

# Unit 11

## Using Linear Regression to Describe Relationships

Objectives:

- To obtain and interpret the slope and intercept of the least squares line for predicting a quantitative response variable  $Y$  from a quantitative explanatory variable  $X$

The correlation between two quantitative variables provides us with a measure the strength and direction of a linear relationship. Obtaining correlation does not require us to identify one of the two variables as being predicted from the other. However, there are situations in which we want to predict one variable from another. A *response variable* or *dependent variable* is one we predict from one or more other variables; an *explanatory variable* or *independent variable* is one from which predictions are made. Typically, we let  $Y$  represent the response (dependent) variable and let  $X$  represent the explanatory (independent) variable. With the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, we might wish to predict yearly income from age, making  $Y$  = "yearly income" the response variable and  $X$  = "age" the explanatory variable. With the Dosage and Reaction Time Data of Table 10-2, it is natural to think of predicting reaction time from drug dosage, making  $Y$  = "reaction time" the response variable and  $X$  = "drug dosage" the explanatory variable.

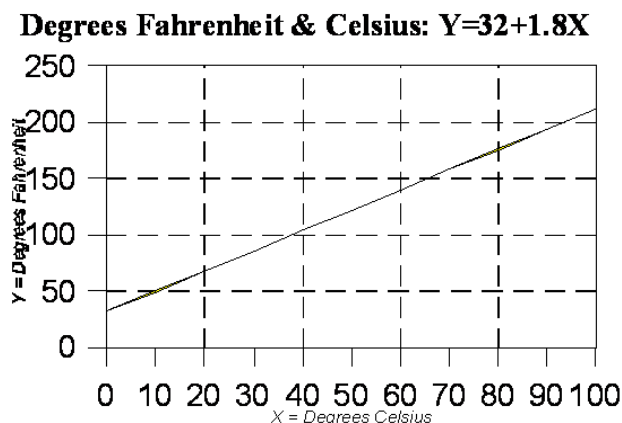
*Regression* refers to the prediction of one quantitative response variable from one or more quantitative explanatory variables. We will be primarily concerned with *simple linear regression*, which refers to the prediction of one response variable  $Y$  from one explanatory variable  $X$  using the equation of a straight line written in the form  $Y = a + bX$ . In this equation,  $a$  is called the *intercept* of the line, and  $b$  is called the *slope* of the line.

To illustrate the role played by the slope and intercept, let us suppose that  $X$  is temperature measured in degrees Celsius and  $Y$  is temperature measured in degrees Fahrenheit. It will always be true that  $Y = 32 + 1.8X$ , since this is the well known formula for changing from degrees Celsius to degrees Fahrenheit. For instance, to convert a temperature of 52 degrees Celsius to degrees Fahrenheit,  $X = 52$  is substituted into the equation, from which we find that 52 degrees Celsius is equal to  $Y = 32 + 1.8(52) = 125.6$  degrees Fahrenheit. The intercept of a line ( $a$ ) is the value of  $Y$  when  $X$  is equal to zero. When degrees Celsius is 0, then the degrees Fahrenheit is  $a = 32$ , since  $32 + 1.8(0) = 32$ . Next, we observe that when  $X = 1$ ,  $Y = 33.8$ ; when  $X = 2$ ,  $Y = 35.6$ ; when  $X = 3$ ,  $Y = 37.4$ ; etc., from which we find that each time  $X$  is increased by 1 unit,  $Y$  is increased by  $b = 1.8$  units. The slope of a line represents the amount of change in  $Y$  each time  $X$  is increased by 1 unit. In this case, each time degrees Celsius is increased by 1 degree, the degrees Fahrenheit is increased by 1.8 degrees.

To graph a line, all we really need to do is plot and connect any two distinct points on the line. For instance, to graph  $Y = 32 + 1.8X$ , which is done in Figure 11-1, we can connect the points (0, 32) and (100, 212), both of which are on the line.

There is a positive linear relationship between degrees Celsius and degrees Fahrenheit, because the slope is positive, implying that  $Y$  increases whenever  $X$  increases. With a negative linear relationship, the slope

Figure 11-1



is negative, implying that  $Y$  decreases whenever  $X$  increases. This is illustrated by the line in Figure 11-2. The equation of this line is  $Y = 10.5 - 0.6X$ , which has a slope of  $b = -0.6$  and an intercept of  $a = 10.5$ . It is easy to check that each time  $X$  is increased by 1,  $Y$  is decreased by 0.6. We have already seen that the slope ( $b$ ) is indeed  $-0.6$ , and it is also easy to see that when  $X = 0$ , then  $Y = 10.5$ , which is the intercept ( $a$ ).

Now, let us return to Figure 7-4, which (if we momentarily ignore the line in the figure) is the scatter plot of  $X =$  "age" and  $Y =$  "yearly income" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1. There appears to be a positive relationship, but this relationship is not a perfect positive one, since the points do not all lie on a straight line. If all the points in a scatter plot do not lie on a straight line, how can we possibly find the equation of a line which will describe this relationship? Since we cannot find the equation of a line which goes through every data point, we do the next best thing by finding the equation of the line which comes closest to all the data points in some sense. One popular method to do this is to find the line which minimizes the squared vertical distances between the data points and the line; this line is called the *least squares line*. The line that has been graphed with the scatter plot in Figure 7-4 is the least squares line.

We shall not attempt to compare the merits of the method of least squares with those of competing methods, nor shall we attempt to derive the equation of the line resulting from the method of least squares. We shall simply state the formulas for the slope and intercept to find the least squares line for predicting  $Y$  from  $X$  with a bivariate data set. The slope in the least squares line can be obtained from

$$b = r \frac{s_Y}{s_X} .$$

From this formula, one sees that the slope  $b$  of the least squares line and the correlation  $r$  must always have the same sign. Intuitively, this should be obvious, since a positive correlation suggests that  $Y$  tends to increase as  $X$  increases (implying a positive slope) and a negative correlation suggests that  $Y$  tends to decrease as  $X$  increases (implying a negative slope). Once the slope of the least squares line is available, the intercept can be obtained from

$$a = \bar{y} - b\bar{x} .$$

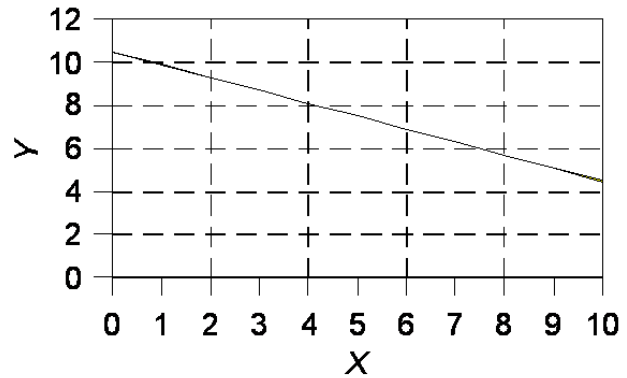
We can use the equation of the least squares line to make predictions about  $Y$  from  $X$  or to describe the relationship between  $X$  and  $Y$ .

Let us illustrate how we can obtain the least squares line for predicting  $Y =$  "yearly income" from  $X =$  "age" with the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1. Using the calculations from Table 10-1, we have previously found that  $\bar{x} = 44.1$ ,  $\bar{y} = 45.4$ ,  $s_x^2 = 11.4691$ ,  $s_y^2 = 15.8953$ , and  $r = +0.4772$ . Using the formula for the slope in the least squares line and then the formula for the intercept in the least squares line, we find that

$$b = r \frac{s_Y}{s_X} = (0.4772) \frac{15.8953}{11.4691} = 0.661 \quad \text{and} \quad a = \bar{y} - b\bar{x} = 45.4 - (0.661)(44.1) = 16.2 .$$

We could write the equation of the least squares line as  $Y = 16.2 + 0.661X$ ; however, in order to emphasize what  $X$  and  $Y$  represent, we might prefer to write the equation as  $inc = 16.2 + 0.661(age)$ , where *inc* is an abbreviation for yearly income in thousands of dollars. The slope of this least squares line,  $b = 0.661$  thousand dollars, is an estimate for the average amount of change in yearly income accompanying an increase of one year in age; that is, for each one year increase in age, the yearly income among the voters represented by the

**Figure 11-2**  
 $Y = 10.5 - 0.6X$



SURVEY DATA is estimated to increase on average by 661 dollars. Figure 7-4, which displays the graph of this least squares line on a scatter plot, provides a visual picture of how well the least squares line fits the data. At a later time, we shall discuss ways for deciding whether or not the fit is a good.

We may use the least squares line to predict yearly income with a given age or to estimate the average yearly income for a given age group, as long as age is within the range of the ages observed in the data. For example, we estimate the average yearly income for 50-year-old voters represented by the SURVEY DATA to be about  $16.2 + 0.661(50) = 49.25$  thousand dollars. If we wanted to predict the yearly income for a particular 30-year-old voter from the voters represented by the SURVEY DATA, our prediction would be  $16.2 + 0.661(30) = 36.1$  thousand dollars (which is of course the same as the estimated average yearly income for 30-year-olds).

As a general rule, estimation and prediction outside the range of the observed values of  $X$  should be avoided. For instance, if we blindly substituted an age of  $X = 10$  years into the least squares line, we would estimate the average yearly income for 10-year-olds to be  $16.2 + 0.661(10) = 22.8$  thousand dollars. We hope you agree that this is a totally meaningless estimate! Examination of the SURVEY DATA will reveal that the ages of the voters in the data ranged from 20 to 62 years old. This implies that our least squares line can be applied only to ages in this range. In fact, we should expect the relationship between age and yearly income to be quite different for at least some ages outside the range from 20 to 62 years.

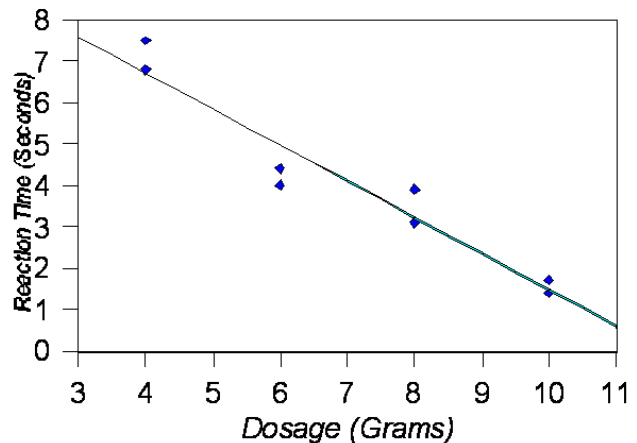
We shall now have you find the least squares line for predicting  $Y =$  “the reaction time to a particular stimulus (in seconds)” from  $X =$  “the dosage of a certain drug (in grams)” with the data of Table 10-2. It should be clear that reaction time is the response variable and dosage is the explanatory variable. In order to find the least squares line, recall that the calculations done in Table 10-3 were used to find that  $\bar{x} = 7$ ,  $\bar{y} = 4.1$ ,  $s_x^2 = 5.714$ ,  $s_y^2 = 4.720$ , and  $r = -0.9628$ . Once you have found the least squares line, graph this line on Figure 10-1, which is a scatter plot of the data. (You should find that the least squares line is  $rct = 10.225 - 0.875(dsg)$ , where  $rct$  represents reaction time in seconds, and  $dsg$  represents dosage in grams; Figure 11-3 is a graph of the least squares line on the scatter plot.)

The slope of this least squares line for predicting reaction time from dosage,  $b = -0.875$  seconds, is an estimate for the average amount of change in reaction time accompanying an increase of one gram in dosage; that is, for each one gram increase in dosage of the drug, the reaction time is estimated to decrease on average by 0.875 seconds. Note that since the relationship between reaction time and drug dosage is a negative one, both the slope  $b$  and the correlation  $r$  are negative. Use the least squares line to estimate the average reaction time with a drug dosage of 9 grams and to predict the reaction time with a dosage of 7 grams. (You should find that the estimated average reaction time with a drug dosage of 9 grams is 2.35 seconds, and that the predicted reaction time with a dosage of 7 grams is 4.1 seconds.)

Earlier, we stated that using the least squares line for estimation and prediction outside the range of the observed values of  $X$  should be avoided. There is one exception to this rule. When our explanatory variable  $X$  is time, and we are attempting to predict a quantitative variable  $Y$  for a future time period, then of course after observing values of  $Y$  for several time periods, our goal is to make predictions for future time periods outside the range of the data. Data observed over a sequence of time periods is called a *time series*. The first two columns of Table 9-1 is an example of time series data for prices, and the first and third columns of Table 9-1 is an example of time series data for quantities. One may attempt to use a least squares line with time series data to make predictions for the future, but such predictions will usually not be very accurate, because very few quantitative variables change over time in a purely linear fashion. Describing how a quantitative variable changes over time almost always requires the use of a nonlinear relationship. Although soon we shall discuss

**Figure 11-3**

**Scatterplot & Least Squares Line for Table 10-2**



ways of deciding whether the assumption of a linear relationship is warranted (as opposed to some other type of relationship), we shall not have time to explore the many different nonlinear relationships which exist.

There is one final remark that we shall make concerning the topics of correlation and linear regression. Recall that we previously cautioned against interpreting a strong correlation between two variables as an indication that changes in one variable cause changes in the other variable. In order to study whether or not changes in a variable  $X$  cause changes in another variable  $Y$ , we must perform a more sophisticated statistical analysis than simply finding a correlation. To establish that a causal effect exists, it is often necessary to be able to control the values of an explanatory variable  $X$ . For instance, when predicting reaction time from drug dosage with the Dosage and Reaction Time Data of Table 10-2, it seems obvious that the dosages were carefully selected by the experimenter. However, when predicting  $Y =$  “yearly income” from  $X =$  “age” with the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, the ages in the SURVEY DATA look random. With the Dosage and Reaction Time Data, the experimenter had control over the values of  $X$  (dosage), whereas with the SURVEY DATA, the experimenter had no control over the ages of the individuals in the study.

**Self-Test Problem 11-1.** In 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. Suppose there is interest in the prediction of grip strength from age.

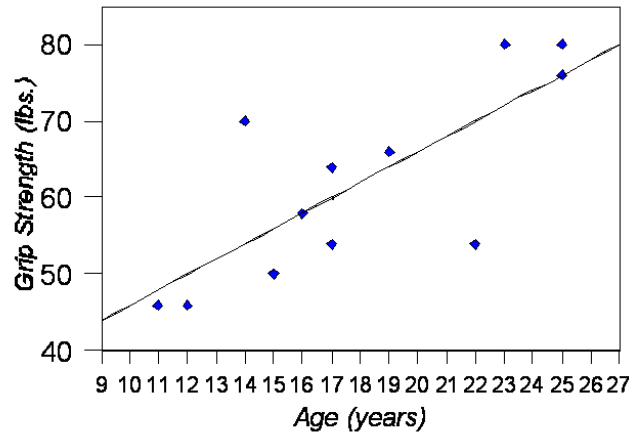
- Identify the response variable  $Y$  and the explanatory variable  $X$ .
- In Self-Test Problem 10-1, it was found that  $\bar{x} = 18$ ,  $\bar{y} = 62$ ,  $s_x^2 = 23.273$ ,  $s_y^2 = 157.091$ , and  $r = +0.7698$ . Find the equation of the least squares line, and write a one sentence interpretation of the slope of the least squares line.
- Use the least squares line to predict the grip strength of a 20-year-old right-handed male.
- Use the least squares line to estimate the average grip strength of 15-year-old right-handed males.
- Give an example of an age for which using this least squares line to predict grip strength would be inappropriate.
- On average, what is the change in grip strength with a five-year increase in age?
- Suppose you are told that a particular right-handed male has a grip strength of 72 lbs.; use the least squares line to estimate this male's age.
- Construct a scatter plot of the data, and graph the least squares line on the scatter plot.
- Do the values for the explanatory variable in the data look like they were controlled by the experimenter, or do they look random?

**Self-Test Problem 11-2.** Suppose that time series data is recorded from the monthly sales of ice cream sales at a particular ice cream parlor. Explain why using a least squares line to predict sales for future months would probably not be very accurate.

#### Answers to Self-Test Problems

- 11-1** (a) Grip strength is the response variable and age is the explanatory variable. (b) The least squares line can be written as  $grp = 26 + 2(age)$ , where  $grp$  represents grip strength in lbs. For each increase of one year in age, grip strength increases on average by about 2 lbs. (c) The predicted grip strength of a 20-year-old right-handed male is 66 lbs. (d) The estimated average grip strength of 15-year-old right-handed males is 56 lbs. (e) It is not appropriate to use this least squares line to predict grip strength for any age outside 11 to 25 years (the age range in the data). (f) For each increase of five years in age, grip increases on average by about  $(5 \times 2)$  10 lbs. (g) The estimated age is about 23 years. (h) See Figure 11-4. (i) The values for the explanatory variable age in the data look random.
- 11-2** Ice cream sales are not likely to change in a linear fashion from month to month, since sales will tend to be higher in summer months and lower in winter months, resulting in a nonlinear cycle within each year.

**Figure 11-4**  
**Scatterplot & Least Squares Line for Table 10-4**



### Summary

A *response variable* or *dependent variable* is one we predict from one or more other variables; an *explanatory variable* or *independent variable* is one from which predictions are made. *Regression* refers to the prediction of one quantitative response variable from one or more quantitative explanatory variables. *Simple linear regression* refers to the prediction of one response variable  $Y$  from one explanatory variable  $X$  using the equation of a straight line written in the form  $Y = a + bX$ . In this equation,  $a$  is called the *intercept* of the line, and  $b$  is called the *slope* of the line. The intercept is the value of  $Y$  when  $X = 0$ ; the slope is the change in  $Y$  whenever  $X$  is increased by one unit. To graph a line, we only need to plot and connect any two distinct points on the line.

One popular method to find the equation of a line which comes closest to all the data points is to use the *least squares line*, which is the line minimizing the squared vertical distances between the data points and the line. The slope of the least squares line can be obtained from

$$b = r \frac{s_Y}{s_X} .$$

Once the slope of the least squares line is available, the intercept can be obtained from

$$a = \bar{y} - b\bar{x} .$$

We may use the least squares line to predict  $Y$  with a given value of  $X$  or to estimate the average value of  $Y$  for a given value of  $X$ , as long as the given value of  $X$  is within the range of the observed data.

Data observed over a sequence of time periods is called a *time series*. Using a least squares line with time series data to make predictions for the future are usually not very accurate, because quantitative variables almost always change over time in a nonlinear fashion.