19 The Assumptions of Linear Regression

Chapter 18 introduced an important statistical technique, linear regression analysis. Like any statistical procedure, regression analysis has assumptions and limitations. In the real world of work, actions, and decisions, analysts sometimes ignore or overlook these assumptions. They do so at some managerial risk, however. All the uses of regression presented in the previous chapter and later ones become less reliable when any of the assumptions is not met.

In Chapter 18, recall that our highway patrol example yielded the following regression equation and coefficient of determination (r^2):

$$\hat{Y} = 72.2 - 2.55X \quad r^2 = .94$$

where $\hat{Y}$ is the predicted speed of all cars and X is the number of police patrol cars on the road. In our discussion in Chapter 18, we found that the predicted values $\hat{Y}_i$ did not exactly equal the real or actual values of Y_i. According to the coefficient of determination, we accounted for 94% rather than all 100% of the variation in Y with $\hat{Y}$. What sorts of other factors account for average car speed in addition to the number of patrol cars on the road?

We can think of several factors. The weather conditions on any given day can slow traffic. The emergence and filling of potholes affect traffic speed. The number of other cars on the road restricts any one car's speed. The curves and hills and visibility on a stretch of highway affect traffic speed. These factors and others omitted from the regression equation probably account for the difference between Y_i and $\hat{Y}_i$. We can express this situation symbolically as

$$\hat{Y} = \alpha + \beta X + \beta_1(X_1, X_2, X_3, X_4)$$

where X_1, X_2, X_3, X_4 are the other factors, and b_1 is some weight (slope) representing their combined effect on Y. (**Note**: In any particular regression problem we can have more or fewer than four omitted factors.)

To simplify matters, we generally refer to all the other factors as e, or error.

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

That is, the value of Y is equal to some constant (α) plus a slope (β) times X plus some error (e).

We introduce this terminology because most assumptions about linear regression are concerned with the error component. In this chapter, we will discuss the assumptions and limitations of linear regression.

Assumption 1

For any value of X, the errors in predicting Y are normally distributed with a mean of zero.

To illustrate, let us assume that we continue the Normal, Oklahoma, patrol car experiment for an entire year. Every day, between one and seven cars is sent out to patrol the local highway, and the average speed of all cars is measured. At the end of the year, let us assume that the overall regression equation remains the same:

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

By the end of the year, we probably have 50 days when four patrol cars were on the road. The average speed for each of these 50 days is listed in Table 19.1 in a frequency distribution.

Using the midpoint of the frequencies to represent each interval, we can calculate the error for each prediction, because we know that the predicted speed for four patrol cars is 62 miles per hour ($72.2 - 2.55 \times 4 = 62$). The error calculations are given in Table 19.2. The mean error for all 50 cars is 0 (add the last column and divide by 50) and is distributed fairly close to normally.

Table 19.1 Frequency Distribution for Patrol Cars and Average Speeds (in mph)

Average Speed	Number of Days
58.5–59.0	1
59.0–59.5	2
59.5–60.0	2
60.0–60.5	4
60.5–61.0	4
61.0–61.5	6
61.5–62.0	6
62.0–62.5	6
62.5–63.0	6
63.0–63.5	4
63.5–64.0	4
64.0–64.5	2
64.5–65.0	2
65.0–65.5	1
65.5–66.0	0

Table 19.2 Error Calculations

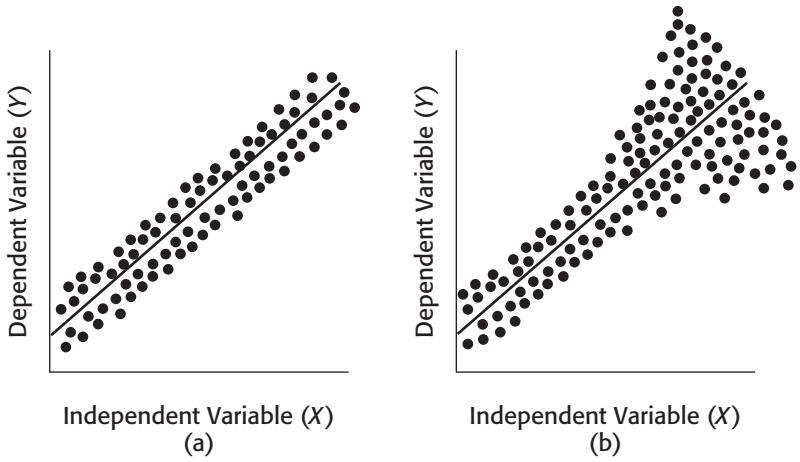
Average Speed	−	Predicted Speed	=	Error	×	Frequency	=	Total Error
58.75	−	62	=	−3.25	×	1	=	−3.25
59.25	−	62	=	−2.75	×	2	=	−5.50
59.75	−	62	=	−2.25	×	2	=	−4.50
60.25	−	62	=	−1.75	×	4	=	−7.00
60.75	−	62	=	−1.25	×	4	=	−5.00
61.25	−	62	=	−.75	×	6	=	−4.50
61.75	−	62	=	−.25	×	6	=	−1.50
62.25	−	62	=	.25	×	6	=	1.50
62.75	−	62	=	.75	×	6	=	4.50
63.25	−	62	=	1.25	×	4	=	5.00
63.75	−	62	=	1.75	×	4	=	7.00
64.25	−	62	=	2.25	×	2	=	4.50
64.75	−	62	=	2.75	×	2	=	5.50
65.25	−	62	=	3.25	×	1	=	3.25

Whenever e has a mean of zero and is normally distributed, statisticians have found that sample slopes (b) have a mean equal to the population slope (β) and are distributed as a t distribution with a standard deviation s_b. When the sample size is fairly large ($\mathbf{N} > 30$), the t distribution resembles the normal distribution, and z scores can be used as estimates of t scores. Because the t distribution is flatter than the normal distribution (the t has greater probability in the tails), it is important to use large samples whenever possible.

Assumption 2

This assumption is called "homoskedasticity," and its violation (nonconstant error) is called "heteroskedasticity": In other words, errors should not get larger as X gets larger. In Figure 19.1(a), errors have the same variance for all values of X_i; in Figure 19.1(b), the errors get larger as the value of X increases. Although Figure 19.1 shows a pattern for which the size of the error is positively related to the value of X (as X increases, e increases), the opposite situation is just as severe. If e decreases as X increases, the data are still heteroskedastic and violate this regression assumption. The same problem can affect the dependent variable; that is, the variance of the error term can increase as the values of Y increase. If this assumption of linear regression is violated, then both the standard error and t statistic associated with the slope coefficient (b) will be inaccurate. This violation can be serious because the standard error and the t statistic are used for testing the statistical significance of the slope; that is, whether there is a relationship between the independent variable X and the dependent variable Y (see Chapter 18). The fact that heteroskedasticity can occur for several reasons makes the topic complex. Aside from error terms that vary depending on the size of observations

Figure 19.1 **Errors and Their Variance**



for a variable, the condition can result from "outliers," or extreme data values (see Chapter 5) and measurement error in either/both the dependent variable or the independent variable. In the case of multiple regression (see Chapter 21), heteroskedasticity can result from the exclusion of one or more relevant *X* variables from the regression equation (as discussed earlier in this chapter) or nonconstant error variance across one or more of the *X* variables.

Statisticians have developed some half dozen different tests for diagnosing heteroskedasticity. One must have an idea of what might be causing heteroskedastic error disturbances to know which of these diagnostic tests should be used to confirm if heteroskedasticity is indeed a problem. Techniques such as weighted least squares (WLS), robust standard errors, and variable transformations (such as logarithmic transformations of the *X* or *Y* variables; see Chapter 20) can be used to correct for heteroskedasticity. Much like the various statistical tests for diagnosing the problem, the user should have a good idea of what is responsible for the heteroskedastic error terms before deciding which of these fixes to use. Given the complexity of these issues, the reader should consult an advanced textbook on regression analysis (see Fox, 2008) for more thorough coverage of how to diagnose and correct for heteroskedasticity.

Assumption 3

Assumption 3 is related to Assumptions 1 and 2. Assumption 3 is that the errors are independent of each other.

Another way of stating this assumption is to say that the size of one error is not a function of the size of any previous errors. We can test for nonindependent errors

by examining the residuals (the predicted value of Y minus the actual value of Y). If they appear to be random with respect to each other, then the errors are independent, and we need not worry. Computer programs exist that can determine whether errors are not random. If this problem ever exists, find a statistician or consult a textbook (see Pindyck and Rubinfeld, 2000). This problem usually causes difficulty only when time series data are used (see Chapter 20). Chapter 22 includes a test for nonrandom errors.

Assumption 4

Both the independent and the dependent variables must be interval variables (see Chapter 2).

The purist position is that regression cannot be performed with nominal or ordinal data. In a practical situation, however, regression with nominal or ordinal dependent and independent variables is possible. First, we will illustrate regression with a nominal independent variable.

The Homegrove City Parks superintendent wants to determine whether brand A or brand B riding mowers are more efficient. He tries three of each riding mower, testing them over normal city parks; he finds the data given in Table 19.3.

If the independent variable (the type of mower) is coded 1 when brand A is used and coded 0 when brand B is used, we have a nominal variable. Nominal variables with values of 1 or 0 are called **dummy variables** (see Chapter 23 for more discussion of dummy variables).

Table 19.3 Mower Brand and Acres of Grass Mowed

Mower	Acres of Grass Mowed
Brand *A*1	52
Brand *A*2	63
Brand *A*3	71
Brand *B*1	54
Brand *B*2	46
Brand *B*3	38

$X_i - \bar{X} = (X_i - \bar{X})$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y} = (Y_i - \bar{Y})$
$1 - .5 = .5$	.25	$52 - 54 = -2$
$1 - .5 = .5$	.25	$63 - 54 = 9$
$1 - .5 = .5$	.25	$71 - 54 = 17$
$0 - .5 = -.5$	.25	$54 - 54 = 0$
$0 - .5 = -.5$	.25	$46 - 54 = -8$
$0 - .5 = -.5$	.25	$38 - 54 = -16$

(continued)

Table 19.3 Mower Brand and Acres of Grass Mowed (*continued*)

$$\Sigma(X_i - \overline{X})^2 = 1.50$$

$(X_i - X)$	×	$(Y_i - Y)$	=	$(X_i - \overline{X})\,(Y_i - \overline{Y})$
.5	×	−2	=	−1
.5	×	9	=	4.5
.5	×	17	=	8.5
–.5	×	0	=	0
–.5	×	−8	=	4.0
–.5	×	−16	=	−8.0

$$\Sigma(X_i - \overline{X})(Y_i - \overline{Y}) = 24$$

$$b = \frac{24}{1.5} = 16$$

$$a = \overline{Y} - b\overline{X} = 54 - (16 \times .5) = 46$$

$$\hat{Y} = 46 + 16X$$

Brand of mower coded as a dummy variable appears with the data for the dependent variable below.

X	Y
1	52
1	63
1	71
0	54
0	46
0	38

$\overline{X} = .5$ $\overline{Y} = 54$

Recall from Chapter 18 that the formula for the slope of the regression line is

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

The calculations follow.

Because X can be only two values, 0 and 1, $\hat{Y}$ can be only two values, 46 and 62. If we test for the statistical significance of the regression slope, we will find out whether the brand A mowers cut significantly more grass than the brand B mowers. Recall that the formula for the standard error of the slope is

$$s_b = \frac{S_{y|x}}{\sqrt{\Sigma(X - \overline{X})^2}} = \frac{S_{y|x}}{\sqrt{1.5}} = \frac{S_{y|x}}{1.22}$$

Recall that $S_{y|x}$ can be calculated by the following formula:

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

The calculations follow.

$Y_i -$	$\hat{Y}_i =$	$(Y_i - \hat{Y}_i)$	$(Y_i - \hat{Y}_i)^2$
52 −	62 =	−10	100
63 −	62 =	1	1
71 −	62 =	9	81
54 −	46 =	8	64
46 −	46 =	0	0
38 −	46 =	−8	64

$$S^2_{y|x} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2} = \frac{310}{4} = 77.5$$

$$S_{y|x} = 8.8$$

Substituting this value into the preceding formula yields

$$s_b = \frac{8.8}{1.22} = 7.2$$

Converting $b = 16.0$ to a t score, we have

$$t = \frac{16.0 - 0}{7.2} = 2.22$$

With a t score of 2.22, the probability of brand A being no better than brand B is approximately .05 (t test, df = 4).

When the mower problem was first presented, you may have thought that this problem could have been solved with a test of means (see Chapter 14). Indeed, it can.

Brand A	Brand B
$\bar{X} = 62$	$\bar{Y} = 46$
$S = 9.5$	$s = 8.0$
s.e. $= 5.48$	s.e. $= 4.61$

$$\text{s.e.}_d = \sqrt{5.48^2 + 4.61^2} = 7.2$$

$$t = \frac{62 - 46}{7.2} = \frac{16}{7.2} = 2.22$$

Notice that we get the same answer that we obtained using regression. This result occurs because a difference of means test is similar to regression with dummy variables (a dummy variable is a nominal variable with codes 1 and 0).

You should also note that the regression intercept (46) is the same value as one of the means; the slope (16) is equal to the difference between the means; and the standard error of the slope (7.2) is equal to the overall standard error in the difference of means test.

Regression can also be performed with a nominal dependent variable. Suppose a personnel office tests the data entry skills of 10 job applicants, who are then hired. After 1 year, five of these processors have been fired. A personnel manager hypothesizes that the processors were fired because they lacked good keyboard skills. The job situation and data processing scores are listed in Table 19.4.

After subjecting these data to a regression computer program, the analyst found the following relationship:

$$\hat{Y} = -1.19 + .0248X \quad s_b = .009$$
$$S_{y|x} = .4 \quad r^2 = .49$$

Clearly a relationship exists ($t = 2.8$, df $= 8$). To understand the results of this regression, we must interpret $\hat{Y}$ as the probability that a processor is not fired. For example, substituting the first person's data processing score into the regression equation, we find

$$\hat{Y} = -1.19 + .0248(85) = -1.19 + 2.11 = .92$$

The probability that the first person will not be fired is .92. Similar calculations could be made for all data processors, and confidence limits could be placed around the probability by using the standard error of the estimate.

Regression with dummy dependent variables does have some pitfalls. If we substitute the word processing score of the fifth person (94) into the regression equation, we find

$$\hat{Y} = -1.19 + .0248(94) = -1.19 + 2.33 = 1.44$$

Table 19.4 Job Situation and Data Processing Score

Job Situation, *Y* (0 = Fired; 1 = Not Fired)	Data Processing Score, *X* (Words per Minute)
1	85
0	48
1	63
0	57
1	94
0	56
1	65
0	58
0	72
1	82

The probability that this person will not be fired is 1.14, a meaningless probability. Using regression with dummy dependent variables often results in probabilities greater than 1 or less than 0. Managerially, we might want to interpret probabilities of more than 1.0 as equal to .99. Similarly, probabilities of less than 0 can be reset to .01. For most management situations, these adjustments will eliminate uninterpretable predictions. Special types of analysis called *probit and logit analysis* can be used to restrict probabilities to values between 0 and 1. These techniques are fairly sophisticated and, therefore, should not be used without expert assistance. (If you want to learn more about them, please see Long, 2007.)

Assumption 5

The final assumption of regression is that the relationships are linear.

Linear relationships are those that can be summarized by a straight line (without any curve). If linear regression is used to summarize a nonlinear relationship, the regression equation will be inaccurate. To determine whether a relationship is linear, we must plot the data on a set of coordinate axes, just as we have been doing in this chapter and in Chapter 18. The data plotted in Figure 19.2 represent linear relationships.

Nevertheless, many relationships that a manager must consider are not linear. For example, the city manager may want to project city revenues for next year. The growth of city revenues may well look like the graph in Figure 19.3. Revenues increase in this example faster than a linear relationship would predict. The graph represents a **logarithmic relationship**. Such relationships and how they can be treated in regression are the subject of the next chapter.

Another relationship sometimes found in the public and nonprofit sectors is the *quadratic* relationship. In situations in which adding another worker will improve the productivity of all workers (because workers can then specialize and be more

Figure 19.2 Linear Relationships

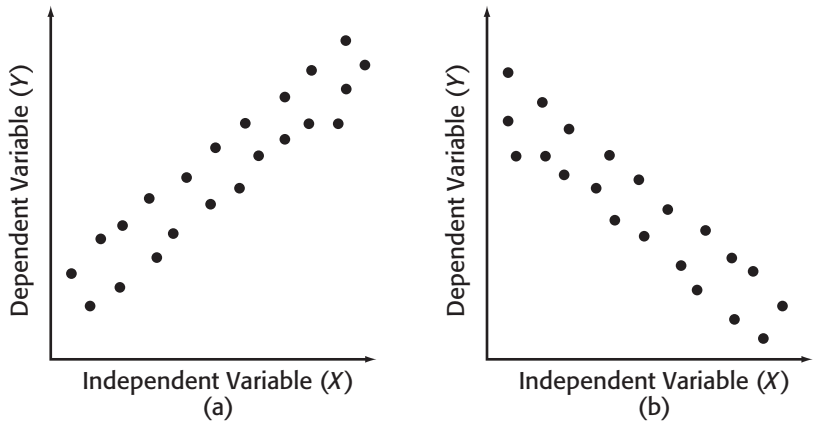


Figure 19.3 Nonlinear Relationship

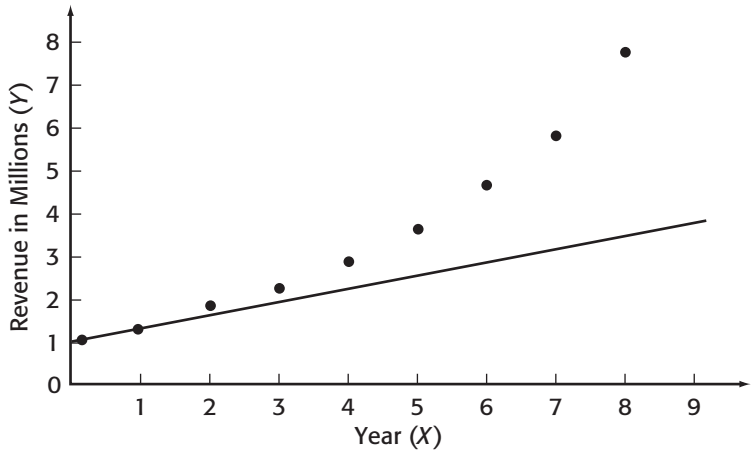


Table 19.5 Data Representing a Quadratic Relationship

Number of Welfare Workers (*X*)	Number of Cases Processed per Day (*Y*)
1	1
2	4
3	9
4	16
5	25

efficient), the relationship between the number of workers and total productivity may be quadratic. This idea is illustrated by the data in Table 19.5. Graphically, this relationship appears as shown in Figure 19.4. Quadratic relationships are discussed in Chapter 21.

Other relationships may be *cubic*. For example, the relationship of the total number of vice police to the number of prostitution arrests is probably cubic. The first few vice police will make very few arrests because they have so much territory to cover. Adding more vice police will raise the productivity of all vice police. As more and more vice police are added, the total arrests will level out (either because all possible prostitutes have been arrested or because they left town). The relationship would appear as shown in Figure 19.5. Cubic relationships will also be discussed in Chapter 21.

The important thing to remember about nonlinear relationships is that linear regression is not a good way to summarize them. Statistical package programs available on computers cannot (or, rather, usually do not) distinguish linear from nonlinear relationships; the onus is on the manager to discern nonlinear relationships.

Figure 19.4 Quadratic Relationship

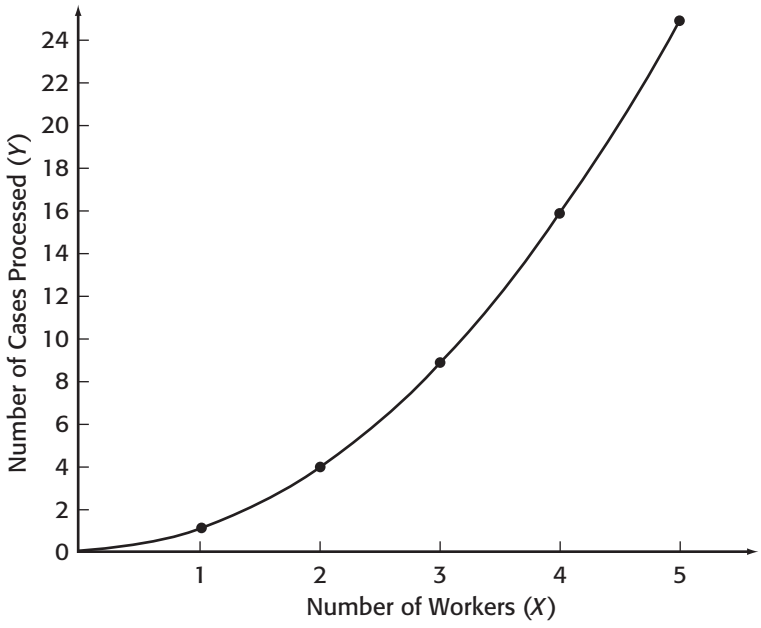


Figure 19.5 Cubic Relationship