18 Introduction to Regression Analysis

Often a public or nonprofit manager wants to know whether two interval-level variables are related. Interval-level data are variables that have a well-defined (equal) interval or unit of measurement, such as money (for example, dollars), time (for example, years), distance (for example, miles), or countable quantities or occasions (for example, number of volunteers or number of computer mouse clicks needed to make a donation on a nonprofit Website). In general, an analyst should not use ordinal- or nominal-level techniques, such as those discussed in Chapters 15 through17, on interval-level data. Treating interval information as ordinal loses much of the information that the data contain. Collapsing interval data to present simple tables can be useful for simple reports and memoranda. However, an analyst should not take interval data (such as number of cars, revenue, highway speeds, hours volunteered, money donated, grant applications submitted, or crime rates) and collapse them into categories for analysis purposes.

A variety of public and nonprofit management problems can be interpreted as relationships between two interval variables. For example, the director of the highway patrol might want to know whether the average speed of motorists on a stretch of highway is related to the number of patrol cars on that stretch of highway. Knowing this information would allow the director to decide rationally whether or not to increase the number of patrol cars. In other situations, the public manager might want to know the relationship between two variables for prediction purposes. For example, a northeastern state is considering a sales tax on beer and would like to know how much revenue the tax would raise in the state. An analyst's strategy might be to see whether a relationship exists between a state's population and its tax revenues from beer sales. If a relationship is found, the analyst could then use the state's population to predict its potential revenue from a beer sales tax. Similarly, Habitat for Humanity might be interested in knowing if weather conditions (measured by inches of rainfall and so on) affect the number of construction volunteers.

This chapter presents an introduction to simple linear regression, a statistical technique to determine the relationship between two interval-level variables. Subsequent chapters build on this foundation.

Relationships between Variables

Relationships between two variables can be classified in two ways: as causal or predictive *and* as functional or statistical. In our first example, the relationship between police cars on the road and motorists' average speed, we have a causal relationship. The implicit hypothesis is that increasing the number of patrol cars on the road will reduce average speeds. In the beer sales tax example, a state's population will predict, or determine, tax revenues from beer sales. The variable that is predicted, or is caused, is referred to as the *dependent* variable (this variable is usually called Y). The variable that is used to predict, or is the cause of, change in another variable is referred to as the *independent* variable (this variable is usually called X).

In the following examples, determine which variable is the dependent variable and which is the independent variable:

A police chief believes that increasing expenditures for police will reduce crime.

independent variable _____

dependent variable _____

A librarian believes that circulation is related to advertising.

independent variable _____

dependent variable _____

MPA candidates who complete the nonprofit concentration perform better as summer interns in nonprofit agencies.

independent variable _____

dependent variable _____

The number of volunteers is affected by the weather.

independent variable _____

dependent variable _____

If you said the dependent variables were crime, circulation, good performance, and number of volunteers, congratulations.

Relationships may also be functional or statistical. A **functional relationship** is a relationship in which one variable (Y) is a direct function of another (X). For example, Russell Thomas, the longtime head of the city motor pool, believes that there is some type of relationship between the number of cars he sends over to Marquette's Tune-Up Shop for tune-ups and the amount of the bill that he receives from Marquette's. Russell finds the information in Table 18.1 for the last five transactions with Marquette's.

Russell knows the first step in determining whether two variables are related is to graph the two variables. When graphing two variables, the independent variable (X) is always graphed along the bottom horizontally, and the dependent variable (Y) is always graphed along the side vertically. See Figure 18.1.

On the axes presented in Figure 18.2, graph the points representing the two variables.

Table 18.1 Data from Marquette's

Number of Cars	Amount of Bill
2	$ 192
1	96
5	480
4	384
2	192

Figure 18.1 The Horizontal Axis Is for *X*

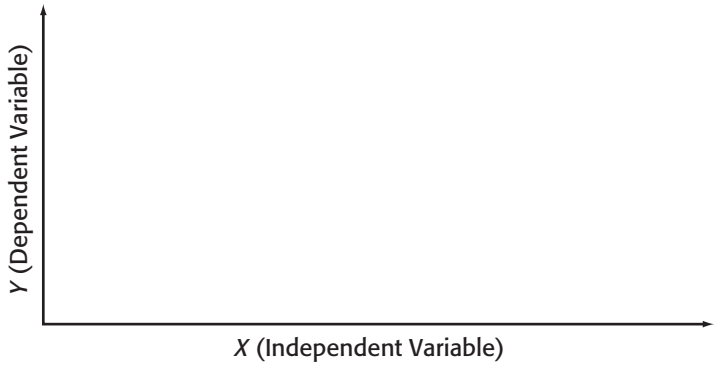
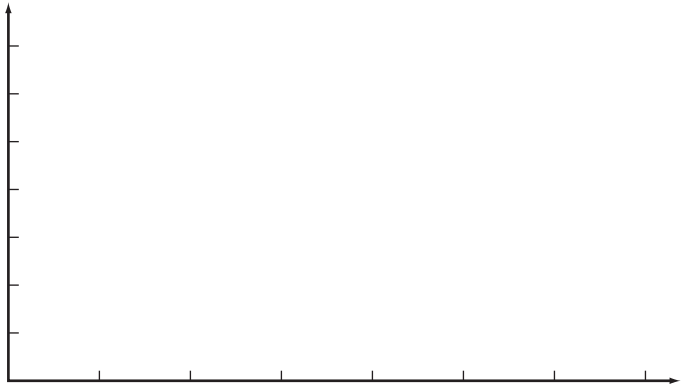


Figure 18.2 Graph Data Here



If you look carefully at the points you graphed in Figure 18.2, you will see that they fall without any deviation along a single line. This is a characteristic of a functional relationship. If someone knows the value of the independent variable, the value of the dependent variable can be predicted exactly. In the preceding example, Marquette's charges the city $96 to tune a car, so the bill is simply $96 times the number of cars.

Unfortunately, very few of the important relationships that public or nonprofit managers must consider are functional. Most relationships are statistical. In a **statistical relationship**, knowing the value of the independent variable lets us estimate a value for the dependent variable, but the estimate is not exact. One process of determining the exact nature of a statistical relationship is called *regression.* We will illustrate how regression can be used to describe relationships with an example.

The Normal, Oklahoma, traffic commissioner believes that the average speed of motorists along Highway 35 within the city limits is related to the number of police cars patrolling that stretch. Average speed is measured by a stationary, unmanned radar gun. The experiment spans 2 months, with measurements taken daily. For a sample of 5 days, the results of the commissioner's experiment are as shown in Table 18.2.

The first step in determining whether a relationship exists is to plot the data on a graph. Plot the given data on the graph in Figure 18.2. After the data are plotted, the analyst can eyeball the data to see whether there is a relationship between the number of cars and the average speed. The graph of the data is shown in Figure 18.3.

Clearly, the graph shows a relationship between the number of police cars on this highway and the motorists' average speed: The more cars on the highway, the lower the average speed. This is termed a negative relationship because the dependent variable (speed) decreases as the independent variable increases. A negative relationship could be stated equally well that the dependent variable increases as the independent variable decreases (see Chapter 3).

Our hypothetical situation in which state population is compared with sales tax revenues from beer sales illustrates a positive relationship (see Table 18.3).

Table 18.2 Commissioner's Data

Number of Police Cars	Average Speed of Motorists
3	64
1	71
4	61
5	58
7	56

Figure 18.3 Graph of Traffic Speed Data

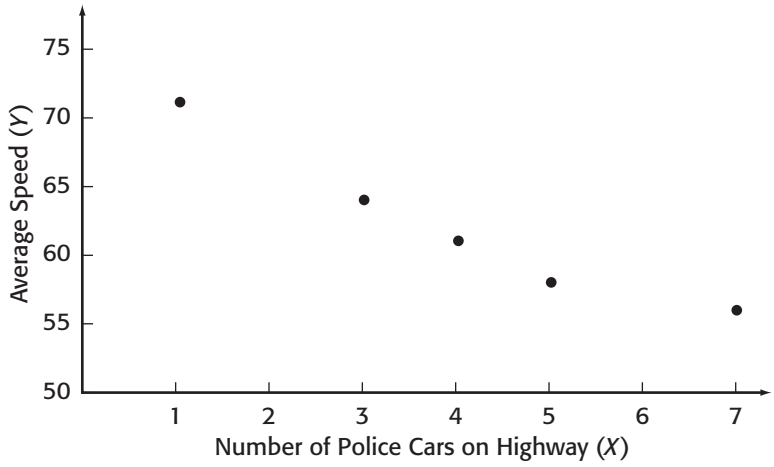


Table 18.3 Relationship between Tax Revenues and Population

State	Population (Millions)	Beer Revenue (Millions)
Pennsylvania	12.4	146
Tennessee	6.1	85
Nevada	2.4	21
Louisiana	4.3	47
North Carolina	9.5	115
Virginia	7.6	90

The graph of these data is shown in Figure 18.4. From the graph, we see that as states' populations increase, so do the states' sales tax revenues from beer sales. Because both variables increase (or decrease) at the same time, the relationship is positive (see Chapter 3).

In many cases (far too many, for most managers), no relationship exists between the two variables. In Table 18.4, the number of police cars patrolling the streets of Normal, Oklahoma, is contrasted with the number of arrests for indecent exposure in Kansas City, Missouri.

A note of explanation is in order. On Wednesday, a regional public administration conference opened in Kansas City. Suspects are held for 24 hours, which also explains the Friday figures. Most of the Saturday incidents occurred at the airport.

Figure 18.4 Relationship between Population and Tax Revenue

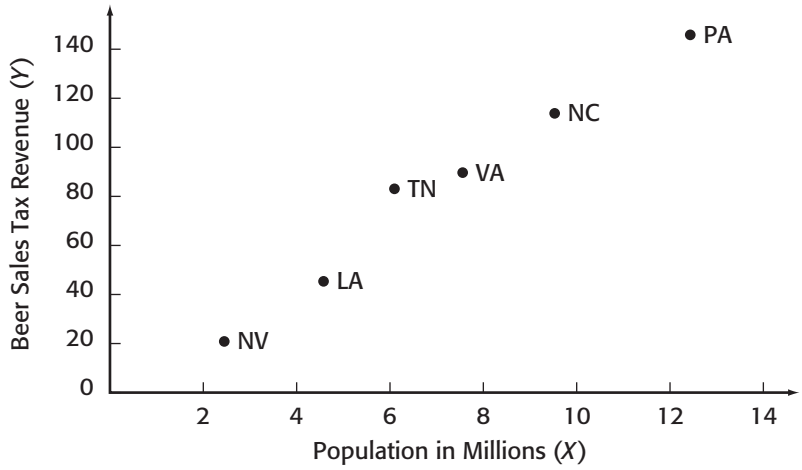


Table 18.4 Patrol Cars and Number of Arrests

Day	Cars on Patrol in Normal	Arrests for Indecent Exposure in Kansas City
Monday	2	27
Tuesday	3	12
Wednesday	3	57
Thursday	7	28
Friday	1	66
Saturday	6	60

The data are graphed in Figure 18.5. Clearly, no relationship exists between the number of police cars patrolling the streets of Normal and arrests for indecent exposure in Kansas City.

Ode to Eyeballing

When an analyst has only a few data points, the relationship between two variables can be determined visually. When the data sets become fairly large, however, eyeballing a relationship is extremely inaccurate. Statistics are needed that summarize the relationship between two variables. One variable, of course, can be summarized by a set of single figures—say, the mean and the standard deviation. *The relationship between two variables can be summarized by a line.*

Figure 18.5 Relationship between Number of Patrol Cars and Indecent Exposure Arrests

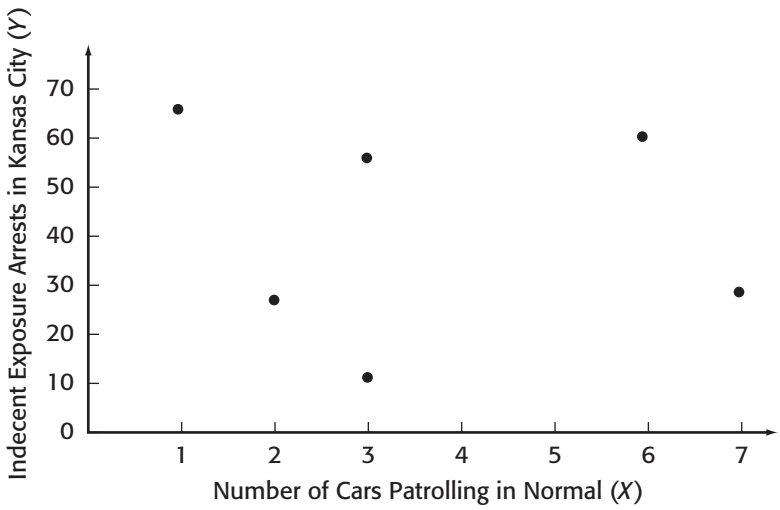
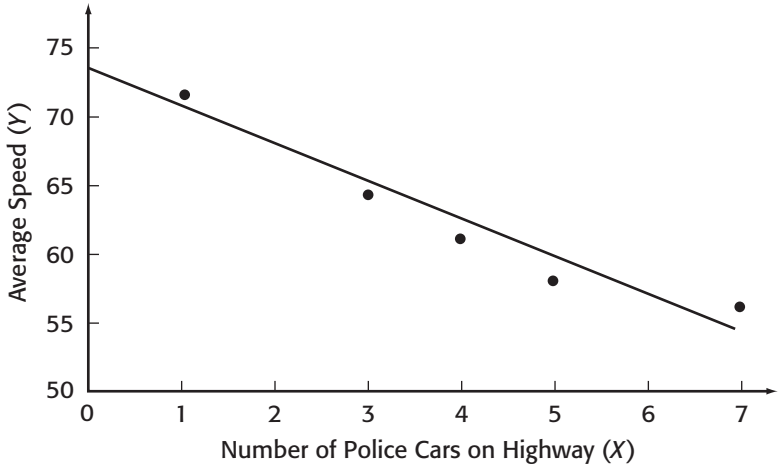


Figure 18.6 Straight Line for Data of Table 18.2



For our example of cars patrolling a stretch of Highway 35 and the average speed of traffic on that portion of highway, a straight line can be drawn that represents the relationship between the data (see Figure 18.6). The line generally follows the pattern of the data, sloping downward and to the right.

Any line can be described by two numbers, and the line describing the relationship between two variables is no exception. Lines *a*, *b*, and *c* in

Figure 18.7 The Lines Differ in Their Slopes

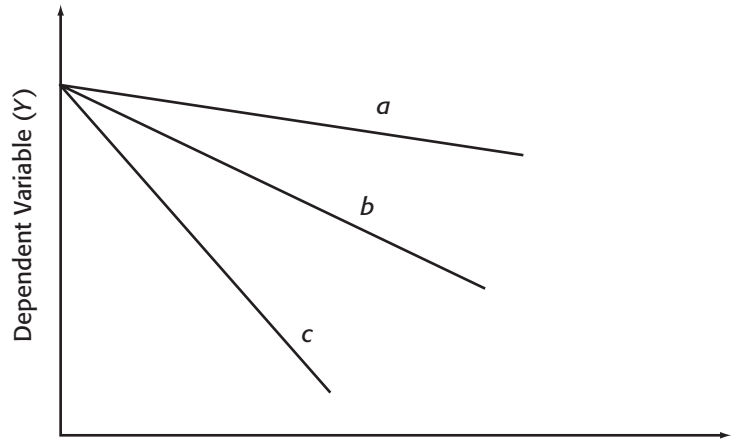


Figure 18.7 differ from each other in terms of how steeply the lines slant from left to right.

The slant of a line is referred to as its slope. The **slope** of any line is defined to be how much the line rises or falls relative to the distance it travels horizontally. Symbolically,

$$\beta = \frac{\Delta Y}{\Delta X}$$

where β (Greek letter beta) is the slope of a line, ΔY (Greek letter delta) is the change in the Y (dependent) variable, and ΔX is the change in the X (independent) variable.

Another way to express this formula is to say that the slope of a line is equal to the ratio of the change in Y for a given change in X (rise over run, for you geometry buffs). The graph in Figure 18.8 shows the slopes of several hypothetical lines.

The second number used to describe a line is the point where the line intersects the *Y-axis* (called the **intercept**). The graph in Figure 18.9 shows several lines with the same slopes but with different intercepts. The intercept, referred to as α (Greek letter alpha) by statisticians, is the value of the dependent variable when the independent variable is equal to zero.

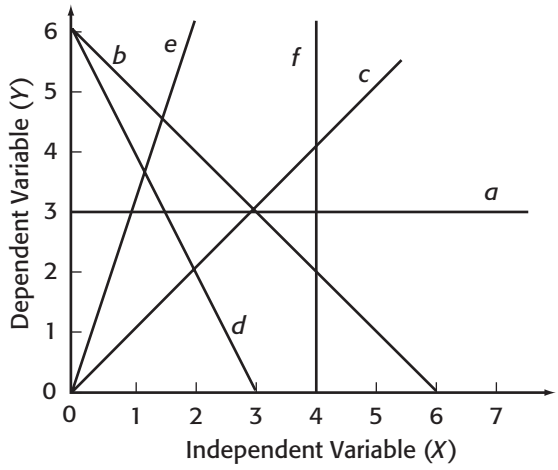
Any line can be fully described by its slope and its intercept:

$$Y = \alpha + \beta X$$

A line describing the relationship between two variables is represented by

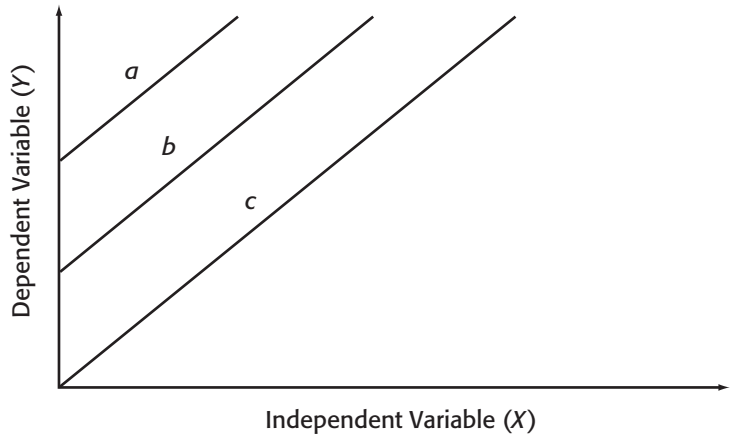
$$\hat{Y} = \alpha + \beta X$$

Figure 18.8 **Several Different Slopes**



Line	Slope
a	0
b	−1
c	+1
d	−2
e	+3
f	∞

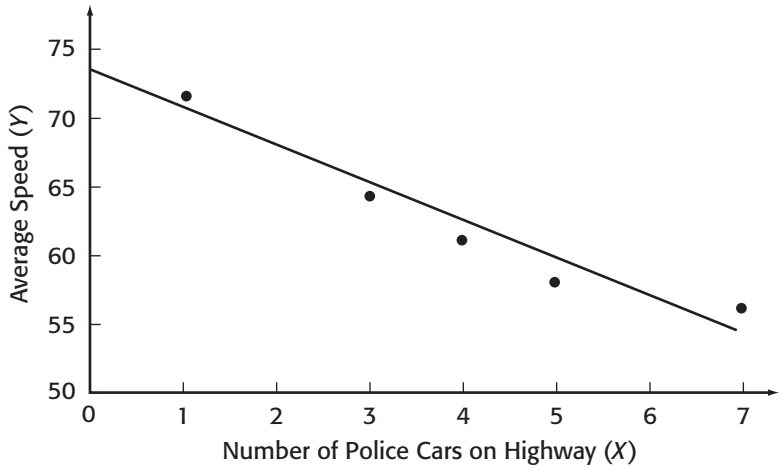
Figure 18.9 **The Lines Have the Same Slope but Different Intercepts**



$\hat{Y}$ *is* a statistician's symbol for the **predicted value** of Y (called "Y hat"). $\hat{Y}$ for any value of X is a function of the intercept (α) and the slope (β), and it may or may not be equal to the actual value of Y.

To illustrate, let us return to our example of traffic speeds and patrol cars. The line drawn through the data in Figure 18.10 represents the relationship between the two variables. If we had only the line, what could we say about the expected average speed if three cars were on the road? $\hat{Y}$, the expected speed, is 65 miles

Figure 18.10 Straight Line for Data of Table 18.2



per hour. (To find this number, draw a line straight up from the three-cars point to the relationship line. From the point where your line touches the relationship line, draw a line parallel to the X-axis to the traffic speed Y-axis line. Your line should touch the Y-axis at 65. This value is $\hat{Y}$.)

Note that the actual value of Y on the one day when three cars were on the road is 64 miles per hour. This fact illustrates the following:

$$\text{predicted value of } Y = \text{real value of } Y + \text{some error}$$

$$\hat{Y}_i = Y_i + e_i$$

or

$$e_i = (\hat{Y}_i - Y_i)$$

Another way of expressing these equations is to say that every value of Y is equal to some predicted value of Y based on X plus some error.*

Linear Regression

The pitfall of just drawing in a line to summarize a relationship is that numerous lines will look as if they summarize the relationship between two variables equally well. Statisticians have agreed that the best line to use to describe a relationship is

*We assume that error can be either negative or positive, so it does not matter whether error is added to $\hat{Y}_i$ (or Y_i) or subtracted from $\hat{Y}_i$ (or $\hat{Y}_i$).

the line that minimizes the squared errors about it—that is, the line that makes the sum of all $(\hat{Y}_i - Y_i)^2$ the smallest possible number. This form of **regression** (or fitting a line to data) is called **ordinary least squares**, or you may call it just **linear regression**.

Linear regression using the principle of minimizing squared errors allows us to find one value of α and one value of β so that a unique regression line of the form $\hat{Y} = \alpha + \beta X$ can be determined. The calculations necessary to find α and β will be illustrated with an example.

Before considering the example, we should note that regression is a technique that is often used to make inferences from a sample to a population. Similar to other situations of inference, slightly different symbols are used. For a population regression, a line is denoted as

$$Y = \alpha + \beta X + \epsilon$$

Sometimes, rather than simply using *Y,* statisticians use the symbol $\mu_{y|x}$, which stands for the mean of *y* given *x,* or the mean value of *y* given the use of *x* to try to predict *y.* The population intercept is denoted α and the slope β. The symbol ϵ represents an error term meant to capture any errors in prediction. In statistics we rarely work with populations but rather with samples. In that case, the symbols are

$$Y = a + bX + e$$

The sample intercept is represented by the symbol *a*, the slope by *b*, and the error term by *e*. Similar to the case with means, the best estimate of the population slope and intercept are the sample slope and intercept.

Through some heavy mathematics based on calculus, statisticians have found that the formula for the slope *b* is as follows:

$$b = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2}$$

Taken a piece at a time, this formula is not as intimidating as it looks. We will use the data on police cars and average speed to calculate *b* (see Table 18.5).

Table 18.5 Relationship between Police Cars and Average Speed

Number of Police Cars (*X*)	Average Speed (*Y*)
3	64
1	71
4	61
5	58
7	56

Step 1: Calculate the mean for both the dependent variable (Y) and the independent variable (X). If you have forgotten how to calculate a mean, please refer to Chapter 5. The mean for Y is 62, and the mean for X is 4.

Step 2: Subtract the mean of the dependent variable from each value of the dependent variable, yielding $(Y_i - \overline{Y})$. Do the same for the independent variable, yielding $(X_i - \overline{X})$.

$X_i - \overline{X}$	$Y_i - \overline{Y}$
$3 - 4 = -1$	$64 - 62 = 2$
$1 - -4 = -3$	$71 - 62 = 9$
$4 - 4 = 0$	$61 - 62 = -1$
$5 - 4 = 1$	$58 - 62 = -4$
$7 - 4 = 3$	$56 - 62 = -6$

Step 3: Multiply $(Y_i - \overline{Y})$ times $(X_i - \overline{X})$. That is, multiply the value that you get when you subtract the mean Y from Y_i by the value you get when you subtract the mean X from X_i.

$(X_i - \overline{X}) \times (Y_i - \overline{Y})$				
-1	$\times$	2	$=$	-2
-3	$\times$	9	$=$	-27
0	$\times$	-1	$=$	0
1	$\times$	-4	$=$	-4
3	$\times$	-6	$=$	-18

Step 4: Sum all the values of $(Y_i - \overline{Y})(X_i - \overline{X})$. You should get a sum of -51. This is the numerator of the formula for b.

Step 5: Use the $(X_i - \overline{X})$ column in Step 3, and square each of the values found in the column.

$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
–1	1
–3	9
0	0
1	1
3	9

Step 6: Sum the squared values of $(X_i - \overline{X})$. The answer is 20.

Step 7: Divide $\Sigma(Y_i - \overline{Y})(X_i - \overline{X})$ or –51, by $\Sigma(X_i - \overline{X})^2$, or 20. This number -2.55 is the slope.

The intercept is much easier to calculate. Statisticians have discovered that

$$\alpha = \mu_y - \beta\mu_x$$

or

$$a = \overline{Y} - b\overline{X}$$

Substituting in the values of 62, −2.55, and 4 for $\overline{Y}$, b, and $\overline{X}$, respectively, we find

$$a = 62-(-2.55) \times 4 = 62-(-10.2) = 62 + 10.2 = 72.2$$

The regression equation that describes the relationship between the number of patrol cars on a stretch of Highway 35 and the average speed of motorists on that stretch of highway is

$$\hat{Y} = 72.2 - 2.55X$$

All sample regressions are of the general form

$$\hat{Y} = a + bX$$

In English, the predicted value of Y ($\hat{Y}$) is equal to X times the slope (b) plus the intercept (a). The slope and the intercept can be positive or negative. In the present example the intercept is positive (72.2) and indicates that if there are no patrol cars on a stretch of Highway 35, the average expected speed of motorists is 72.2. The slope is negative (−2.55) and indicates that for every patrol car on the road, the average speed of motorists is expected to decrease by 2.55.

Some Applications

The regression equation provides a wealth of information. Suppose the traffic commissioner wants to know the estimated average speed of traffic if six patrol cars are placed on duty. Another way of stating this question is, What is the value of $\hat{Y}$ (the estimated average speed) if the value of X (the number of cars) is 6? Using the formula

$$\hat{Y} = 72.2 - 2.55X$$

substitute 6 for X to obtain

$$\hat{Y} = 72.2 - 2.55 \times 6 = 72.2 - 15.3 = 56.9$$

The best estimate of the average speed for all cars on a stretch of Highway 35 is 56.9 if six patrol cars are placed on that stretch.

How much would the mean speed for all cars decrease if one additional patrol car were added? The answer is 2.55 miles per hour. The **regression coefficient** is the ratio of change in $\hat{Y}$ to the change in X. Where the change in X is 1 (car), the change in $\hat{Y}$ is −2.55 (miles per hour). In a management situation, this is how the slope should be interpreted. It is how much $\hat{Y}$ will change if X is changed (increased) 1 unit. Remember, though, that the regression line gives estimates, and there is error (e) in predicting actual Y scores.

What would the average speed be if no patrol cars were on the road? Substituting 0 into the regression equation for X, we find

$$\hat{Y} = 72.2 - 2.55X = 72.2 - 2.55(0) = 72.2$$

When X is 0, the value of $\hat{Y}$ *is* 72.2, or the intercept. The intercept is defined as the value of $\hat{Y}$ when X is equal to zero.

An Example

Most analysts rely on computer programs to calculate regression equations. We expect that you will do so, too. However, just for practice, we ask you to calculate the regression equation for the population and beer sales tax example. Recall that for six states, the data are as given in Table 18.6. In the space provided, calculate the slope and the intercept of the regression line.

Table 18.6 Relationship between Tax Revenues and Population

Population, X (millions)	Beer Revenue, Y (millions)
12.4	146
6.1	85
2.4	21
4.3	47
9.5	115
7.6	90

$\overline{X} =$ _____ $\overline{Y} =$ _____

$X_i - \overline{X}$	$(X_i - \overline{X})^2$	$Y_i - \overline{Y}$
12.4 − _____ = _____	_____	146 − _____ = _____
6.1 − _____ = _____	_____	85 − _____ = _____
2.4 − _____ = _____	_____	21 − _____ = _____
4.3 − _____ = _____	_____	47 − _____ = _____
9.5 − _____ = _____	_____	115 − _____ = _____
7.6 − _____ = _____	_____	90 − _____ = _____

$(X_i - \overline{X}) \times (Y_i - \overline{Y})$		
_____	_____	
_____	× _____	= _____
_____	× _____	= _____
_____	× _____	= _____
_____	× _____	= _____
_____	× _____	= _____
_____	× _____	= _____

$$\Sigma(X_i - \overline{X})(Y_i - \overline{Y}) = \_\_\_\_\_$$
$$\Sigma(X_i - \overline{X})^2 = \_\_\_\_\_$$
$$\beta = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2} = \_\_\_\_\_$$
$$\alpha = \overline{Y} - \beta\overline{X} = \_\_\_\_\_$$

The answer to this exercise is presented at the end of the chapter, following the problems.

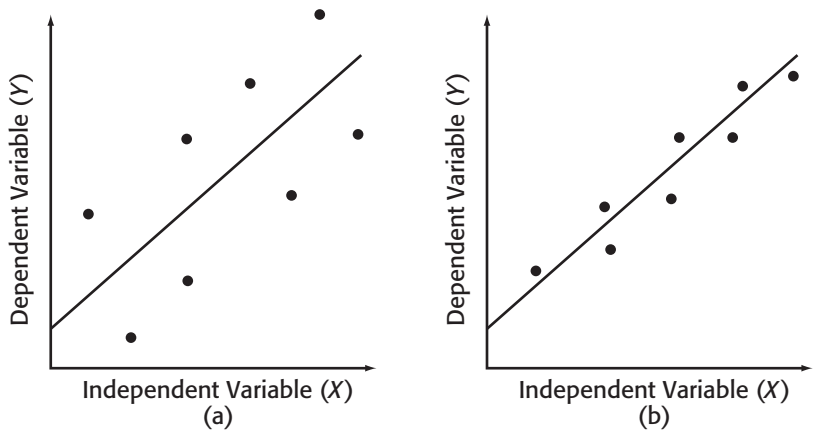
What would be the best estimate of Colorado's beer sales tax revenue if it had such a tax and had 5.5 million people?

Measures of Goodness of Fit

Any relationship between two variables can be summarized by linear regression. A regression line per se, however, does not tell us how well the regression line summarizes the data. To illustrate, the two sets of data in Figure 18.11 can both be summarized with the same regression line. In the graph of part (b), however, the data points cluster closely about the line; in the graph of part (a), the data points are much farther from the line. We can say that the regression line of (b) fits the data better than does the regression line of (a), even though the "best-fitting" regression line for the data in both Figures 18.11(a) and 18.11(b) is the same.

The distance a point is from the regression line is referred to as **error**. Recall that the regression line gives the value of $\hat{Y}_i$, whereas the data point represents

Figure 18.11 **Differences in Goodness of Fit**



Y_i. In the following sections, we will discuss various ways that statisticians have devised to measure the goodness of fit of the regression line to the data. All these methods are based on the error.

Error in the context of regression analysis means unexplained variance (i.e., spread or dispersion in the values of Y that cannot be accounted for or "explained" by changes in X). A regression equation will almost always have some error, so trying to eliminate error entirely is not realistic. What are some of the causes of error in a regression equation?

First, a single independent variable rarely accounts for all of the variation in a dependent variable. For example, the unemployment rate might explain a substantial portion of the variation in demand for services at a local food bank, but other variables, including higher retail food prices and climate (such as cold weather), might also account for some of the variation in demand. Data values will not fall perfectly along a regression line if the independent variable explains only some of the variation in the dependent variable. This is why analysts often perform regression with several independent variables ("multiple regression"), a subject covered in Chapter 21.

Second, individual cases within our data do not always conform to the overall relationships we find when using regression analysis. For example, if most drivers slow down when the number of police cars on patrol goes up, a small number of drivers may throw caution to the wind and continue traveling at a high rate of speed. Even when a regression equation reveals a relationship between an independent and a dependent variable, deviations from the general pattern for individual cases are almost always inevitable.

Third, almost always our measurements of important variables in public and nonprofit administration contain error. Measuring such concepts as organization effectiveness and mission salience is very difficult. The presence of error in measurement contributes to errors in the regression equation. Fourth, variables may be related but not in a linear fashion; we look at curvilinear regression in Chapter 21. Estimating a linear regression when the relationship is curvilinear will generate substantial error.

The Standard Error of the Estimate

Statisticians commonly use three measures of fit in regression analysis. The first, called the **residual variation**, or the variance of the estimate, is equal to the sum of the squared error divided by $n - 2$. Symbolically,

$$S^2_{y|x} = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - 2}$$

$S^2_{y|x}$ is called the residual variance of Y given X.

Another way of stating what $S^2_{y|x}$ represents is to call it the average squared error of the regression estimates. Although the residual variation is rarely used as a measure of fit, its square root $(S_{y|x})$ is. This measure, called the **standard error**

of the estimate (or sometimes root mean square error), is an estimate of the variation in $\hat{Y}$, the predicted value of Y. The standard error of the estimate can be used to place confidence intervals around an estimate that is based on a regression equation.

To illustrate the utility of this measure of fit of the regression line, we need to calculate the residual variance for a set of data. We will use the police cars and speed data of Table 18.5. In the worked-out example presented earlier, we found that the number of patrol cars on the highway was related to the average speed of all cars and that

$$\overline{X} = 4 \quad \overline{Y} = 62 \quad \hat{Y} = 72.2 - 2.55X$$

To calculate the residual variation, follow these steps:

Step 1: Using the values of X and the regression equation, calculate a $\hat{Y}$ value for every X value.

$X \times b$	$Xb + a = \hat{Y}$
3×-2.55	$-7.65 + 72.2 = 64.6$
1×-2.55	$-2.55 + 72.2 = 69.7$
4×-2.55	$-10.2 + 72.2 = 62.0$
5×-2.55	$-12.75 + 72.2 = 59.5$
7×-2.55	$-17.85 + 72.2 = 54.4$

Step 2: Using the $\hat{Y}$ values and the Y values, calculate the total error for each value of Y.

$Y - \hat{Y}$
$64 - 64.6 = -0.6$
$71 - 69.7 = \ \ 1.3$
$61 - 62.0 = -1.0$
$58 - 59.5 = -1.5$
$56 - 54.4 = -1.6$

Step 3: Square the errors found in Step 2, and then sum these squares.

$(Y - \hat{Y})^2$	
.36	$\Sigma(Y - \hat{Y})^2 = 7.86$
1.69	
1.00	
2.25	
2.56	

Step 4: Divide the sum of the squared errors by $n - 2$ to find the residual variation (or average squared error).

$$S^2_{y|x} = \frac{7.86}{3} = 2.62$$

Step 5: Take the square root of this number to find the standard error of the estimate.

$$S_{y|x} = \sqrt{2.62} = 1.62$$

The standard error of the estimate may be interpreted as the amount of error that one makes when predicting a value of Y for a given value of X. The standard error of the estimate, however, applies only to predicting error at the exact middle of the distribution [that is, where X is equal to the mean of X; for a good explanation of why this is true, see Gujarati and Porter (2008), Chapter 5]. To predict a confidence limit around any single point, the following transformation of the standard error of the estimate is used:

$$S_{y|x} \times \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)S_x^2}}$$

where X_0 is the value of X being predicted, $\overline{X}$ is the mean of X, S_x is the standard deviation of X, and n is the sample size. Confidence limits can be placed around any predicted value of Y by using the following formula:

$$Y \pm t \times S_{y|x}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)S_x^2}}$$

where *t is* the t score associated with whatever confidence limits are desired.

Suppose the highway commissioner wanted to predict the average speed of all cars when three patrol cars were on the road. Using the regression equation, he would find

$$Y = 72.2 - 2.55(3) = 72.2 - 7.65 = 64.55$$

This estimate of the average speed is not exact; it can be in error by a certain amount. To put 90% confidence limits around this estimate (64.55), we need to know the t score associated with 90% confidence limits. Simple (bivariate) regressions have $n - 2$ degrees of freedom, so we check Table 3 in the Statistical Tables for the .05 level (.05 + .05 = .10) with 3 degrees of freedom and find the value 2.35. Because we already know the value of x (it is 3), all we need is the standard deviation of x to do the calculations. That value is 2.23 (you need not believe us—you can calculate this yourself from the raw data). Thus the formula reduces to

$$64.55 \pm (2.35 \times 1.62) \times \sqrt{1.00 + (1/5) + [(3-4)^2/(5-1)(2.23)^2]}$$
$$64.55 \pm 3.81 \times \sqrt{1.00 + .2 + .05}$$
$$64.55 \pm 3.81 \times \sqrt{1.25}$$
$$64.55 \pm 3.81 \times 1.12$$
$$64.55 \pm 4.27$$
$$60.28 \text{ to } 68.82$$

We can be 90% sure that the mean speed of all cars (when three patrol cars are on the road) is between 60.28 and 68.82 miles per hour.

The Coefficient of Determination

The second goodness of fit measure adjusts for the total variation in *Y*. This measure, the coefficient of determination, is the ratio of the explained variation to the total variation in *Y*. Explained variation is nothing more than the total variation in the dependent variable minus the error variation. Statisticians have defined the ratio of explained to unexplained variation as equal to

$$r^2 = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2}$$

This measure is called the **coefficient of determination**, or r^2. In bivariate (one independent variable) regression, we use r^2. In multiple (more than one independent variable) regression, the symbol R^2 is used (see Chapter 21). The coefficient of determination ranges from zero (the data do not fit the line at all) to one (the data fit the line perfectly).

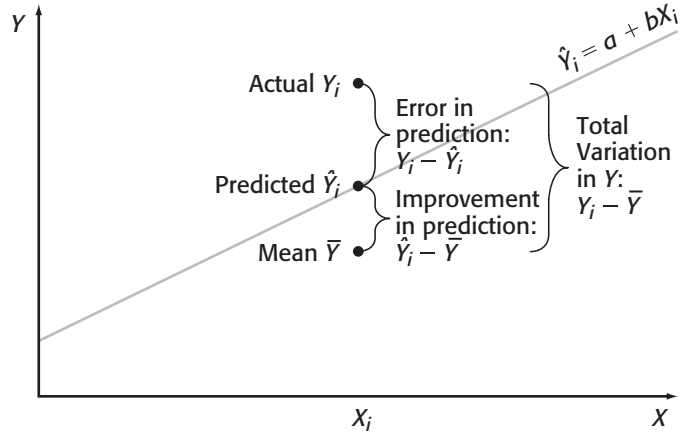
The best way to interpret the coefficient of determination is as follows. If someone wanted you to guess the next value of *Y* but gave you no information, your best guess as to what *Y* is would be $\overline{Y}$, the mean. The amount of error in this guess would be $(Y_j - \overline{Y})$. The total squared error for several guesses of *Y* would be $\Sigma(Y_i - \overline{Y})^2$. If someone asked you to guess the next value of *Y* and gave you both the corresponding value of *X* and a regression equation, your best guess as to the value of Y_j would be $\hat{Y}_j$. How much of an improvement would $\hat{Y}$ be over just guessing the mean? Obviously, it is $(\hat{Y}_j - \overline{Y})$, or the difference between the estimated value of Y_j (or $\hat{Y}_j$) and the mean. The total improvement in squared error for several guesses would be $\Sigma(Y_i - \overline{Y})^2$. As you can tell, the coefficient of determination is the ratio of the reduction of the error by using the regression line to the total error by guessing the mean. Figure 18.12 shows the improvement in prediction achieved by using $\hat{Y}_i$ rather than $\overline{Y}$ to predict Y_j. The improvement in prediction is essential to calculating the coefficient of determination.

To calculate the coefficient of determination, follow these steps:

Step 1: Using the regression equation and each value of *X*, estimate a predicted value of *Y* ($\hat{Y}$). Such estimates were just made in the previous example; they are

X	$\hat{Y}$
3	64.6
1	69.7
4	62.0
5	59.5
7	54.4

Figure 18.12 Coefficient of Determination



Step 2: From each value of $\hat{Y}$, subtract the mean value of Y (in this case, 62), and square these differences.

$\hat{Y} - \bar{Y} = (\hat{Y} - \bar{Y})$	$(\hat{Y} - \bar{Y})^2$
$64.6 - 62 = 2.6$	6.8
$69.7 - 62 = 7.7$	59.3
$62.0 - 62 = .0$	.0
$59.5 - 62 = -2.5$	6.3
$54.4 - 62 = -7.6$	57.8

Step 3: Sum these squared differences to find the numerator of the coefficient of determination. In this case, the answer is 130.2.

Step 4: Subtract the mean value of Y from the individual values of Y, and square these differences.

$Y - \bar{Y} = (Y - \bar{Y})$	$(\hat{Y} - \bar{Y})^2$
$64 - 62 = 2$	4
$71 - 62 = 9$	81
$61 - 62 = -1$	1
$58 - 62 = -4$	16
$56 - 62 = -6$	36

Step 5: Sum these squared differences to get the denominator of the coefficient of determination. The answer is 138.

Step 6: To find the coefficient of determination, divide the value found in Step 3 by the value found in Step 5.

$$r^2 = \frac{130.2}{138} = .94$$

To interpret the coefficient of determination, we can say that the number of patrol cars on a stretch of Highway 35 can explain 94% of the variance in the average speed of cars on that stretch of highway.

The square root of the coefficient of determination is called the **correlation coefficient**, or *r*. The value of *r* ranges from –1.0 for perfect negative correlation to +1.0 for perfect positive correlation. Despite its frequent use in many academic disciplines, the correlation coefficient has no inherent value because it is difficult to interpret. The coefficient of determination is far more useful.

The Standard Error of the Slope

The third measure of goodness of fit is the standard error of the slope. If we took several samples with an independent and a dependent variable and calculated a regression slope (b) for each sample, the sample slopes would vary somewhat. The standard deviation of these slope estimates is called the **standard error of the slope** estimate. The formula for the standard error of the slope estimate is

$$s_b = \frac{S_{y|x}}{\sqrt{\Sigma(X_i - \overline{X})^2}}$$

The standard error of the slope can be used in the same manner as other standard errors—to place a confidence interval around the slope estimate. The standard error of the slope estimate is calculated as follows:

Step 1: Calculate $S_{y|x}$, the standard error of the estimate. You will find that $S_{y|x}$ for the data we have been considering (see Table 18.1) is equal to 1.62.

Step 2: From each value of X, subtract the value of $\overline{X}$, and square these differences.

$X - \overline{X}$	=	$(X - \overline{X})$	$(X - \overline{X})^2$
3 − 4	=	−1	1
1 − 4	=	−3	9
1 − 4	=	0	0
4 − 4	=	1	1
7 − 4	=	3	9

Step 3: Sum all the squared differences: $\Sigma(X_i - \overline{X})^2 = 20$.

Step 4: Take the square root of the number found in Step 3: $\sqrt{20} = 4.47$.

Step 5: Divide the standard error of the estimate (1.62) by the number found in Step 4 to get the standard error of the slope estimate.

$$s_b = \frac{1.62}{4.47} = .36$$

The standard error of the slope estimate can be used just like any other standard error. We can place 90% confidence limits around the slope estimate. The procedure for using the sample slope to place a 90% confidence limit around the slope estimate is

$$\begin{aligned} b &\pm t \times s_b(\text{df} = 3) \\ -2.55 &\pm 2.65 \times .36 \\ -2.55 &\pm .85 \\ -3.40 &\text{ to } -1.70 \end{aligned}$$

We can be 90% sure that the population slope falls between -1.70 and -3.40.

The standard error of the slope can also be used to answer the following question: What is the probability that one could draw a sample with a slope equal to the value of b obtained in a regression equation if the slope in the population equals zero? This is called *testing the statistical significance of the slope.* If $\beta = 0$, then there is no relationship between the variables in the population. If it is probable that the sample was drawn from such a population, we could not reject the null hypothesis that no relationship exists between the independent variable and the dependent variable.

To determine the probability in our example that a sample with a slope of –2.55 could have been drawn from a population where $\beta = 0$, we convert b into a t score by using 0 as the mean and by using the standard error of the slope

$$t = \frac{X - \mu}{s}$$

$$t = \frac{b - \beta}{s_b} = \frac{-2.55 - 0}{.36} = -7.1$$

A t value of 7.1 with 3 degrees of freedom is greater than the value for .005 ($t = 5.841$). The probability that a sample with a slope of -2.55 could have been drawn from a population with a slope of zero is less than .005. If there are no major research design problems (and there appear to be none here), the public or nonprofit manager would be justified in concluding that a relationship exists. (Research design is the subject of Chapter 3.)

Sometimes the entire population is used to calculate a regression line (for example, an analysis based on all 50 U.S. states). In such cases, the preceding exercise of testing for statistical significance does not make theoretical sense

because these procedures assume that only a sample of the data are available. Many analysts test for statistical significance anyway to illustrate that the relationship is not trivial, or that it is very unlikely to occur by chance (i.e., the independent variable and the dependent variable are actually related).

Although you may not immediately see a link between the standard error of the slope and the coefficient of determination, the two are closely related. Think about why this is the case. The coefficient of determination reveals the amount of variation in the dependent variable that is explained by the independent variable. When we test the statistical significance of the slope and are unable to reject the null hypothesis ($\beta = 0$), the amount of variation in the dependent variable explained by the independent variable is typically small. In contrast, when we are able to reject the null hypothesis ($\beta = 0$), the coefficient of determination will be larger, because the independent variable does indeed explain variation in the dependent variable.

To illustrate this point, we will perform a regression using the data in Table 18.4. Recall that when we graphed these data, we found no evidence of a relationship between the number of police cars on patrol in Normal, Oklahoma, and arrests for indecent exposure in Kansas City, Missouri.

$$\hat{Y} = 4.24 - .0138X$$

$$s_b = .0541 \qquad r^2 = .017$$

$$t = \frac{-.0138 - 0}{.0541} = .26$$

The regression equation confirms our initial finding of no relationship between the two variables. The slope coefficient is clearly not statistically significant. A t value of .26 with 4 degrees of freedom fails to exceed the t value associated with alpha at .05 ($t = 2.132$). The r^2 for the model is .017, indicating that the number of police cars on patrol in Normal explains less than 2% of the variation in indecent exposure arrest rates in Kansas City.

Generally speaking, if the independent variable does a poor job of explaining variation in the dependent variable, the r^2 value will also be quite low. A statistically insignificant slope and low r^2 are signs that the independent variable does a poor job of explaining variation in the dependent variable.

Chapter Summary

Regression is a technique that can be used to describe the statistical relationship between two interval variables. This chapter illustrated the use of simple (bivariate) linear regression.

The relationship between two variables can be summarized by a line, and any line can be fully described by its slope and its intercept. The slope of a line is equal to the ratio of the change in Y to a given change in X. The intercept of