/30 = 44.1$			$\bar{y} = 1362/30 = 45.4$	
		$s_x^2 = 3814.70/29 = 131.5414$			$s_y^2 = 7327.20/29 = 252.6621$	
		$c = +2522.80/29 = +86.9931$				

With a scatter plot similar to Figures 9-1g and 9-1h, we should expect the covariance to be close to zero, since there appears to be neither a positive relationship nor a negative relationship between  $X$  and  $Y$ ; in fact, there appears to be no relationship at all. However, do not be fooled into thinking that a covariance close to zero implies there is no relationship between  $X$  and  $Y$ . A scatter plot similar to Figure 9-1j would incline us to believe that there is a nonlinear relationship between  $X$  and  $Y$ , but we would expect the covariance to be close to zero, since the relationship appears to be neither a positive one nor a negative one.

In Table 10-1, the fact that we found the covariance between age and yearly income in the SURVEY DATA to be positive suggests that age and yearly income increase together and decrease together. The scatter plot of Figure 7-4 visually displays this positive relationship and also seems to suggest that the relationship might be considered linear but weak.

As another example of bivariate data, suppose  $X$  = “the dosage of a certain drug (in grams)” and  $Y$  = “the reaction time to a particular stimulus (in seconds)” are recorded for each subject involved in a study, with the results displayed in Table 10-2. This data is repeated in the first

two columns of Table 10-3, where we shall have you calculate the means, the variances, and the covariance. In Table 10-3, complete the last five columns, find the sum of the entries for each column, and obtain the statistics listed at the bottom of the table. (You should find that  $\bar{x} = 7$ ,  $\bar{y} = 4.1$ ,  $s_x^2 = 5.714$ ,  $s_y^2 = 4.720$ , and  $c = -5$ .)

The covariance  $c$  is really only capable of giving us some information about whether the relationship between two quantitative variables appears to be positive or negative; the covariance does not give us any sense about the strength of a relationship. We can get some measure of the strength of a linear relationship by using a *correlation*. Many different types of correlations are available, and often the word *correlation* is used to refer to a measure of the strength of a relationship between two variables even when one or both of the variables is not quantitative. Presently, we are going to focus only on one particular type of correlation used to measure the strength of a linear relationship between two quantitative variables  $X$  and  $Y$ .

One popular measure of the strength and direction of a linear relationship between two quantitative variables  $X$  and  $Y$  is called the *Pearson product moment correlation*. When there is no chance of confusion with any other type of correlation, we shall simply refer to the Pearson product moment correlation as the correlation, and we shall use the letter  $r$  to denote this correlation. The correlation  $r$  between two quantitative variables  $X$  and  $Y$  is defined to be the covariance divided by the product of the standard deviations, that is, we may write

<b>Table 10-2</b>								
<b>Dosage and Reaction Time Data</b>								
The dosage of a stimulant drug and the reaction time to a stimulus are recorded for each of several subjects injected with the drug.								
Dosage (grams)	4	4	6	6	8	8	10	10
Reaction Time (seconds)	7.5	6.8	4.0	4.4	3.9	3.1	1.4	1.7

<b>Table 10-3</b>						
<b>Calculation of the Means, Variances, and Covariance for the Dosage and Reaction Time Data of Table 10-2</b>						
$x$	$y$	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
4	7.5					
4	6.8					
6	4.0					
6	4.4					
8	3.9					
8	3.1					
10	1.4					
10	1.7					
$\bar{x} =$ _____		$\bar{y} =$ _____				
$s_x^2 =$ _____		$s_y^2 =$ _____		$c =$ _____		

$$r = \frac{c}{s_x s_y} .$$

While the covariance  $c$  could be equal to any real number, the value of  $r$  will always be between  $-1$  and  $+1$ , although we are not going to prove this fact here.

Roughly speaking, a value of  $r$  close to  $0$  suggests no linear relationship, while a value of  $r$  close to  $-1$  suggests a negative linear relationship, and a value of  $r$  close to  $+1$  suggests a positive linear relationship. A perfect positive linear relationship corresponds to  $r = +1$ , and a perfect negative linear relationship corresponds to  $r = -1$ . Recall once again that Figures 9-1a and 9-1b are each a scatter plot illustrating a perfect linear relationship between two quantitative variables, since in both figures the data points all lie exactly on a straight line. We would of course find that  $r = +1$  for Figure 9-1a and that  $r = -1$  for Figure 9-1b. In Figure 9-1c we might expect that  $r$  is less than but relatively close to  $+1$ , and in Figure 9-1d we might expect that  $r$  is greater than but relatively close to  $-1$ . The scatter plots of Figures 9-1e and 9-1f each illustrate a weaker linear relationship than those of Figures 9-1c and 9-1d respectively; consequently, we might expect that in Figure 9-1e  $r$  is closer to zero but still positive, and that in Figure 9-1f  $r$  is also closer to zero but still negative.

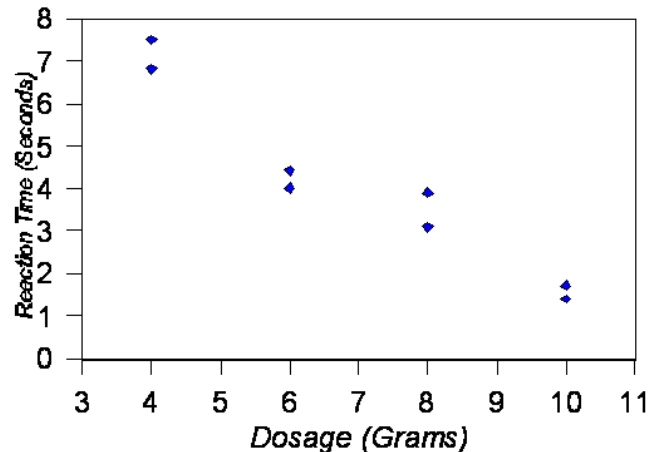
The scatter plots of Figures 9-1g and 9-1h each show no real relationship; consequently, in both cases we would expect that  $r$  will be close to  $0$ . The scatter plots of Figures 9-1i and 9-1j each show a nonlinear relationship. While we have previously noted that the relationship displayed by Figure 9-1i could be called negative, we cannot really call the relationship displayed by Figure 9-1j either positive or negative; as a result, we would expect that  $r$  will be negative for Figure 9-1i and that  $r$  will be very close to  $0$  for Figure 9-1j. As with the covariance, do not be fooled into thinking that a correlation close to zero implies there is no relationship between  $X$  and  $Y$ ; a correlation close to zero only implies that there is no linear relationship between  $X$  and  $Y$ .

To calculate the correlation  $r$  between age and yearly income in the SURVEY DATA, we obtain from our earlier calculations in Table 10-1 that  $s_x = 11.4691$ ,  $s_y = 15.8953$ , and  $c = +86.9931$ . Consequently, the correlation between age and

yearly income is  $r = +86.9931 / [(11.4691)(15.8953)] = +0.4772$ . Use the calculations from Table 10-3 to find the correlation between dosage and reaction time in the data of Table 10-2. (You should find that  $r = -0.9628$ .)

**Figure 10-1**

**Scatterplot for the Data of Table 10-2**



**Table 10-4**

**Age and Grip Strength Data**

The age (years) and right-hand grip strength (pounds of force) are recorded for each of several right-handed males.

Age	15	17	19	11	16	22	17	25	12	14	25	23
Grip Strength	50	54	66	46	58	54	64	80	46	70	76	80

**Self-Test Problem 10-1.** The Age and Grip Strength Data, displayed in Table 10-4, is used in a study of the relationship between age and grip strength among right-handed males. Let  $X$  represent the variable “age,” and let  $Y$  represent the variable “grip strength”.

- Find the mean for each variable and the variance for each variable.
- Find the covariance and correlation between the two variables.
- Based on the value of the correlation, which of Figures 9-1a to 9-1j would you expect a scatter plot of this data to resemble? Why?
- Construct a scatter plot of the data with age on the horizontal axis, and decide whether you think the relationship should be considered linear or non-linear.

It is easy to be tempted to think that a correlation between  $+0.75$  and  $+1$ , or a correlation between  $-0.75$  and  $-1$ , must indicate a reasonably strong linear relationship, and also that a correlation between  $-0.25$  and  $+0.25$  must indicate a weak linear relationship. However, this is not necessarily the case. The number of observations  $n$  is an important consideration in whether not a given value of  $r$  should be considered significant. This important fact is often forgotten when evaluating the strength of a linear relationship. Precisely how to decide when a correlation should be considered significant is a topic which we shall address in a future unit. For now, however, in order to demonstrate the important role  $n$  plays in interpreting correlation, we point out that, in general, a large value of  $r$  should not necessarily be considered significant if  $n$  is small; also, when  $n$  is large, what may appear to be a small value of  $r$  may very well be considered significant. For instance, a value of  $r = 0.35$  with  $n = 100$  would be considered very significant, whereas a value of  $r = 0.80$  with  $n = 5$  would not be considered significant. Quoting the value of a correlation is popular with some researchers, but if such a correlation value is not accompanied by the value of  $n$  and/or some other indication of its significance, the value of the correlation is of little value and is open to misinterpretation.

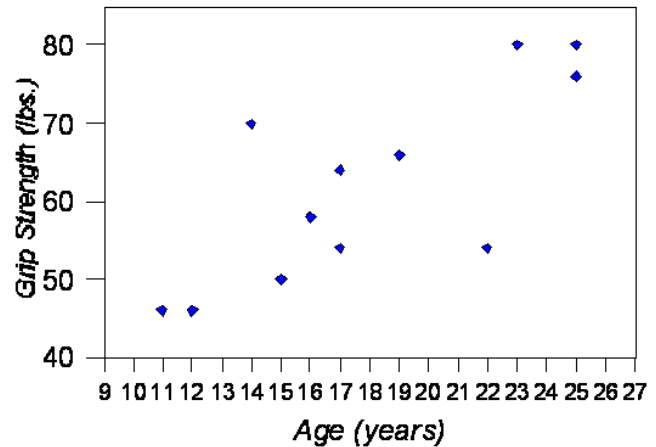
Recall that the correlation between drug dosage and reaction time in the data of Table 10-2 was found to be  $r = -0.9628$ . While this appears to represent a very strong negative linear relationship, the fact that  $n = 8$  should make us cautious about jumping to such a conclusion. We shall have to wait until a future unit before we learn how to decide whether or not the linear relationship between age and grip strength should be considered significant.

A few other cautionary remarks about correlation are in order. The fact that there is a linear relationship (or any other type of relationship) between two variables  $X$  and  $Y$  does not necessarily imply a cause-and-effect relationship. For instance, we should not be surprised to observe a high positive correlation between thumb size and reading ability among school children in grades one through seven, but this certainly should not lead us to conclude that a large thumb size causes higher reading ability or vice versa. The reason for a high correlation in this situation is of course because both of the variables involved are related to a third variable: age.

A correlation between two variables which results from each variable being highly correlated to one or more other variables is called a *spurious correlation*. (Spurious means "wrongly attributed.") We may find that there is a high positive correlation between smoking and risk of lung cancer; but while smoking may be a contributory factor to lung cancer, it is possible for lung cancer to be caused by any number of factors. Conceivably, there may exist some third factor which simultaneously influences both a person's desire to smoke and risk of lung cancer.

Figure 10-2

Scatterplot for the Data of Table 10-4



**Self-Test Problem 10-2.** Figure 7-3 displays a scatter plot for the variables “weekly radio hours” and “weekly TV hours” in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1. Based on this scatter plot, do you think

- the relationship between “weekly radio hours” and “weekly TV hours” is linear or non-linear?
- the correlation between “weekly radio hours” and “weekly TV hours” is positive, negative, or close to zero?

**Self-Test Problem 10-3.** At the Department of Motor Vehicles in a state, a scatter plot is created from automobile accident records. The variable “miles away from home” is scaled on the horizontal axis, and the variable “number of accidents in the past year” is scaled on the vertical axis. A large negative correlation is found.

- (a) Is it appropriate to conclude that drivers tend to be safer from automobile accidents when they are farther away from home? Why or why not?
- (b) Why do you think this correlation might be considered to be a spurious correlation?

**Self-Test Problem 10-4.** Decide whether obtaining a Pearson product moment correlation  $r$  is appropriate in each of the situations described, and give a reason for your decision.

- (a) Data on type of job (white collar or blue collar) and diastolic blood pressure are recorded for each of several residents in a city, in order to study the relationship between type of job and diastolic blood pressure.
- (b) Weekly study time and weekly TV time are recorded for each of several high school students, in order to study the relationship between study time and time spent watching TV.
- (c) Eye color (classified as blue, brown, or green) and hair color (classified as brown, black, blond, and red) are recorded for a large number of people, in order to study a possible relationship between eye color and hair color.
- (d) Weekly study time and weekly TV time are recorded for each of several high school students, in order to study the difference in distribution between study time and time spent watching TV.
- (e) Monthly income is recorded for each of several high school students, in order to study the distribution monthly incomes among high school students.
- (f) Grade point average and weekly TV time are recorded for each of several high school students, in order to study the relationship between study time and grade point average.

#### Answers to Self-Test Problems

- 10-1** (a) The mean age is  $\bar{x} = 18$  years, and the mean grip strength is  $\bar{y} = 62$  lbs.; the variance of the ages is  $s_x^2 = 256/11 = 23.273$ , and the variance of the grip strengths is  $s_y^2 = 1728/11 = 157.091$ . (b) The covariance is  $c = +512/11 = +46.545$ , and the correlation is  $r = +0.7698$ . (c) Since the correlation is positive and seems relatively large, one might expect the scatter plot to resemble Figure 9-1c. (d) Figure 10-2 displays the scatter plot; it appears reasonable to say that the relationship is linear.
- 10-2** (a) It appears reasonable to say that the relationship is linear. (b) The scatter plot appears to show a reasonably strong negative correlation between weekly radio hours and weekly TV hours; the actual value of the correlation is  $-0.6384$ .
- 10-3** (a) The fact that there is a linear relationship (or any other type of relationship) between two variables  $X$  and  $Y$  does not necessarily imply a cause-and-effect relationship. (b) Each of the two variables “miles away from home” and “number of accidents in the past year” is going to be heavily influenced by a third variable “time spent driving”; the reason more accidents tend occur at shorter distances away from home is because people tend to spend considerably more time driving close to home than away from home.
- 10-4** (a) The correlation  $r$  would not be appropriate, since the relationship between two variables which are not both quantitative is of interest. (b) The correlation  $r$  would be appropriate, since the relationship between two variables which are both quantitative is of interest. (c) The correlation  $r$  would not be appropriate, since the relationship between two variables which are not both quantitative is of interest. (d) The correlation  $r$  would not be appropriate, since the interest is in the difference in distribution between the two quantitative variables and not in the relationship. (e) The correlation  $r$  would not be appropriate, since only one quantitative variable is of interest. (f) The correlation  $r$  would be appropriate, since the relationship between two variables which are both quantitative is of interest.

### Summary

We represent a list of  $n$  pairs of observations of quantitative variables  $X$  and  $Y$  by

$$(x_1, y_1) (x_2, y_2) \dots (x_n, y_n)$$

Such data is often called *bivariate data*. The *covariance*  $c$  between two quantitative variables  $X$  and  $Y$  is defined to be

$$c = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1} .$$

The covariance can be used as a measure of whether the relationship between the variables  $X$  and  $Y$  tends to be positive or negative. If the relationship between  $X$  and  $Y$  tends to be positive, then  $(x - \bar{x})$  and  $(y - \bar{y})$  will often tend to have the same sign, which, in turn, implies that the covariance  $c$  will be positive. If the relationship between  $X$  and  $Y$  tends to be negative, then  $(x - \bar{x})$  and  $(y - \bar{y})$  will often tend to have opposite signs, which, in turn, implies that the covariance  $c$  will be negative. The covariance being close to zero implies there is no linear relationship between  $X$  and  $Y$  but does not necessarily imply that there is no relationship between  $X$  and  $Y$ . The covariance provides us with information about the direction of a relationship but not about the strength of the relationship.

The *Pearson product moment correlation*, referred to simply as the correlation when there is no chance of confusion, is defined to be the covariance divided by the product of the standard deviations. Using  $r$  to represent this correlation, we may write

$$r = \frac{c}{s_x s_y} .$$

The value of  $r$  will always be between  $-1$  and  $+1$ . Roughly speaking, a value of  $r$  close to  $0$  suggests no linear relationship, while a value of  $r$  close to  $-1$  suggests a negative linear relationship and a value of  $r$  close to  $+1$  suggests a positive linear relationship. A perfect positive linear relationship corresponds to  $r = +1$ , and a perfect negative linear relationship corresponds to  $r = -1$ . When there is no real relationship, or when we cannot really call the relationship either positive or negative, we would expect that  $r$  will be very close to  $0$ . The number of observations  $n$  is an important consideration in whether not a given value of  $r$  should be considered significant. A correlation between two variables which results from each variable being highly correlated to a third variable is called a *spurious correlation*.