Chapter 12

Linear Regression and Correlation

Objectives

In this chapter we study correlation and regression. We will

- demonstrate how the correlation coefficient is calculated and interpreted.
- show how least-squares regression models are fit to data.
- examine the relationship between the regression line, sample correlation, and the prediction of means.
- show how to test whether a regression relationship is statistically significant.
- extend regression ideas to multiple regression, analysis of covariance, and logistic regression.

12.1 Introduction

In this chapter we discuss some methods for analyzing the relationship between two quantitative variables, X and Y. **Linear regression** and **correlation analysis** are techniques based on fitting a straight line to the data.

Examples

Data for regression and correlation analysis consist of pairs of observations (X, Y). Here are two examples.

Example 12.1.1

Amphetamine and Food Consumption Amphetamine is a drug that suppresses appetite. In a study of this effect, a pharmacologist randomly allocated 24 rats to three treatment groups to receive an injection of amphetamine at one of two dosage levels, or an injection of saline solution. She measured the amount of food consumed by each animal in the 3-hour period following injection. The results (gm of food consumed per kg body weight) are shown in Table 12.1.1.[1]

Figure 12.1.1 shows a **scatterplot** of

$$Y = \text{Food consumption}$$

against

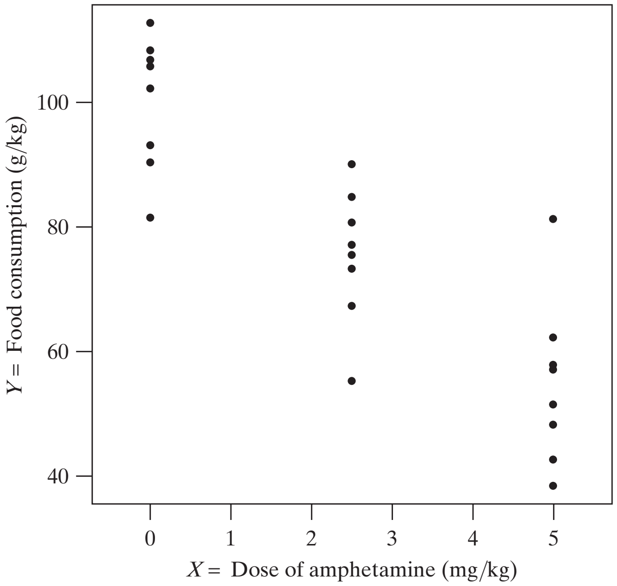
$$X = \text{Dose of amphetamine}$$

The scatterplot suggests a definite dose-response relationship, with larger values of X tending to be associated with smaller values of Y.* ■

Table 12.1.1 Food consumption (Y) of rats (gm/kg)

	X = Dose of amphetamine (mg/kg)		
	0	2.5	5.0
	112.6	73.3	38.5
	102.1	84.8	81.3
	90.2	67.3	57.1
	81.5	55.3	62.3
	105.6	80.7	51.5
	93.0	90.0	48.3
	106.6	75.5	42.7
	108.3	77.1	57.9
Mean	100.0	75.5	55.0
SD	10.7	10.7	13.3
No. of animals	8	8	8

Figure 12.1.1 Scatterplot of food consumption against dose of amphetamine



Example 12.1.2

Arsenic in Rice Environmental pollutants may enter the food supply as contaminants leach into the soil from which the food is grown. It is hypothesized that naturally occurring silicon in rice plants may inhibit the absorption of some pollutants. In a study to investigate compounds that mitigate arsenic absorption in rice, researchers sampled 32 rice plants and measured the concentration of arsenic in the polished rice (μg/kg rice) as well as the concentration of silicon in the straw (g/kg straw) of each plant.[2] Figure 12.1.2 shows a scatterplot of

$$Y = \text{rice arsenic concentration}$$

*In many dose-response relationships, the response depends linearly on log(dose) rather than on dose itself. We have chosen a linear portion of the dose-response curve to simplify the exposition.

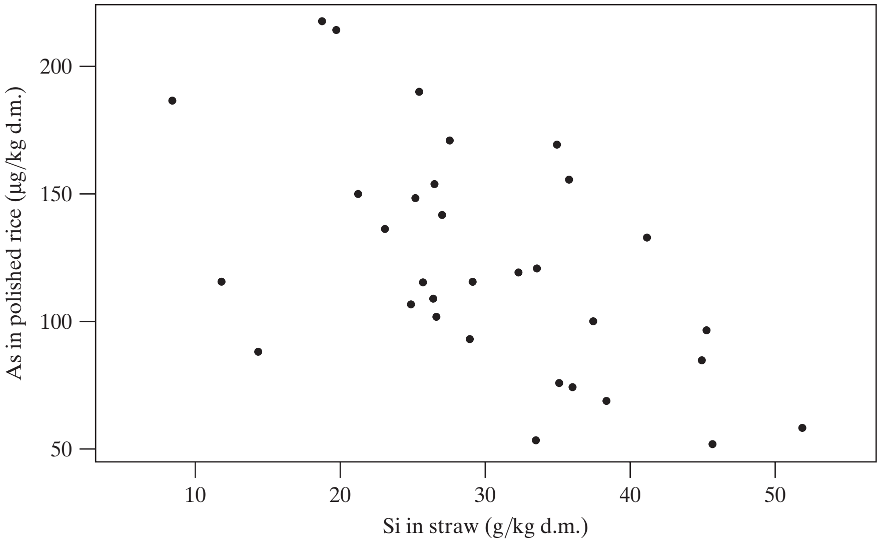


Figure 12.1.2 Scatterplot of rice arsenic concentration against straw silicon concentration

against

$$X = \text{straw silicon concentration}$$

The scatterplot suggests that higher straw silicon concentrations (X) tend to be associated with lower rice arsenic concentrations (Y). ■

12.2 The Correlation Coefficient

Suppose we have a sample of n pairs for which each pair represents the measurements of two variables, X and Y. If a scatterplot of Y versus X shows a general linear trend, then it is natural to try to describe the strength of the linear association. In this section we will learn how to measure the strength of linear association using the **correlation coefficient**. The following example illustrates the kind of situation we wish to consider.

Example 12.2.1

Length and Weight of Snakes In a study of a free-living population of the snake *Vipera bertis*, researchers caught and measured nine adult females.[3] Their body lengths and weights are shown in Table 12.2.1 and are displayed as a scatterplot in Figure 12.2.1. The number of observations is $n = 9$. ■

The scatterplot shown in Figure 12.2.1 shows a clear upward trend. We say that weight shows a **positive association** with length, indicating that greater lengths are associated with greater weights. Thus, snakes that are longer than the average length of $\bar{x} = 63$ tend to be heavier than the average weight of $\bar{y} = 152$. The line superimposed on the plot is called the **least-squares line** or **fitted regression line** of Y on X. We will learn how to compute and interpret the regression line in Section 12.3.

Measuring Strength of Linear Association

How strong is the linear relationship between snake length and weight? Are the data points tightly clustered around the regression line, or is the scatter loose? To answer these questions we will compute the **correlation coefficient**, a scale-invariant numeric measure of the strength of linear association between two quantitative variables. Being scale invariant means that the correlation coefficient is unaffected

Table 12.2.1

	Length X (cm)	Weight Y (g)
	60	136
	69	198
	66	194
	64	140
	54	93
	67	172
	59	116
	65	174
	63	145
Mean	63	152
SD	4.6	35.3

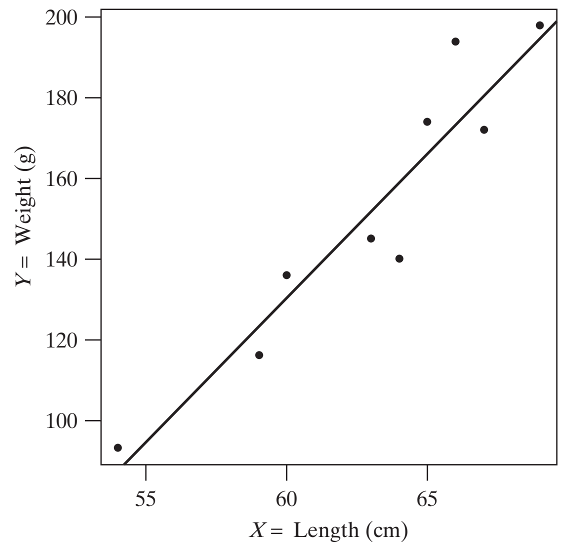


Figure 12.2.1 Body length and weight of nine snakes with fitted regression line

by any changes in measurement scales. That is, the correlation between length and weight will be the same whether measured in centimeters and grams or inches and pounds. To understand how the correlation coefficient works, consider again the snake length and weight example. Rather than plotting the original data, Figure 12.2.2 plots the standardized data (z-scores) displayed in Table 12.2.2; note that this plot looks identical to our original plot except now our scales are unit-less.

Dividing the plot into quadrants based on the sign of the standardized score, we see that most of these points fall into the upper-right and lower-left quadrants. Points falling in these quadrants will have standardized scores whose *products* are positive. Likewise, points falling in the upper-left and lower-right quadrants will have standardized score products that are negative. Computing the sum of these products provides a numeric measure of where our points fall (i.e., which quadrants are dominant). In our case, since there is a positive association between length and weight, most points fall in the positive product quadrants; thus, the sum of the

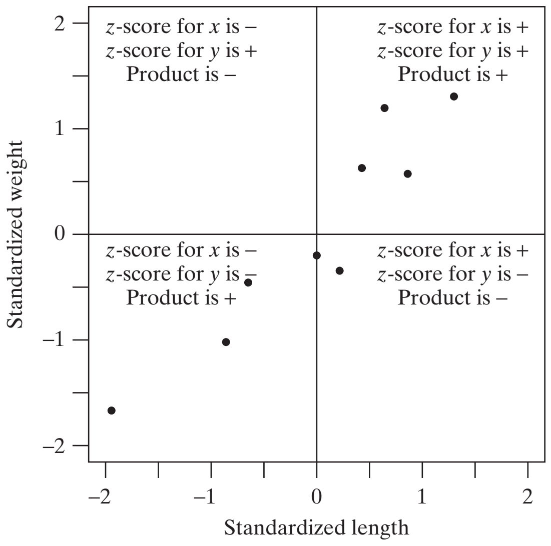


Figure 12.2.2 Scatterplot of standardized weight versus standardized length

Table 12.2.2 Standardized snake weights, lengths, and their products

	Weight	Length	Standardized weight	Standardized length	Product of standardized values
	X	Y	$z_x = \frac{x - \bar{x}}{s_x}$	$z_y = \frac{y - \bar{y}}{s_y}$	$z_x z_y$
	60	136	−0.65 . . .	−0.45 . . .	0.29 . . .
	69	198	1.29 . . .	1.30 . . .	1.68 . . .
	66	194	0.65 . . .	1.19 . . .	0.77 . . .
	64	140	0.22 . . .	−0.34 . . .	−0.07 . . .
	54	93	−1.94 . . .	−1.67 . . .	3.24 . . .
	67	172	0.86 . . .	0.57 . . .	0.49 . . .
	59	116	−0.86 . . .	−1.02 . . .	0.88 . . .
	65	174	0.43 . . .	0.62 . . .	0.27 . . .
	63	145	0.00 . . .	−0.20 . . .	0.00 . . .
Sum	567	1368	0.00	0.00	7.5494
Mean	63.000	152.000	0.00	0.00	
SD	4.637	35.338	1.00	1.00	

Values in the table are truncated for ease of reading. Because the summary values will be used in subsequent calculations, they include more digits than one would typically report when following our rounding conventions.

products of standardized scores is positive. If a negative relationship were present, most of the points would fall in the negative quadrants and the sum would be negative. And, if there were no *linear* relationship, the points would fall in evenly in all four quadrants so that the positive and negative products would balance and their sum would be zero.

The correlation coefficient is based on this sum. It is computed as the average product of standardized scores (using $n - 1$ rather than n to compute the average):*

The correlation coefficient, r

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

From this formula it is clear that X and Y enter r symmetrically; therefore, if we were to interchange the labels X and Y of our variables, r would remain unchanged. In fact, this is one of the advantages of the correlation coefficient as a summary statistic: In interpreting r, it is not necessary to know (or to decide) which variable is labeled X and which is labeled Y.

*By substituting $\sqrt{\sum_{i=1}^{n} (x - \bar{x})^2/(n - 1)}$ for s_x and $\sqrt{\sum_{i=1}^{n} (y - \bar{y})^2/(n - 1)}$ for s_y, the equation for the correlation coefficient can be rewritten as $r = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x - \bar{x})^2 \sum_{i=1}^{n}(y - \bar{y})^2}}$.

Interpreting the Correlation Coefficient

Mathematically, the correlation coefficient is unit free and always between -1 and 1. The sign of the correlation indicates the sign of the relationship and matches the sign of the slope of the regression line: positive (increasing) or negative (decreasing). The closer the correlation is to -1 or 1, the stronger the linear relationship between X and Y. A correlation equal to -1 or 1 indicates a perfect linear relationship between the two variables—a scatterplot of such data would display the data falling exactly on a straight line. Interestingly, a correlation of zero does not necessarily mean that there is no relationship between X and Y—it only means that there is no *linear* relationship between X and Y. The preceding computation of the correlation indicates that the sum of the products of standardized values will be zero whenever the positive and negative products balance; this can happen in many ways. Figure 12.2.3 displays several examples with a variety of correlation coefficient values.

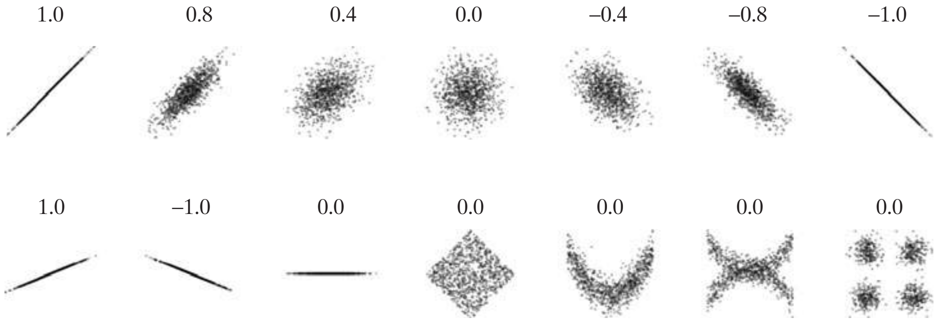


Figure 12.2.3 Scatterplots of data with a variety of sample correlation values

Example 12.2.2

Length and Weight of Snakes In Table 12.2.2 we showed that for the snake data the sum of the products of the standardized scores is 7.5494. Thus, the correlation coefficient for the lengths and weights of our sample of nine snakes is about 0.94.

$$r = \frac{1}{9-1} \times 7.5494 \approx 0.94$$ ■

In this example we may also refer to the value 0.94 as the **sample correlation**, since the lengths and weights of these nine snakes comprise a sample from a larger population. The sample correlation is an estimate of the **population correlation** (often denoted by the Greek letter “rho,” ρ)—in this case the correlation coefficient for the entire population of adult female *Vipera bertis* snakes. In order to regard the sample correlation coefficient r as an estimate of a population parameter, it must be reasonable to assume that both the X and the Y values were selected at random, as in the following **bivariate random sampling model**:

Bivariate Random Sampling Model:

We regard each pair (x_i, y_i) as having been sampled at random from a population of (x, y) pairs.

In the bivariate random sampling model, the observed X’s are regarded as a random sample and the observed Y’s are also regarded as a random sample, so that the marginal statistics $\bar{x}, \bar{y}, s_x$, and s_y are estimates of corresponding population values μ_x, μ_y, σ_x, and σ_y.

For many investigations the random sampling model is reasonable, but the additional assumption of a bivariate random sampling model is not. This is generally the case when the values of X are specified by the experimenter as in Example 12.1.1 where the researchers assigned rats to one of three dosages of amphetamine. This type of sampling model is called the random subsampling model and is defined in Section 12.4. In these cases the sample correlation coefficient is not an appropriate estimate of the population correlation.

Inference Concerning Correlation

We have described how the correlation coefficient describes a data set within the bivariate random sampling model. Now we shall consider statistical inference based on r for data from this model. Consider the following example.

Testing the Hypothesis H_0: $\rho = 0$

In some investigations it is not a foregone conclusion that there is any relationship between X and Y. It then may be relevant to consider the possibility that any apparent trend in the data is illusory and reflects only sampling variability. In this situation it is natural to formulate the null hypothesis

H_0: X and Y are uncorrelated in the population.

or, equivalently

H_0: There is no linear relationship between X and Y.

A t test of H_0 is based on the test statistic

$$t_s = r\sqrt{\frac{n - 2}{1 - r^2}}$$

Critical values are obtained from Student's t distribution with

$$\text{df} = n - 2$$

The following example illustrates the application of this t test.

Example 12.2.3

Blood Pressure and Platelet Calcium It is suspected that calcium in blood platelets may be related to blood pressure. As part of a study of this relationship, researchers recruited 38 subjects whose blood pressure was normal (that is, not abnormally elevated).[4] For each subject two measurements were made: pressure (average of systolic and diastolic measurements) and calcium concentration in the blood platelets. The data are shown in Figure 12.2.4. The sample size is $n = 38$ and the sample correlation is $r = 0.5832$.

Is there evidence that blood pressure and platelet calcium are linearly related? We will test the null hypothesis

$$H_0: \rho = 0$$

against the nondirectional alternative

$$H_A: \rho \neq 0$$

These hypotheses are translations of the verbal hypotheses

H_0: Platelet calcium is not linearly related to blood pressure.

H_A: Platelet calcium is linearly related blood pressure.

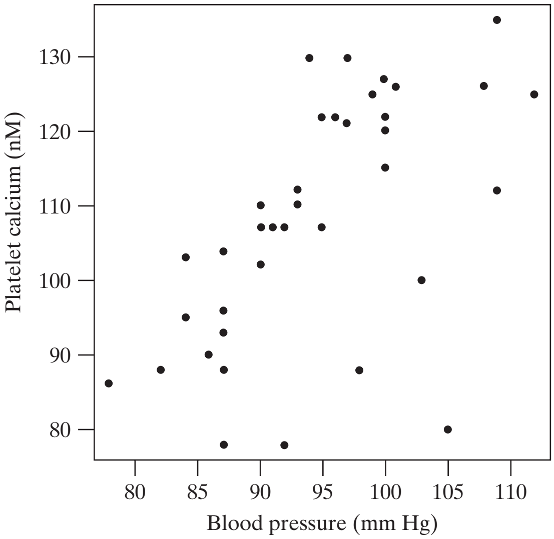


Figure 12.2.4 Blood pressure and platelet calcium for 38 persons with normal blood pressure

Let us choose $\alpha = 0.05$. The test statistic is

$$t_s = 0.5832\sqrt{\frac{38 - 2}{1 - 0.5832^2}} = 4.308$$

From Table 4 with df $= n - 2 = 36 \approx 40$, we find $t_{40,0.0005} = 3.551$. Thus, we find P-value $< 0.0005 \times 2 = 0.001$ (since H_A is nondirectional) and we reject H_0. The data provide strong evidence that platelet calcium is linearly related to blood pressure ($t_s = 4.308$, df $= 36$, P-value < 0.001). ■

Why $n - 2$? The t statistic in the hypothesis test for the preceding population correlation coefficient has an associated df $= n - 2$. The origin of the $n - 2$ is easy to explain. Any two points determine a straight line, yet such a small data set ($n = 2$) provides no information about the inherent variability in the scatter of the points (or, equivalently, the strength of association between X and Y). It is not until we observe a third point that we are able to begin estimating the strength of any relationship. As in our earlier contexts related to t distributions and F-distributions (Chapters 6, 7, 8, and 11), the degrees of freedom is the number of pieces of information provided by the data about the "noise" from which the investigator wants to extract the "signal."

Confidence Interval for ρ (Optional)

If the sample size is large, it is possible to construct a confidence interval for ρ. The sampling distribution of the sample correlation coefficient, r, is skewed, so in order to construct the confidence interval we apply what is known as the Fisher transformation of r:

$$z_r = \frac{1}{2}\ln\left[\frac{1 + r}{1 - r}\right]$$

where ln is the natural logarithm (base e). We can then construct a 95% confidence interval for $\frac{1}{2}\ln\left[\frac{1 + \rho}{1 - \rho}\right]$ as

$$z_r \pm 1.96\frac{1}{\sqrt{n - 3}}$$

Finally, we can convert the limits of the confidence interval for $\frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right]$ into a confidence interval for ρ by solving for ρ in the equations given by

$$\frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right] = z_r \pm 1.96\frac{1}{\sqrt{n-3}}$$

Intervals with other confidence levels are constructed analogously. For example, to construct a 90% confidence interval, replace 1.96 with 1.645. The construction of a confidence interval for a correlation coefficient is illustrated in Example 12.2.4.

Example 12.2.4

Blood Pressure and Platelet Calcium For the data of Example 12.2.3 the sample size is $n = 38$ and the sample correlation is $r = 0.5832$. The Fisher transformation of r gives

$$z_r = \frac{1}{2}\ln\left[\frac{1+0.5832}{1-0.5832}\right] = \frac{1}{2}\ln\left[\frac{1.5832}{0.4168}\right] = 0.6673$$

A 95% confidence interval for $\frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right]$ is

$$0.6673 \pm 1.96\frac{1}{\sqrt{38-3}}$$

or 0.6673 ± 0.3313, which is $(0.3360, 0.9986)$.

Setting

$$\frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right] = 0.3360 \text{ gives } \rho = \frac{e^{2(0.3360)}-1}{e^{2(0.3360)}+1} = 0.32$$

Setting

$$\frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right] = 0.9986 \text{ gives } \rho = \frac{e^{2(0.9986)}-1}{e^{2(0.9986)}+1} = 0.76$$

We are 95% confident that the correlation between blood pressure and platelet calcium in the population is between 0.32 and 0.76. Thus, a 95% confidence interval for ρ is $(0.32, 0.76)$. ■

Correlation and Causation

We have noted earlier that an observed association between two variables does not necessarily indicate any causal connection between them. It is important to remember this caution when interpreting correlation. The following example shows that even strongly correlated variables may be causally unrelated.

Example 12.2.5

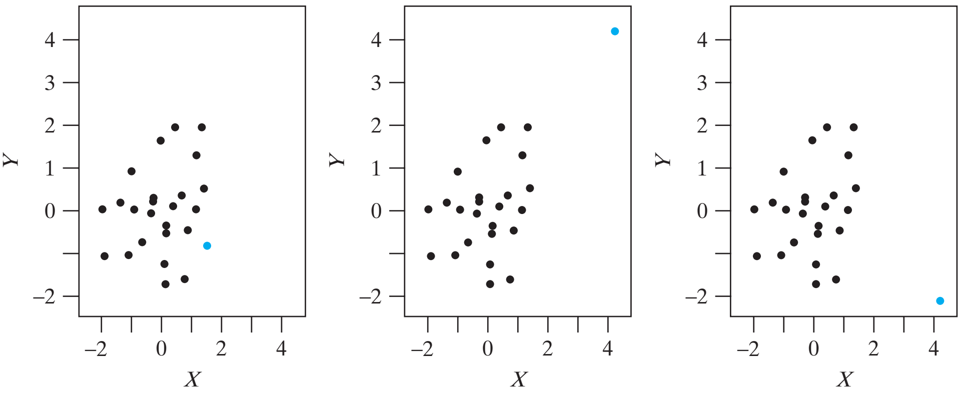
Reproduction of an Alga Akinetes are sporelike reproductive structures produced by the green alga *Pithophora oedogonia.* In a study of the life cycle of the alga, researchers counted akinetes in specimens of alga obtained from an Indiana lake on 26 occasions over a 17-month period. Low counts indicated germination of the akinetes. The researchers also recorded the water temperature and the photoperiod (hours of daylight) on each of the 26 occasions. The data showed a rather strong negative correlation between akinete counts and photoperiod; the correlation coefficient was $r = -0.72$. The researchers, however, recognized that this observed correlation might not reflect a causal relationship. Longer days (increasing photoperiod) also tend to bring higher temperatures, and the akinetes might actually be responding to temperature rather than photoperiod. To resolve the question, the researchers conducted laboratory experiments in which temperature and photoperiod were varied independently; these experiments showed that temperature, not photoperiod, was the causal agent.[5] ■

As Example 12.2.5 shows, one way to establish causality is to conduct a controlled experiment in which the putative causal factor is varied and all other factors are either held constant or controlled by randomization. When such an experiment is not possible, indirect approaches using statistical analysis can shed some light on potentially causal relationships. (One such approach will be illustrated in Example 12.8.3.)

Cautionary Notes

To describe the results of testing a correlation coefficient, investigators often use the term *significant*, which can be misleading. For instance, a statement such as "A highly significant correlation was noted" is easily misunderstood. It is important to remember that statistical significance simply indicates rejection of a null hypothesis; it does not necessarily indicate a large or important effect. A "significant" correlation may in fact be quite a weak one; its "significance" means only that it cannot easily be explained away as a chance pattern. From the formula $t_s = r\sqrt{\frac{n-2}{1-r^2}}$ we can see that for a fixed value of r, t_s increases as n increases. Thus, if the sample size is large enough, t_s will be large enough for the correlation to be "significant" no matter how small r is. It is always wise to assess the practical significance of any result by considering a confidence interval for the population parameter of interest.

The correlation coefficient is highly sensitive to extreme points. For example, Figure 12.2.5(a) shows a scatterplot of 25 points with a correlation of $r = 0.2$; one of the points has been plotted as a blue dot. Figure 12.2.5(b) shows the same points, except that the point plotted as a blue dot has been changed. The change of that single point causes the correlation coefficient to climb from 0.2 to 0.6. Figure 12.2.5(c) shows a third version of the data. In this case $r = -0.1$. These three graphs illustrate how a single point can greatly influence the size of the correlation coefficient. It is important to always plot the data before using r (or any other statistic) to summarize the data.



(a) $r = 0.2$

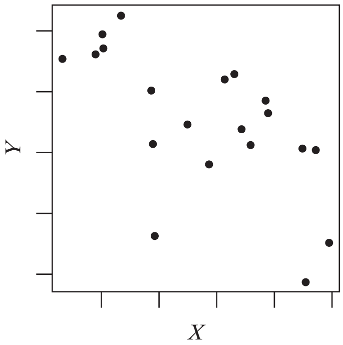
(b) $r = 0.6$

(c) $r = -0.1$

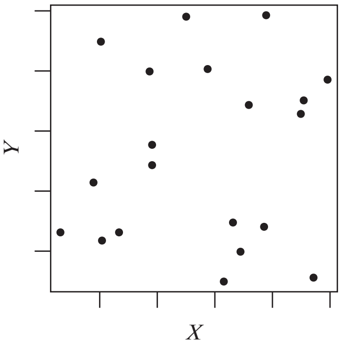
Figure 12.2.5 The effect of outliers on the sample correlation coefficient

Exercises 12.2.1–12.2.10

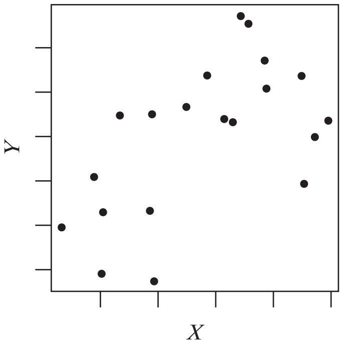
12.2.1 Arrange the following plots in order of their correlations (from closest to −1 to closest to +1).



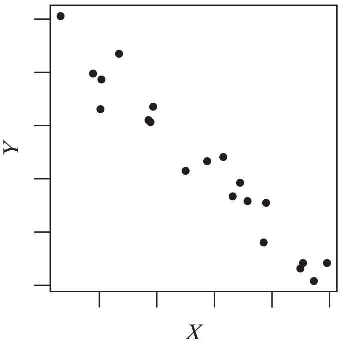
(a)



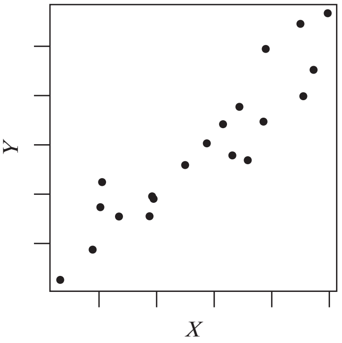
(b)



(c)



(d)



(e)

12.2.2 Consider the following data.

	X	y
	6	6
	1	7
	3	3
	2	2
	5	14
Mean	3.4	6.4
SD	2.1	4.7

(a) Plot the data. Does there appear to be a relationship between X and Y? Is it linear or nonlinear? Weak or strong?

(b) Compute the sample correlation coefficient between X and Y.

(c) Is there significant evidence that X and Y are correlated? Conduct a test using $\alpha = 0.05$.

12.2.3 In a study of natural variation in blood chemistry, blood specimens were obtained from 284 healthy people. The concentrations of urea and of uric acid were measured for each specimen, and the correlation between these two concentrations was found to be $r = 0.2291$. Test the hypothesis that the population correlation coefficient is zero against the alternative that it is positive.[6] Let $\alpha = 0.05$.

12.2.4 Researchers measured the number of neurons in the CA1 region of the hippocampus in the brains of eight persons who had died of causes unrelated to brain function. They found that these data were negatively correlated with age. The sample value of r was −0.63.[7]

(a) Is this correlation coefficient significantly different from zero? Conduct a test using $\alpha = 0.10$.

(b) Suppose in part (a) you found that the correlation does significantly differ from zero. Does this provide evidence that aging is a cause for CA1 neuron loss? If not, what could be said? Briefly explain.

12.2.5 Twenty plots, each 10 × 4 meters, were randomly chosen in a large field of corn. For each plot, the plant density (number of plants in the plot) and the mean cob weight (gm of grain per cob) were observed. The results are given in the table.[8]

PLANT DENSITY X	COB WEIGHT Y	PLANT DENSITY X	COB WEIGHT Y
137	212	173	194
107	241	124	241
132	215	157	196
135	225	184	193
115	250	112	224
103	241	80	257
102	237	165	200
65	282	160	190
149	206	157	208
85	246	119	224

Preliminary calculations yield the following results:

$$\bar{x} = 128.05 \qquad \bar{y} = 224.10$$
$$s_x = 32.61332 \qquad s_y = 24.95448$$
$$r = -0.94180$$

(a) Is there significant evidence for a linear relationship between cob weight and plant density? Carry out an appropriate test using $\alpha = 0.05$.

(b) Is this study an observational study or an experiment?

(c) Farmers are interested in whether manipulating plant density can alter cob weight. Could these data be used to answer this question? If not, what could be said? Briefly explain.

12.2.6 Laetisaric acid is a compound that holds promise for control of fungus diseases in crop plants. The accompanying data show the results of growing the fungus *Pythium ultimum* in various concentrations of laetisaric acid. Each growth value is the average of four radial measurements of a *P. ultimum* colony grown in a petri dish for 24 hours; there were two petri dishes at each concentration.[9]

(a) Is there significant evidence for a linear relationship between fungus growth and acid concentration? Carry out an appropriate test using $\alpha = 0.05$.

(b) Is this study an observational study or an experiment?

(c) It is suggested that acid could be used to retard fungus growth. Could these data be used to verify this claim? If not, what could be said? Briefly explain.

	LAETISARIC ACID CONCENTRATION X (μG/ml)	FUNGUS GROWTH Y (mm)
	0	33.3
	0	31.0
	3	29.8
	3	27.8
	6	28.0
	6	29.0
	10	25.5
	10	23.8
	20	18.3
	20	15.5
	30	11.7
	30	10.0
Mean	11.500	23.642
SD	10.884	7.8471
	$r = -0.98754$	

12.2.7 To investigate the dependence of energy expenditure on body build, researchers used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure for each man during conditions of quiet sedentary activity. The results are shown in the table.[10] (See also Exercise 12.5.5.)

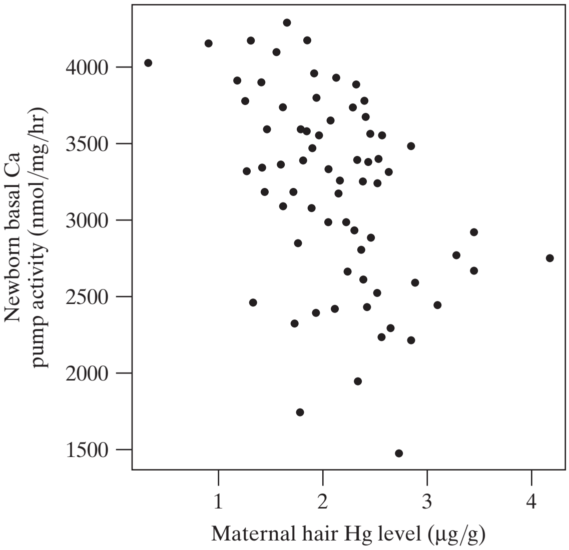
SUBJECT	FAT-FREE MASS X (kg)	ENERGY EXPENDITURE Y (kcal)
1	49.3	1,894
2	59.3	2,050
3	68.3	2,353
4	48.1	1,838
5	57.6	1,948
6	78.1	2,528
7	76.1	2,568
Mean	62.400	2,168.429
SD	12.095	308.254
	$r = 0.98139$	

(a) The correlation between energy expenditure and fat-free mass is very large (near 1). It is 0.98139, but the sample size is quite small, only 7. Is there enough evidence to claim the correlation is different from zero? Carry out an appropriate test using $\alpha = 0.05$.

(b) Is this study an observational study or an experiment?

(c) Persons who exercise could increase their fat-free mass. Could these data be used to claim that their energy expenditure would also increase? If not, what could be said? Briefly explain.

12.2.8 Cellular ability to regulate homeostasis is measured by basal Ca pump activity. Deregulation of calcium homeostasis can trigger serious effects of cell functioning. Can maternal mercury exposure measured by mercury deposits in hair (μg/g) affect newborn's basal Ca pump activity (nmol/mg/hr)? The following data summaries and graph are from a human study involving a sample of 75 newborns and their mothers.[11]



$$\bar{x} = 2.11183 \qquad \bar{y} = 3196.8196$$
$$s_x = 0.61166 \qquad s_y = 611.34876$$
$$r = -0.45289$$

(a) It is a good habit to always plot our data before analysis. Examining the preceding scatterplot, does there seem to be a linear trend in the data? Is it increasing or decreasing? Is it weak or strong?

(b) Examining the plot, we see there is a mother with a maternal hair level around 4.2 μg/g. If her child's basal Ca pump activity were changed from about 2800 to about 2000 nmol/mg/hr, would the sample correlation increase or decrease?

(c) Is there evidence that newborn basal Ca pump activity linearly decreases with maternal hair level? Carry out an appropriate test using $\alpha = 0.05$.

(d) In part (c) you should have found that there is strong evidence for a linearly decreasing relationship between X and Y. Explain how the evidence can be so strong even though the graph displays substantial scatter and the sample correlation is not close to -1.

(e) Based on your answer to part (c) and the design of this study, what can we say regarding the primary research question: Is there statistical evidence that maternal mercury exposure measured by mercury deposits in hair (μg/g) *affects* newborn's basal Ca pump activity (nmol/mg/hr)?

12.2.9 For each of the following examples, explain whether or not it is reasonable to treat the sample correlation coefficient, r, as an estimate of a population correlation coefficient ρ. Briefly justify your answer.

(a) The blood chemistry data from Exercise 12.2.3.

(b) The CA1 neuron data from Exercise 12.2.4.

(c) The cob weight data from Exercise 12.2.5.

(d) The fungus growth data from Exercise 12.2.6.

(e) The basal Ca pump activity from Exercise 12.2.8.

12.2.10 (optional) For each of the following data sets, compute a 95% confidence interval for the population correlation coefficient.

(a) The blood chemistry data from Exercise 12.2.3.

(b) The cob weight data from Exercise 12.2.5.

(c) The energy expenditure data from Exercise 12.2.7.

12.3 The Fitted Regression Line

In Section 12.2 we learned how the correlation coefficient describes the strength of linear association between two numeric variables, X and Y. In this section we will learn how to find and interpret the line that best summarizes their linear relationship.

Example 12.3.1

Ocean Temperature Consider a data set for which there is a perfect linear relationship between X and Y for example, temperature measured in X = Celsius and Y = Fahrenheit. Figure 12.3.1 displays 20 weekly ocean temperatures (in both °C and °F) for a coastal California city along with a line that perfectly describes the relationship:* $y = 32 + \frac{9}{5}x$. A summary of the data appears in Table 12.3.1.[12]

*This equation is the Celsius to Fahrenheit conversion formula.

Figure 12.3.1 Scatterplot of Y = ocean temperature in °F versus X = ocean temperature in °C. The mean value $(\bar{x}, \bar{y})$ is denoted with a ▲

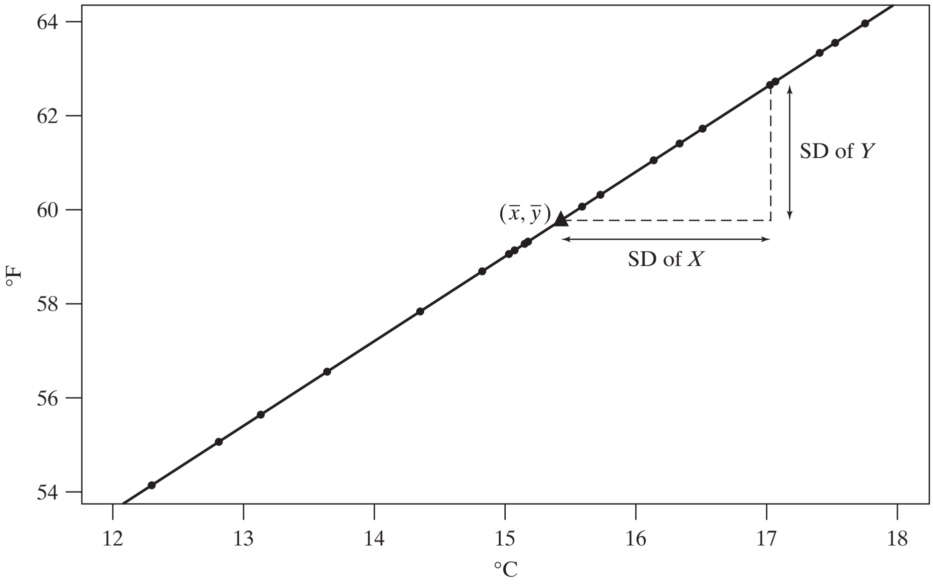


Table 12.3.1 Summary of water temperature data

	X = temperature (°C)	Y = temperature (°F)
Mean	15.43	59.77
SD	1.60	2.88

Because X and Y are measuring the same variable (temperature), it stands to reason that a water specimen that is 1 SD above average in °C ($s_x = 1.60$) will also be 1 SD above average in °F ($s_y = 2.88$). Combined, these values can describe the slope of the line that fits these data exactly:

$$\frac{\text{rise}}{\text{run}} = \frac{s_y}{s_x} = \frac{2.88}{1.60} = 1.80$$

In this example we also happen to know the equation of the line that describes the Celsius to Fahrenheit conversion. The slope of this line is $9/5 = 1.80$, the same value we found previously. ■

The SD Line

In perfect linear relationships (i.e., when $r = \pm 1$) the line that fits the data exactly will have slope $\pm s_y/s_x$ (the sign of the slope matches the sign of the correlation coefficient) and passes through the point $(\bar{x}, \bar{y})$. This line is sometimes referred to as the **SD line**. Our previous temperature example displays this property. But what about situations in which r is not exactly ± 1, that is, when the relationship between X and Y is less than perfectly linear?

Example 12.3.2

Arsenic in Rice In Section 12.1 we observed a scatterplot indicating that the amount of arsenic in rice and silicon in rice straw appears to be linearly related ($r = -0.566$). Figure 12.3.2 displays a scatterplot of these data along with the SD line (dashed line). At first glance the SD line appears to be a good fit to these data; however, further investigation suggests otherwise. Suppose we wanted to estimate the mean arsenic concentration in rice for plants with straw silicon concentrations of

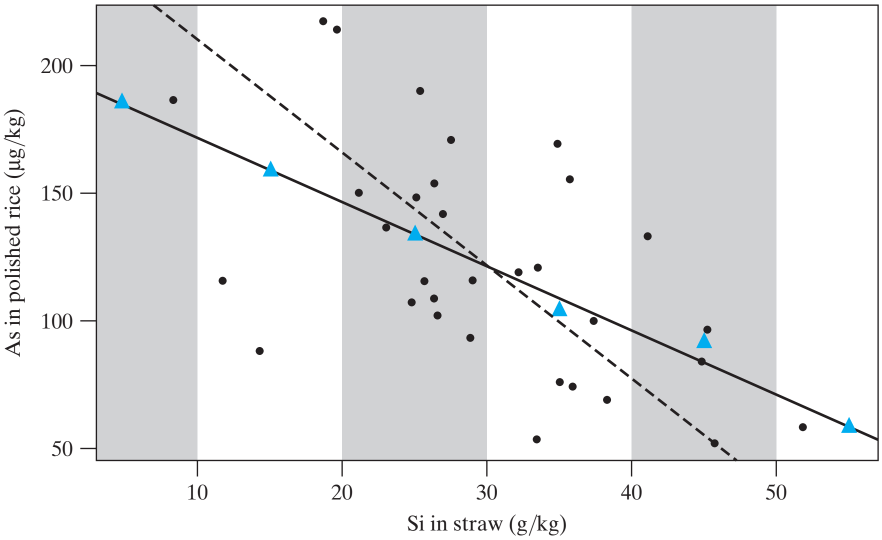


Figure 12.3.2 Concentrations of arsenic in rice versus silicon in straw for 32 rice plants. The dashed and solid lines are the SD and fitted regression lines, respectively. Each ▲ symbol indicates the mean rice arsenic concentration for a range of straw silicon concentrations specified by the shading

15 g/kg. The SD line suggests an estimated mean rice arsenic concentration of approximately 190 µg/kg. Another way to estimate this value would be to simply use the mean rice arsenic concentration for plants in our sample that have straw silicon concentrations around 15 g/kg. The mean arsenic concentration in rice for straw silicon concentrations between 10 and 20 g/kg is 158.6 µg/kg (denoted by a ▲ on the graph), which is considerably less than the 190 µg/kg value given by the SD line. Similarly, for plants with straw silicon concentrations around 45 g/kg, the SD line indicates an arsenic level of about 55 µg/kg while the mean arsenic level for plants with silicon between 40 and 50 g/kg in our sample is 91.4 µg/kg, a much larger value. ■

The rice arsenic example shows that the SD line tends to overestimate the mean value of Y for below average X values and underestimate the mean value of Y for above average X values. Figure 12.3.3 shows an even more exaggerated example for a data set with a correlation even farther from ±1; it is near zero ($r = -0.05$). Recall that a correlation of zero indicates no linear relationship between X and Y. This lack of linear relationship is demonstrated by the fact that the

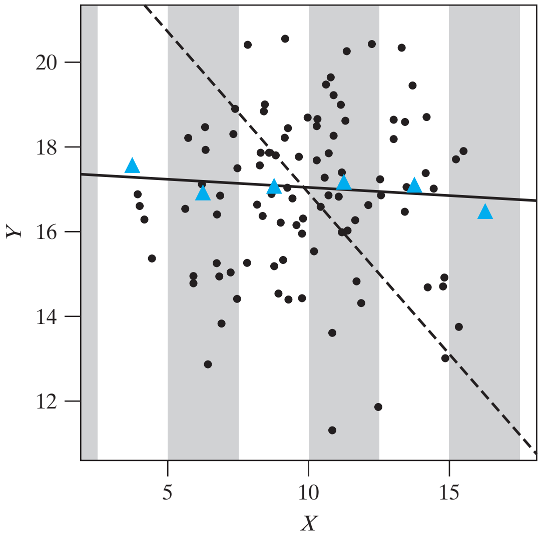


Figure 12.3.3 Scatterplot, SD line (dashed), and fitted regression line (solid) for a sample of 100 data (x, y) values with a correlation near zero. The ▲ symbols indicate the mean Y values for ranges of X values specified by the shading

mean value of Y is about the same (≈ 17) regardless of the value of X (most of the ▲'s in the plot are near 17).

If the SD line can be such a poor summary, why bother studying it? Because it is an ideal starting place based on a perfect linear relationship. With a perfect (positive) linear relationship, the SD line is the best fitting line and has a slope of s_y/s_x. Our examples illustrate that if the relationship is not perfect, the relationship between the mean Y values and X values has a flatter slope. Mathematically, it can be shown that the line that is best suited to predicting Y (in a certain sense)—the so called **least-squares** or **fitted regression** line—has a slope equal to $r(s_y/s_x)$ and passes through the point $(\bar{x}, \bar{y})$. That is, for X values one standard deviation above average, the mean Y value will only be r standard deviations above average (assuming that r is positive; if r is negative, then for X values one standard deviation above average, the mean Y value will be r standard deviations below average).

Example 12.3.3

Arsenic in Rice A summary and scatterplot of our rice arsenic data appear in Table 12.3.2 and Figure 12.3.4. In this example we estimate that plants with straw silicon concentrations that are $s_x = 10.04$ g/kg above average (i.e., one standard deviation above average) will have rice arsenic concentrations that are 25.19 μg/kg lower than average ($r \times s_y = -0.566 \times 44.50 = -25.19$). Equivalently, the slope of the fitted regression line is

$$r(s_y/s_x) = -0.566 \times (44.50/10.04) = -2.51\,(\mu\text{g As/kg rice})/(\text{g Si/kg straw})$$

meaning that each additional 1-g/kg increase in straw silicon concentration is associated with a 2.51-μg/kg decrease in the rice arsenic concentration, on average. ■

Table 12.3.2 Summary of rice arsenic data

	X = Si in straw (g/kg)	Y = As in rice (μg/kg)
Mean	29.85	122.25
SD	10.04	44.50
	$r = -0.566$	

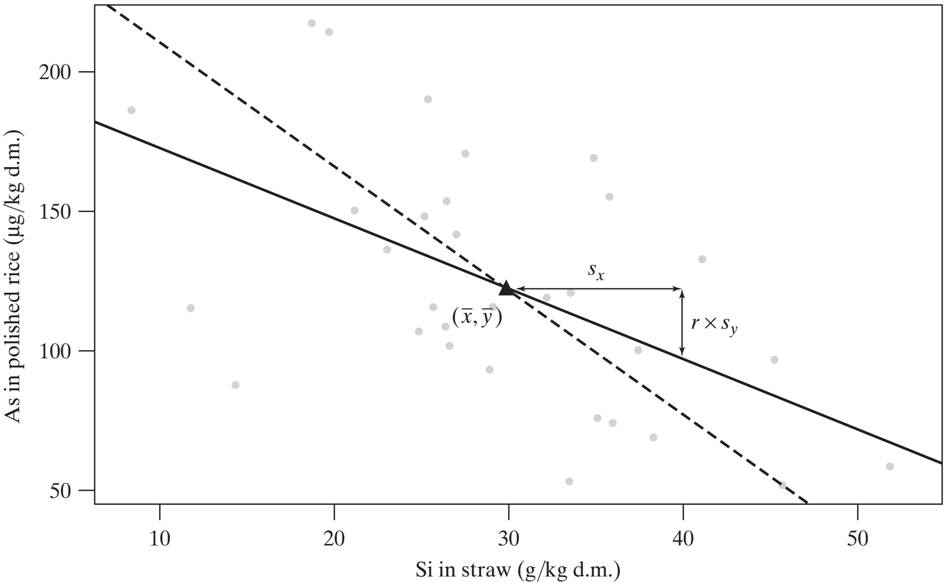


Figure 12.3.4 Concentrations of arsenic in rice versus silicon in straw for 32 rice plants with SD line (dashed) and fitted regression line (solid)

Equation of the Fitted Regression Line

The equation of a straight line can be written as

$$Y = b_0 + b_1 X$$

where b_0 is the y-intercept and b_1 is the slope of the line. The slope b_1 is the rate of change of Y with respect to X.

The fitted regression line of Y on X is written $\hat{y} = b_0 + b_1 x$. We write $\hat{y}$ (read "Y-hat") in place of Y to remind us that this line is providing only estimated or predicted Y values; unless the correlation is ± 1, we don't expect the data values to fall exactly on the line. The fitted regression line estimates the mean value of Y for any given value of X. We discuss this concept of the regression line as a *line of averages* in further detail below.

The slope and intercept of the least-squares* regression line are calculated from the data as follows:

Least-Squares Regression Line of Y on X

$$\text{Slope: } b_1 = r\left(\frac{s_y}{s_x}\right)$$

$$\text{Intercept: } b_0 = \bar{y} - b_1\bar{x}$$

Previously we saw the motivation for the formula for the slope, b_1. The formula for the intercept is also easy to motivate. We can rewrite the Y-intercept formula as

$$\bar{y} = b_0 + b_1\bar{x}$$

which shows that *regression line passes through the joint mean* $(\bar{x}, \bar{y})$ of our data.

We illustrate the use of these formulas by continuing our rice arsenic example.

Example 12.3.4

Arsenic in Rice Previously we found the slope of the regression line to be $b_1 = r(s_y/s_x) = -2.51$ (μg As/kg rice)/(g Si/kg straw). Using this value we find the Y-intercept,

$$b_0 = 122.25 - (-2.51) \times 29.85 = 197.17\,\mu\text{g/kg}$$

Thus, our fitted regression line is $\hat{y} = 197.17 - 2.51x$ as previously displayed in Figure 12.3.4. ■

Note that the Y-intercept, the point $(0, b_0) = (0, 197.17)$, does not appear on the scatterplot in Figure 12.3.4 as the X-scale limits do not extend to zero; they range from about 5 to 55 to produce a plot for which the data fill the picture nicely.

Graph of Averages

If we have several observations of Y at a given level of X, we can estimate the population mean Y value for the given X value ($\mu_{Y|X}$) by simply using the sample average of Y, $\bar{y}$, for that given value of X; we can denote this sample average as $\bar{y}|X$.[†]

*There are other methods of finding fitted regression lines. In this text, we consider only the least-squares regression line, which aims to minimize the squared vertical distances between the data values and the fitted line.

[†] A more detailed exposition of these "conditional means" appears in Section 12.4.

Sometimes we are able to calculate a sample average, $\bar{y}$, for each of several X values. A graph of $\bar{y}|X$ is known as a **graph of averages**, since it shows the (observed) average of Y for different values of X.

Example 12.3.5

Amphetamine and Food Consumption Figure 12.3.5 is a graph of averages for the food consumption data in Table 12.1.1, showing the average y value for each of the 3 levels of X. Note that the 3 $\bar{y}$'s almost lie on a line. This supports the use of the linear model with these data. ■

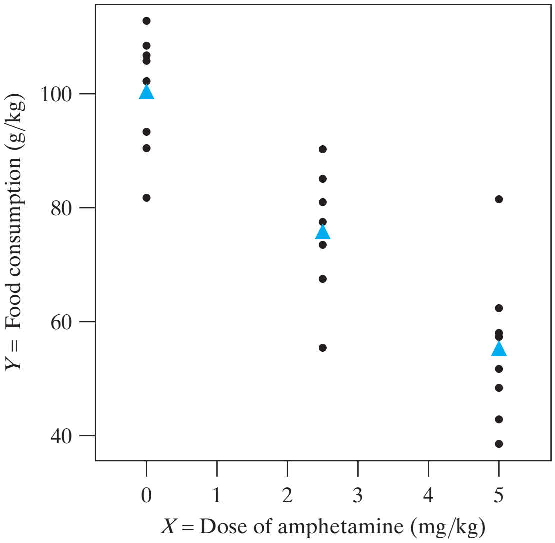


Figure 12.3.5 Graph of averages (▲) for food consumption data from Example 12.1.1 with the original data plotted as black dots

If the $\bar{y}$'s in a graph of averages fall exactly on a line, then that line is the regression line and $\mu_{Y|X}$ is estimated with $\bar{y}|X$. Usually, however, the $\bar{y}$'s are not perfectly collinear. In this case, the regression line is a *smoothed* version of the graph of averages, resulting in a fitted model in which all of the estimates of $\mu_{Y|X}$ fall on a line. By smoothing the graph of averages into a line, we use information from *all* the observations to estimate $\mu_{Y|X}$ at any level of X.

Example 12.3.6

Amphetamine and Food Consumption If we apply the preceding regression formulas to the food consumption data in Table 12.1.1, we obtain $b_0 = 99.3$ and $b_1 = -9.01$. Thus, the estimate of $\mu_{Y|X=0}$ is 99.3 g/kg. This estimate differs slightly from $\bar{y}|X = 0$, which is 100.0 g/kg. The estimate 99.3 makes use of (1) the 8 y values when $X = 0$ (which averaged to 100.0) and (2) the linear trend established by the other 16 data points, which showed higher food consumption associated with lower doses. Likewise, $\mu_{Y|X=2.5}$ is $99.3 - 9.01 \times 2.5 = 76.78$ g/kg, which differs slightly from $\bar{y}|X = 2.5$, which is 75.5 g/kg, and $\mu_{Y|X=5}$ is $99.3 - 9.01 \times 5 = 54.25$ g/kg, which differs slightly from $\bar{y}|X = 5$, which is 55.0 g/kg. ■

The idea of smoothing the graph of averages into a straight line carries over to the setting in which we have only a single observation at each level of X, as is the case with the rice arsenic example. When we draw a line through a set of (X, Y) data, we are expressing a belief that the underlying dependence of the mean value of Y on X is smooth, even though the data may show the relationship only roughly. Linear regression is one formal way of providing a smooth description of the data.

The Residual Sum of Squares

We now consider a statistic that describes the scatter of the points about the fitted regression line. The equation of the fitted line is $y = b_0 + b_1x$. Thus, for each observed x_i in our data there is a predicted Y value of

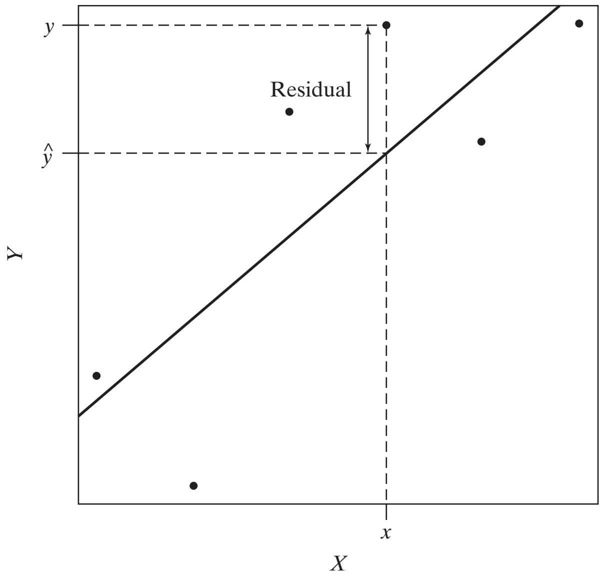
$$\hat{y}_i = b_0 + b_1x_i$$

Also associated with each observed pair (x_i, y_i) is a quantity called a **residual**, defined as

$$e_i = y_i - \hat{y}_i$$

Figure 12.3.6 shows $\hat{y}$ and the residual for a typical data point (x_i, y_i). It can be shown that the sum of the residuals, taking into account their signs, is always zero, because of "balancing" of data points above and below the fitted regression line. The *magnitude* (absolute value) of each residual is the vertical distance of the data point from the fitted line.

Figure 12.3.6 $\hat{y}$ and the residual for a typical data point (x, y)



Note that a residual is calculated in terms of *vertical* distance. In using the regression model $\hat{y} = b_0 + b_1x$ we are thinking of the variable X as a predictor and the variable Y as a response that depends on X. We care primarily about how close each observed value, y_i, is to its predicted value, $\hat{y}_i$. Thus, we measure vertical distance from each point to the fitted line. A summary measure of the distances of the data points from the regression line is the **residual sum of squares**, or **SS(resid)**, which is defined as follows:

Residual Sum of Squares

$$\text{SS(resid)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}e_i^2$$

It is clear from the definition that the residual sum of squares will be small if the data points all lie very close to the line.

The following example illustrates the computation of SS(resid).

Example 12.3.7

Arsenic in Rice For the rice arsenic data, Table 12.3.3 indicates how SS(resid) would be calculated from its definition. The values displayed are abbreviated to improve readability. ■

Table 12.3.3 Calculation of SS(resid) for a portion of the rice arsenic data

Obs #	x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8.3	186.2	176.2 . . .	10.0 . . .	99.50 . . .
2	11.8	115.5	167.6 . . .	−52.1 . . .	2716.00 . . .
3	14.3	87.9	161.2 . . .	−73.3 . . .	5373.93 . . .
4	18.7	217.2	150.2 . . .	67.0 . . .	4492.74 . . .
5	19.7	213.8	147.8 . . .	66.0 . . .	4356.67 . . .
6	21.2	150.0	144.0 . . .	6.0 . . .	35.53 . . .
7	23.0	136.2	139.4 . . .	−3.2 . . .	10.26 . . .
8	25.1	148.3	134.1 . . .	14.2 . . .	200.46 . . .
9	26.4	153.4	130.8 . . .	22.6 . . .	512.49 . . .
⋮	⋮	⋮	⋮	⋮	⋮
27	38.3	69.0	101.0 . . .	−32.1 . . .	1028.99 . . .
28	41.1	132.8	94.0 . . .	38.8 . . .	1503.19 . . .
29	45.2	96.6	83.6 . . .	12.9 . . .	167.11 . . .
30	44.9	84.5	84.5 . . .	0.0 . . .	0.00 . . .
31	45.7	51.7	82.5 . . .	−30.8 . . .	948.51 . . .
32	51.8	58.6	67.1 . . .	−8.5 . . .	71.69 . . .
Sum				0.0	41727.11 = SS(resid)

The Least-Squares Criterion

Many different criteria can be proposed to define the straight line that "best" fits a set of data points. The classical criterion is the least-squares criterion:

Least-Squares Criterion

The "best" straight line is the one that minimizes the residual sum of squares.

The formulas given for b_0 and b_1 were derived from the least-squares criterion by applying calculus to solve the minimization problem. (The derivation is given in Appendix 12.1.) The fitted regression line is also called the "least-squares line."

The least-squares criterion may seem arbitrary and even unnecessary. Why not fit a straight line by eye with a ruler? Actually, unless the data lie nearly on a straight line, it can be surprisingly difficult to fit a line by eye. The least-squares criterion provides an answer that does not rely on individual judgment and that (as we shall see in Sections 12.4 and 12.5) can be usefully interpreted in terms of estimating the distribution of Y values for each fixed X. Furthermore, we will see in Section 12.8 that the least-squares criterion is a versatile concept, with applications far beyond the simple fitting of straight lines.

The Residual Standard Deviation

A summary of the results of the linear regression analysis should include a measure of the closeness of the data points to the fitted line. A measure derived from the residual sum of squares and easier to interpret is the **residual standard deviation**, denoted s_e, which is defined as follows:

Residual Standard Deviation

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n-2}} = \sqrt{\frac{\text{SS(resid)}}{n-2}}$$

The residual standard deviation tells how far above or below the regression line points tend to be. Thus, the residual standard deviation specifies how far off predictions made using the regression model tend to be. Notice the factor in the denominator $n - 2$, rather than the usual $n - 1$. The following example illustrates the calculation of s_e.

Example 12.3.8

Arsenic in Rice For the rice arsenic data, we use SS(resid) from Example 12.3.7 to calculate

$$s_e = \sqrt{\frac{41727.11}{32-2}} = \sqrt{1390.90} = 37.30 \text{ µg/kg}$$

Thus, predictions for the concentrations of arsenic in rice based on the regression model tend to err by about 37.30 µg/kg on average. ■

Note that the formula for s_e is closely analogous to the formula for s_y:

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Both these SDs measure variability in Y, but the residual SD measures variability around the *regression line* and the ordinary SD measures variability around the mean, $\bar{y}$. Roughly speaking, s_e is a measure of the typical vertical distance of the data points from the regression line. (Notice that the unit of measurement of s_e is the same as that of Y—for instance, µg/kg in the case of the rice arsenic data or grams in the case of the snake data from Example 12.2.1.) Figure 12.3.7 shows the scatterplot and regression line for the snake data from Example 12.2.1 with the residuals represented as vertical lines and the residual SD indicated as a vertical ruler line. Note that the residual SD roughly indicates the magnitude of a typical residual. Finding the equation of this line and the residual standard deviation appears as an exercise at the end of this section.

In many cases, s_e can be given a more definite quantitative interpretation. Recall from Section 2.6 that for a "nice" data set, we expect roughly 68% of the observations to be within 1 SD of the mean (and similarly for 95%, 2 SDs). Recall also that these rules work best if the data follow approximately a normal distribution. Similar interpretations hold for the residual SD: For "nice" data sets that are not too small, we expect roughly 68% of the observed Y's to be within $\pm 1s_e$ of the regression line. In other words, we expect roughly 68% of the data points to be within a vertical distance of s_e above and below the regression line (and similarly for 95%,

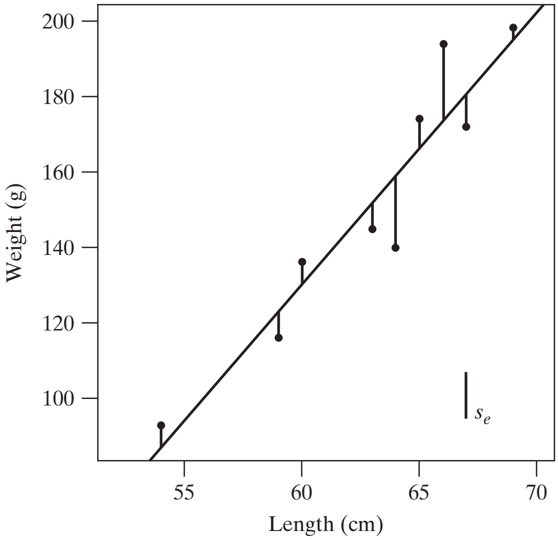


Figure 12.3.7 Weight versus length of nine snakes showing the residuals and a line segment denoting the magnitude of the residual SD

$\pm 2s_e$). These rules work best if the residuals follow approximately a normal distribution. The rice arsenic data we've been working with are well-suited to illustrate the 68% rule.

Example 12.3.9

Arsenic in Rice For the rice arsenic data, the fitted regression line is $\hat{y} = 197.17 - 2.51x$ and the residual standard deviation is $s_e = 37.30$. Figure 12.3.8 shows the data and the regression line. The dashed lines are a vertical distance of s_e from the regression line. Of the 32 data points, 22 are within the dashed lines; thus, 22/32 or $\approx$ 69% of the observed Y's are within $\pm 1s_e$ of the regression line. ■

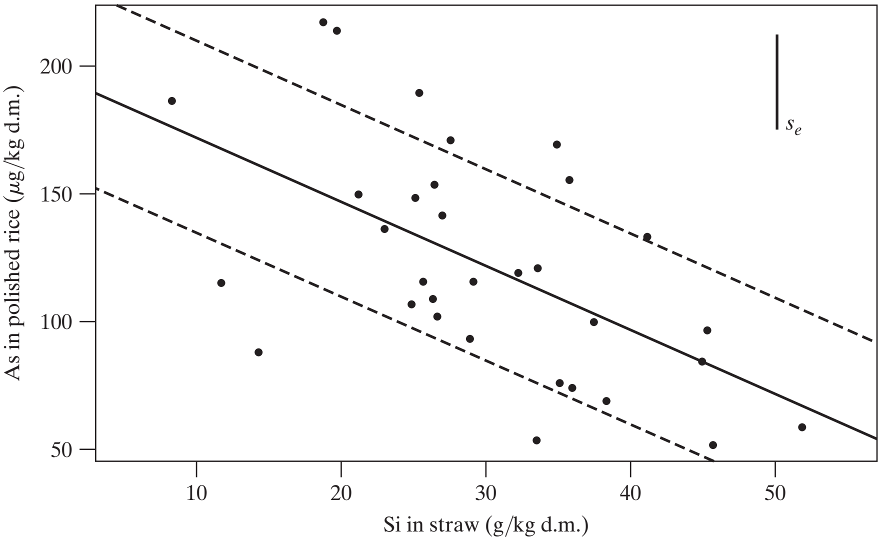


Figure 12.3.8 Arsenic in rice versus silicon in straw for 32 rice plants. The dashed lines are a vertical distance of s_e from the regression line

The Coefficient of Determination

We have said that the magnitude of r describes the tightness of the linear relationship between X and Y and have seen how its value is related to the slope of the regression line. When squared, it also provides an additional and very interpretable summary of the regression relationship. The **coefficient of determination**, r^2, describes the proportion of the variance in Y that is explained by the linear relationship between Y and X. This interpretation follows from the following fact (proved in Appendix 12.2).

Fact 12.3.1: Approximate Relationship of r to s_e and s_y

The correlation coefficient r obeys the following approximate relationship:

$$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

(The approximation in Fact 12.3.1 is best for large n, but it holds reasonably well even for n as small as 10.) The numerator, $s_y^2 - s_e^2$, can be roughly interpreted as the total variance in Y explained by the regression line: It is the difference between the variance in Y and the residual variance—the variance left over after fitting the regression line to the data. If the line fits the data very well, then s_e^2 will be close to zero so this numerator will be close to s_y^2; in this case r^2 will be close to 1. At the other extreme, if the line is a very poor fit, then s_e^2 will be close to s_y^2 and the numerator will be close to 0; in this case r^2 will be close to 0. The denominator, s_y^2, is the variance of Y; thus the ratio, r^2, is the proportion of the variance of Y that is explained by the regression relationship between Y and X. Note that because $-1 \le r \le 1$, $0 \le r^2 \le 1$. The following examples illustrate the interpretation and an application of r^2 in context.

Example 12.3.10

Arsenic in Rice For the rice arsenic data, we found $r = -0.566$, so $r^2 = 0.320$ or 32.0%. Thus, 32% of the variance in rice arsenic concentration is explained by the linear relationship between rice arsenic concentration and straw silicon concentration. ■

Example 12.3.11

Amphetamine and Food Consumption The standard deviation of food consumption for our entire sample of 24 rats (i.e., combining rats across all three doses of amphetamine) was $s_y = 21.84$ g/kg. Further, suppose r^2 was given to be 0.739. What is the estimated standard deviation of food consumption for rats given 4-mg/kg doses of amphetamine? That is, what is the value of $s_{Y|X=4}$?

To answer this question we first must recognize that the value of X is irrelevant; the residual standard deviation s_e describes the standard deviation of Y values for any given X value, and therefore for $X = 4$. Thus, we need to find the value of s_e. From Fact 12.3.1 we have

$$r^2 \approx 1 - \frac{s_e^2}{s_y^2}$$

After a little algebra, we find that the (approximate) standard deviation of food consumption for rats given 4-mg/kg doses of amphetamine is

$$s_e \approx s_y\sqrt{1 - r^2} = 21.84\sqrt{1 - 0.739} = 11.16\,\text{g/kg}$$ ■

Exercises 12.3.1–12.3.10

12.3.1 In a study of protein synthesis in the oocyte (developing egg cell) of the frog *Xenopus laevis*, a biologist injected individual oocytes with radioactively labeled leucine. At various times after injection, he made radioactivity measurements and calculated how much of the leucine had been incorporated into protein. The results are given in the accompanying table; each leucine value is the content of labeled leucine in two oocytes. All oocytes were from the same female.[13]

	TIME	LEUCINE
	0	0.02
	10	0.25
	20	0.54
	30	0.69
	40	1.07
	50	1.50
	60	1.74
Mean	30.00	0.830
SD	21.60	0.637
	$r = 0.993$	
	SS(resid) = 0.035225	

(a) Plot the data. Does there appear to be a relationship between X and Y? Is it linear or nonlinear? Weak or strong?

(b) Use linear regression to estimate the rate of incorporation of the labeled leucine.

(c) Draw the regression line on your graph.

(d) Calculate the residual standard deviation.

12.3.2 In an investigation of the physiological effects of alcohol (ethanol), 15 mice were randomly allocated to three treatment groups, each to receive a different oral dose of alcohol. The dosage levels were 1.5, 3.0, and 6.0 gm alcohol per kg body weight. The body temperature of each mouse was measured immediately before the alcohol was given and again 20 minutes afterward. The accompanying table shows the drop (before minus after) in body temperature for each mouse. (The negative value −0.1 refers to a mouse whose temperature rose rather than fell.)[14]

ALCOHOL		DROP IN BODY TEMPERATURE (°C)					
DOSE (gm/kg)	LOG(DOSE) X	INDIVIDUAL VALUES (Y)					MEAN
1.5	0.176	0.2	1.9	−0.1	0.5	0.8	0.66
3.0	0.477	4.0	3.2	2.3	2.9	3.8	3.24
6.0	0.778	3.3	5.1	5.3	6.7	5.9	5.26

(a) Plot the mean drop in body temperature versus dose. Plot the mean drop in body temperature versus log(dose). Which plot appears more nearly linear?

(b) Plot the individual (x, y) data points [where X = log(dose)].

(c) For the regression of Y on X = log (dose) preliminary calculations yield the following: $\bar{x} = 0.477$, $\bar{y} = 3.05333$, $s_x = 0.25439$, $s_y = 2.13437$, $r = 0.91074$. Calculate the fitted regression line and the (approximate) residual standard deviation.

(d) Draw the regression line on your graph.

(e) Is this study an example of an observational study or an experiment? How can you tell?

(f) Could data from this study be used to determine whether or not alcohol lowers body temperature? Briefly explain.

12.3.3 Consider the cob weight data from Exercise 12.2.5.

(a) Use the summaries in Exercise 12.2.5 to calculate the fitted regression line and approximate residual standard deviation.

(b) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

(c) SS(resid) = 1337.3. Use this value to compute the residual standard deviation. How does it compare to the approximate value determined in part (a)?

(d) Interpret the value of s_e in the context of this setting.

(e) What proportion of the variation in cob weights is explained by the linear relationship between cob weight and density?

12.3.4 Consider the Fungus growth data from Exercise 12.2.6.

(a) Calculate the linear regression of Y on X.

(b) Plot the data and add the regression line to your graph. Does the line appear to fit the data well?

(c) SS(resid) = 16.7812. Use this to compute s_e. What are the units of s_e?

(d) Draw a ruler line on your graph to show the magnitude of s_e. (See Figure 12.3.8).

12.3.5 Consider the Energy Expenditure data from Exercise 12.2.7.

(a) Calculate the linear regression of Y on X.

(b) Plot the data and add the regression line to your graph. Does the line appear to fit the data well?

(c) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

(d) SS(resid) = 21026.1. Use this to compute s_e. What are the units of s_e?

12.3.6 The rowan (*Sorbus aucuparia*) is a tree that grows in a wide range of altitudes. To study how the tree adapts to its varying habitats, researchers collected twigs with attached buds from 12 trees growing at various altitudes in North Angus, Scotland. The buds were brought back to the laboratory and measurements were made of the dark respiration rate. The accompanying table shows the altitude of origin (in meters) of each batch of buds and the dark respiration rate (expressed as μl of oxygen per hour per mg dry weight of tissue).[15]

	ALTITUDE OF ORIGIN X (m)	RESPIRATION RATE Y (μl/hr × mg)
	90	0.11
	230	0.20
	240	0.13
	260	0.15
	330	0.18
	400	0.16
	410	0.23
	550	0.18
	590	0.23
	610	0.26
	700	0.32
	790	0.37
Mean	433.333	0.21000
SD	214.617	0.07710
	$r = 0.88665$	
	SS(resid) = 0.013986	

(a) Calculate the linear regression of Y on X.

(b) Plot the data and the regression line.

(c) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

(d) Calculate the residual standard deviation.

12.3.7 Scientists studied the relationship between the length of the body of a bullfrog and how far it can jump. Eleven bullfrogs were included in the study. The results are given in the table.[16]

(a) Calculate the linear regression of Y on X.

(b) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

(c) What proportion of the variation in maximum jump distances can be explained by the linear relationship between jump distance and frog length?

(d) Calculate the residual standard deviation and specify the units.

(e) Interpret the value of the residual standard deviation in the context of this setting.

BULLFROG	LENGTH X (mm)	MAXIMUM JUMP Y (cm)
1	155	71.0
2	127	70.0
3	136	100.0
4	135	120.0
5	158	103.3
6	145	116.0
7	136	109.2
8	172	105.0
9	158	112.5
10	162	114.0
11	162	122.9
Mean	149.6364	103.9909
SD	14.4725	17.9415
	$r = 0.28166$	
	SS(resid) = 2,963.61	

12.3.8 The peak flow rate of a person is the fastest rate at which the person can expel air after taking a deep breath. Peak flow rate is measured in units of liters per minute and gives an indication of the person's respiratory health. Researchers measured peak flow rate and height for each of a sample of 17 men. The results are given in the table.[17]

SUBJECT	HEIGHT X (cm)	PEAK FLOW RATE Y (l/min)
1	174	733
2	183	572
3	176	500
4	169	738
5	183	616
6	186	787
7	178	866
8	175	670
9	172	550
10	179	660
11	171	575
12	184	577
13	200	783
14	195	625
15	176	470
16	176	642
17	190	856
Mean	180.4118	660.0000
SD	8.5591	117.9952
	$r = 0.32725$	
	SS(resid) = 198,909	

(a) Calculate the linear regression of Y on X.

(b) What proportion of the variation in flow rate is explained by the linear regression of flow rate on height?

(c) For each subject, calculate the predicted peak flow rate, using the regression equation from part (a).

(d) For each subject, calculate the residual, using the results from part (c).

(e) Calculate s_e and specify the units.

(f) What percentage of the data points are within $\pm s_e$ of the regression line? That is, what percentage of the 17 residuals are in the interval $(-s_e, s_e)$?

12.3.9 For each of the following data sets, prepare a plot like Figure 12.3.8, showing the data, the fitted regression line, and two lines whose vertical distance above and below the regression line is s_e. What percentage of the data points are within $\pm s_e$ of the regression line? What percentage of the data points do you expect to find within $\pm s_e$ of the regression line? How do these values compare?

(a) The body temperature data of Exercise 12.3.2.

(b) The corn yield data of Exercise 12.3.3.

12.3.10 Suppose a large sample of (x, y) pairs were used to fit the regression of Y on X. Now suppose we observed 100 further (x, y) pairs. About how many of these new observations would you expect to be farther than $2s_e$ from the regression line?

12.4 Parametric Interpretation of Regression: The Linear Model

One use of regression analysis is simply to provide a concise description of the data. The quantities b_0 and b_1 locate the regression line and s_e describes the scatter of the points about the line.

For many purposes, however, data description is not enough. In this section we consider inference from the data to a larger population. In previous chapters we have spoken of one or several populations of Y values. Now, to encompass the X variable as well, we need to expand the notion of a population.

Conditional Populations and Conditional Distributions

A **conditional population** of Y values is a population of Y values associated with a fixed, or given, value of X. Within a conditional population we may speak of the **conditional distribution** of Y. The mean and standard deviation of a conditional population distribution are denoted as

$$\mu_{Y|X} = \text{Population mean } Y \text{ value for a given } X$$
$$\sigma_{Y|X} = \text{Population SD of } Y \text{ values for a given } X$$

(Note that the "given" symbol "|" is the same one used for conditional probability in Chapters 3 and 10.) The following example illustrates this notation.

Example 12.4.1

Amphetamine and Food Consumption In the rat experiment introduced in Example 12.1.1, the response variable Y was food consumption and the three values of X (dose) were $X = 0$, $X = 2.5$, and $X = 5$. In Example 12.3.5 we examined the graph of averages and considered the food consumption data as three independent samples (as for an ANOVA). In the ANOVA context we denote the three population means as μ_1, μ_2, and μ_3. In regression notation these means would be denoted as

$$\mu_{Y|X=0} \quad \mu_{Y|X=2.5} \quad \mu_{Y|X=5}$$

respectively. Similarly, the three population standard deviations, which would be denoted as σ_1, σ_2, and σ_3 in an ANOVA context, would be denoted as

$$\sigma_{Y|X=0} \quad \sigma_{Y|X=2.5} \quad \sigma_{Y|X=5}$$

respectively. In other words, the symbols

$$\mu_{Y|X} \text{ and } \sigma_{Y|X}$$

represent the mean and standard deviation of food consumption values for rats that are given dose X of amphetamine. ■

In observational studies, conditional distributions pertain to subpopulations rather than experimental treatment groups, as in the following example.

Example 12.4.2 Height and Weight of Young Men Consider the variables

$$X = \text{Height}$$

and

$$Y = \text{Weight}$$

for a population of young men. The conditional means and standard deviations are

$$\mu_{Y|X} = \text{Mean weight of men who are } X \text{ inches tall}$$
$$\sigma_{Y|X} = \text{SD of weights of men who are } X \text{ inches tall}$$

Thus, $\mu_{Y|X}$ and $\sigma_{Y|X}$ are the mean and standard deviation of weight in the *subpopulation* of men whose height is X. Of course, there is a different subpopulation for each value of X. ■

The Linear Model

When we conduct a linear regression analysis, we think of Y as having a distribution that depends on X. The analysis can be given a parametric interpretation if two conditions are met. These conditions, which constitute the **linear model**, are given in the following box.

The Linear Model

1. *Linearity.* $Y = \mu_{Y|X} + \varepsilon$ where $\mu_{Y|X}$ is a linear function of X; that is

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

   Thus, $Y = \beta_0 + \beta_1 X + \varepsilon$.

2. *Constancy of standard deviation.* $\sigma_{Y|X}$ does not depend on X. We denote this constant value as σ_ε.

In the linear model $Y = \beta_0 + \beta_1 X + \varepsilon$, the ε term represents **random error**. We include this term in the model to reflect the fact that Y varies, even when X is fixed. The variability of Y for a fixed value of X is measured by the conditional standard deviation of Y, $\sigma_{Y|X}$. But, because the linear model stipulates that this standard deviation is the same for every value of X, we commonly use the notation σ_ε to represent this standard deviation and refer to it as the standard deviation of the random error.

The following two examples show the meaning of the linear model.

Example 12.4.3

Amphetamine and Food Consumption For the rat food consumption experiment, the linear model asserts that (1) the population mean food consumption is a linear function of dose, and that (2) the population standard deviation of food consumption values is the same for all doses. Notice that the second condition is closely analogous to the condition in ANOVA that the population SDs are equal: $\sigma_1 = \sigma_2 = \sigma_3$. The linear model also allows for the fact that there is variability in Y when X is fixed. For example, there were 8 observations for which $X = 5$. The 8 y-values averaged 55.0, but none of the observations was equal to 55.0; there was substantial variability within the 8 y-values. This variability is quantified by the SD of 13.3. ■

Example 12.4.4

Height and Weight of Young Men We consider an idealized fictitious population of young men whose joint height and weight distribution fits the linear model exactly. For our fictitious population we will assume that the conditional means and SDs of weight given height are as follows:

$$\mu_{Y|X} = -145 + 4.25X$$
$$\sigma_\varepsilon = 20$$

Thus, the regression parameters of the population are $\beta_0 = -145$ and $\beta_1 = 4.25$. (This fictitious population resembles that of U.S. 17-year-olds.)[18] Thus, the model is $Y = -145 + 4.25X + \varepsilon$.

Table 12.4.1 shows the conditional means and SDs of Y = weight for a few selected values of X = height. Figure 12.4.1 shows the conditional distributions of Y given X for these selected subpopulations.

Table 12.4.1 Conditional means and SDs of weight given height in a population of young men*

Height (in) X	Mean weight (lb) $\mu_{Y\|X}$	Standard deviation of weights (lb) $\sigma_{Y\|X}$
64	127	20
68	144	20
72	161	20
76	178	20

*Note that all values of $\sigma_{Y|X}$ are the same; they equal $\sigma_\varepsilon = 20$.

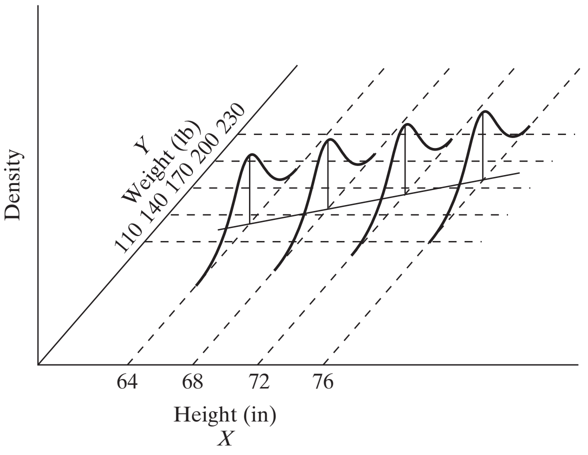


Figure 12.4.1 Conditional distributions of weight given height in a population of young men

Note, for example, that if height $= 68$ (in), then the mean weight is 144 (lb) and the SD of the weights is 20 (lb). For this subpopulation, $Y = 144 + \varepsilon$. If a particular young man who is 68 inches tall weighs 145 pounds, then $\varepsilon = 1$ for him. If another 68-inch-tall young man weighs 140 pounds, then $\varepsilon = -4$ in his case. Of course, β_0, β_1, and ε are generally not observable. This example is fictitious. ■

Remark. Actually, the term *regression* is not confined to linear regression. In general, the relationship between $\mu_{Y|X}$ and X is called the *regression of Y on X*. The linearity assumption asserts that the regression of Y on X is linear rather than, for instance, a curvilinear function.

Estimation in the Linear Model

Consider now the analysis of a set of (X, Y) data. Suppose we assume that the linear model is an adequate description of the true relationship of Y and X. Suppose further that we are willing to adopt the following **random subsampling model**:

Random Subsampling Model

For each observed pair (x, y), we regard the value y as having been sampled at random from the conditional population of Y values associated with the X value x.

Within the framework of the linear model and the random subsampling model, the quantities b_0, b_1, and s_e calculated from a regression analysis can be interpreted as estimates of population parameters:

b_0 is an estimate of β_0
b_1 is an estimate of β_1
s_e is an estimate of σ_ε

Example 12.4.5 Length and Weight of Snakes From the summaries of the snake data of Example 12.2.1 and 12.2.2, we can compute the following regression coefficients $b_0 = -301$, $b_1 = 7.19$, and $s_e = 12.5$ (computing these yourself from the provided summaries would be a good exercise). Thus,

-301 is our estimate of β_0
7.19 is our estimate of β_1
12.5 is our estimate of σ_ε ■

The application of the linear model to the snake data has yielded two benefits. First, the slope of the regression line, 7.19 gm/cm, is an estimate of a morphological parameter ("weight per unit length"), which is of potential biological interest in characterizing the population of snakes. Second, we have obtained an estimate (12.5 g) of the variability of weight among snakes of fixed length, even though no direct estimate of this variability was possible because no two of the observed snakes were the same length.

Interpolation in the Linear Model

In Section 12.3 we regarded the regression line as a line of averages. The idea of smoothing the graph of averages into a straight line can be extended to the setting in which we have only a single observation at each level of X. When we draw a line through a set of (X, Y) data, we are expressing a belief that the underlying dependence of Y on X is smooth, even though the data may show the relationship only roughly. Linear regression is one formal way of providing a smooth description of the data as illustrated in the following example.

Example 12.4.6

Arsenic in Rice What are the mean and standard deviation of arsenic concentrations in rice for plants with straw silicon concentrations of 33 g/kg? None of our observed plants had a straw silicon concentration of 33 g/kg. If there were some observations with this much silicon, we could average the associated arsenic concentrations to obtain one answer to our question, but because there is an apparent linear relationship between X and Y, we can use the line to obtain an even better estimate of the mean rice arsenic concentration that uses all of the data. In Example 12.3.4 we found the regression equation to be $\hat{y} = 197.17 - 2.51x$ and $s_e = 37.30$. Thus the estimated mean arsenic concentration for straw with 33 g/kg silicon is $197.17 - 2.51 \times 33 = 114.35\,\mu\text{g/kg}$ with a standard deviation of $s_e = 37.30\ \mu\text{g/kg}$. Figure 12.4.2 shows the interpolation graphically. ■

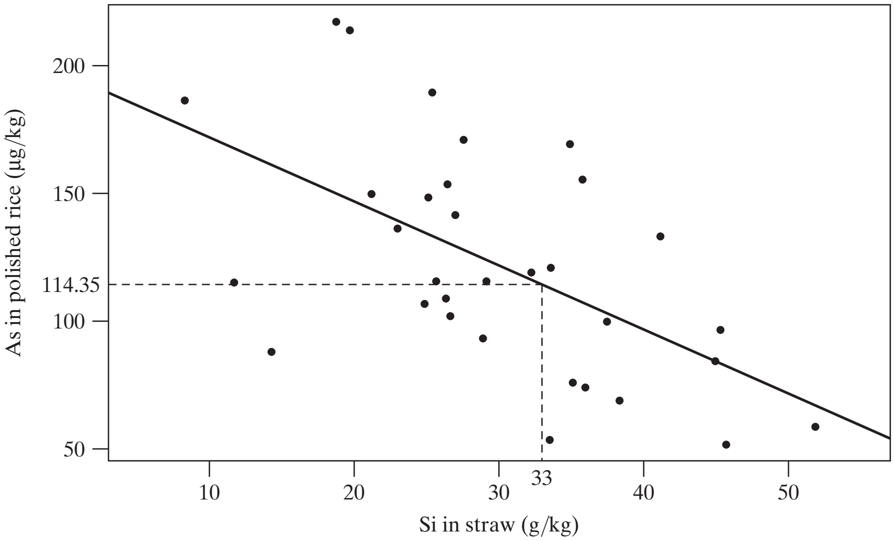


Figure 12.4.2 Concentrations of arsenic in rice versus silicon in straw for 32 rice plants

Note that estimation of the mean uses the linearity condition of the linear model, while estimation of the standard deviation uses the condition of constant standard deviation. In some situations only the linearity condition may be plausible, and then only the mean would be estimated.

Example 12.4.6 is an example of **interpolation**, because the X values we chose ($X = 33$ for the rice arsenic and 3.5 for the food consumption examples) were within the range of observed values of X. By contrast, **extrapolation** is the use of a regression line (or other curve) to predict Y for values of X that are outside the range of the data. Extrapolation should be avoided whenever possible, because there is usually no assurance that the relationship between $\mu_{Y|X}$ and X remains linear for X values outside the range of those observed. Many biological relationships are linear for only part of the possible range of X values. The following is an example.

Example 12.4.7 **Amphetamine and Food Consumption** The dose-response relationship for the rat food consumption experiment of Example 12.1.1 looks approximately like Figure 12.4.3.[19] The data cover only the linear portion of the relationship. Clearly it would be unwise to extrapolate the fitted line out to $X = 10$ or $X = 15$. ■

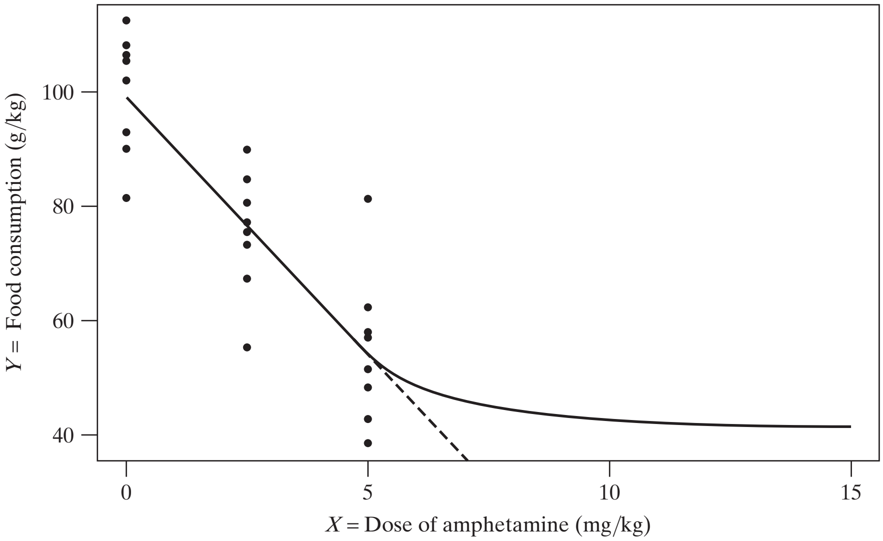


Figure 12.4.3 Dose-response curve (mean response versus dose) for rat food consumption experiment

Prediction and the Linear Model

Consider the setting of using height, X, to predict weight, Y, for a large group of young men for whom the average weight is 150 pounds. Suppose a young man is chosen at random and we must predict his weight.

1. If we don't know anything about the height of the man, then the best estimate we can give of his weight is the overall average weight, $\bar{y} = 150$.
2. Suppose we learn that the man's height is 76 inches. If we know that the average weight of all 76-inch-tall men in the group is 180 pounds, then we can use this conditional average, $\bar{y}|x = 76$, as our prediction of the man's weight. We expect this prediction, which essentially is using the graph of averages (but without smoothing), to be more accurate than the one given in part 1.
3. Suppose we learn that the man's height is 76 inches and we also know that the least-squares regression equation is $Y = -140 + 4.3X$. Then we can use the value $x = 76$ to get a prediction, which would be $-140 + 4.3 \times 76 = 186.8$.

Is the prediction in 3 better than the prediction made in 2? Since using the regression equation amounts to smoothing the graph of averages, we expect prediction 3 to be better than prediction 2 *to the extent that we believe that there is a linear relationship between height and weight.* Prediction 3 has the advantage of using information from all the data points, not just those for which $x = 76$. Method 3 also has the advantage of allowing for predictions when the x value (the height) is not one that is in the original data set (as discussed in the preceding subsection "Interpolation in the Linear Model"), so that $\bar{y}|x$ is not known. However, method 3 will give poor predictions if the linear relationship does not hold. Thus it is very important to think about such relationships, and to explore them graphically, before using a regression model.

Exercises 12.4.1–12.4.9

12.4.1 For the data in Exercise 12.2.6 there were two observations for which $X = 0$. The average response (Y value) for these points is $\frac{33.3 + 31.0}{2} = 32.15$. However, the intercept of the regression line, b_0, is not 32.15. Why not? Why is b_0 a better estimate of the average fungus growth when laetisaric acid concentration is zero than 32.15?

12.4.2 Refer to the body temperature data of Exercise 12.3.2. Assuming that the linear model is applicable, estimate the mean and the standard deviation of the drop in body temperature that would be observed in mice given alcohol at a dose of 2 gm/kg. [*Tip*: Is the X variable dose or log(dose)?]

12.4.3 Refer to the cob weight data of Exercises 12.2.5 and 12.3.3. Assume that the linear model holds.

(a) Estimate the mean cob weight to be expected in a plot containing (i) 100 plants; (ii) 120 plants.

(b) Assume that each plant produces one cob. How much grain would we expect to get from a plot containing (i) 100 plants? (ii) 120 plants?

12.4.4 (*Continuation of Exercise 12.4.3*). For the cob weight data, SS(resid) = 1,337.3. Estimate the standard deviation of cob weight in plots containing (i) 100 plants; (ii) 120 plants.

12.4.5 Refer to the fungus growth data of Exercise 12.2.6. For these data, SS(resid) = 16.7812. Assuming that the linear model is applicable, find estimates of the mean and standard deviation of fungus growth at a laetisaric acid concentration of 15 μg/ml.

12.4.6 Refer to the energy expenditure data of Exercise 12.2.7. Assuming that the linear model is applicable, estimate the 24-hour energy expenditure of a man whose fat-free mass is 55 kg.

12.4.7 Refer to the Ca pump activity of Exercise 12.2.8. For these data SS(resid) = 21,984,623.

(a) Assuming that the linear model is applicable, estimate the mean and standard deviation basal Ca pump activity for children born to mothers with a hair Hg level of 3 μg/g.

(b) Using the values computed in part (a) to support your answer, would it be surprising for a mother with a hair Hg level of 3 μg/g to give birth to a child with a basal Ca pump activity above 4000 nmol/mg/hr?

12.4.8 Refer to the bullfrog data of Exercise 12.3.7. Assuming that the linear model is applicable, estimate the maximum jump length of a bullfrog whose body length is 150 mm.

12.4.9 Refer to the peak flow data of Exercise 12.3.8. Assuming that the linear model is applicable, find estimates of the mean and standard deviation of peak flow for men 180 cm tall.

12.5 Statistical Inference Concerning β_1

The linear model provides interpretations of b_0, b_1, and s_e that take them beyond data description into the domain of statistical inference. In this section we consider inference about the true slope β_1 of the regression line. The methods are based on the condition that the conditional population distribution of Y for each value of X is a normal distribution. This is equivalent to stating that in the linear model of $Y = \beta_0 + \beta_1 X + \varepsilon$, the ε values come from a normal distribution.

The Standard Error of b_1

Within the context of the linear model, b_1 is an estimate of β_1. Like all estimates calculated from data, b_1 is subject to sampling error. The standard error of b_1 is calculated as follows:

Standard Error of b_1

$$\mathrm{SE}_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$$

The following example illustrates the calculation of SE_{b_1}.

Example 12.5.1 **Length and Weight of Snakes** For the snake data, we found in Table 12.2.2 that $n = 9$, $s_x = 4.637$, and in Example 12.4.5 that $s_e = 12.5$. The standard error of b_1 is

$$\mathrm{SE}_{b_1} = \frac{12.5}{4.637\sqrt{9-1}} = 0.9531$$

To summarize, the slope of the fitted regression line (from Example 12.4.5) is

$$b_1 = 7.19 \text{ gm/cm}$$

and the standard error of this slope is

$$\mathrm{SE}_{b_1} = 0.95 \text{ gm/cm}$$

■

Structure of the SE. Let us see how the standard error of b_1 depends on various aspects of the data. In the same way that $\mathrm{SE}_{\bar{Y}}$ depends on the variability in the Y data (s_y) and the sample size (n), SE_{b_1} depends on the scatter of the data about the regression line (s_e) and the size of the sample (n). The formula for SE_{b_1} supports our intuition showing that data with less scatter about the regression line (smaller s_e) and larger sample sizes (larger n) produce more precise estimates of β_1 (i.e., a smaller SE_{b_1}). While variability in Y and sample size are the only two factors that affect our ability to estimate a population mean precisely ($\mathrm{SE}_{\bar{Y}}$), there is a third factor that is important for precise estimation of β_1: the variability of the X data. The more spread out our X values (larger s_x), the more precise our estimate of β_1 will be. The dependence on the spread in the X values is illustrated in Figure 12.5.1, which shows two data sets with the same value of s_e and the same value of n, but different values of s_x. Imagine using a ruler to fit a straight line by eye; it is intuitively clear that the data in case (b)—with the larger s_x—would determine the slope of the line more precisely.

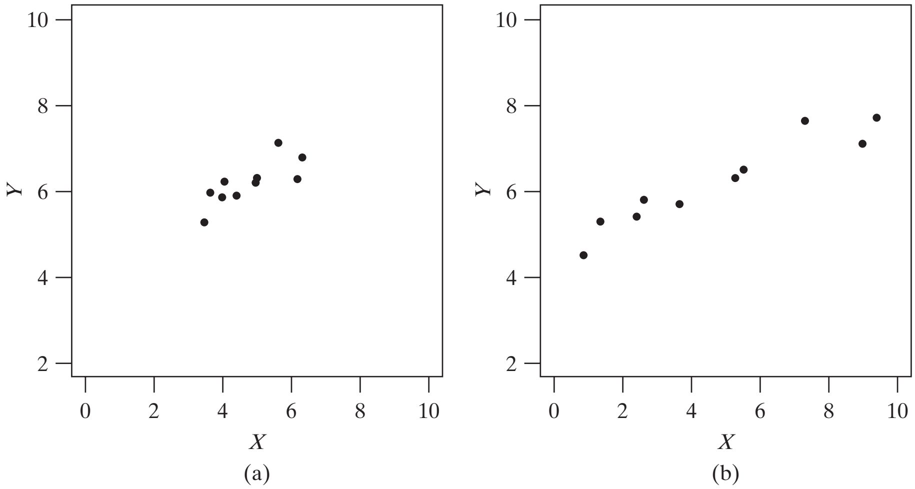


Figure 12.5.1 Two data sets with the same value of n and of s_e but different s_x: (a) smaller s_x and (b) larger s_x

As another way of thinking about this, imagine holding your arms out in front of you, extending the index finger on each hand, and balancing a meter stick on your two fingers. If you move your hands far apart from each other, balancing the meter stick is easy—this is like case (b). However, if you move your hands close together, balancing the meter stick becomes more difficult—this is like case (a). Having the base of support spread out increases stability. Likewise, having the x values spread out decreases the standard error of the slope.

Implications for Design. The previous discussion implies that, for the purpose of gaining precise information about β_1, it is best to have the values of X as widely dispersed as possible. This fact can guide the experimenter when the design of the experiment includes choosing values of X. Other factors also play a role, however. For instance, if X is the dose of a drug, the criterion of widely dispersed X's would lead to using only two dosages, one very low and one very high. But in practice an experimenter would want to have at least a few observations at intermediate doses, to verify that the relation is actually linear within the range of the data.

Confidence Interval for β_1

In many studies the quantity β_1 is a biologically meaningful parameter and a primary aim of the data analysis is to estimate β_1. A confidence interval for β_1 can be constructed by the familiar method based on the SE and Student's t distribution. For instance, a 95% confidence interval is constructed as

$$b_1 \pm t_{0.025}\,\mathrm{SE}_{b_1}$$

where the critical value $t_{0.025}$ is determined from Student's t distribution with

$$\mathrm{df} = n - 2$$

Intervals with other confidence coefficients are constructed analogously; for instance, for a 90% confidence interval one would use $t_{0.05}$.

Example 12.5.2

Length and Weight of Snakes Let us use the snake data to construct a 95% confidence interval for β_1. We found that $b_1 = 7.19186$ and $\mathrm{SE}_{b_1} = 0.9531$. There are $n = 9$ observations; we refer to Table 4 with $\mathrm{df} = 9 - 2 = 7$, and obtain

$$t_{7,0.025} = 2.365$$

The confidence interval is

$$7.19186 \pm 2.365 \times 0.9531$$

or

$$4.94 \text{ gm/cm} < \beta_1 < 9.45 \text{ gm/cm}$$

We are 95% confident that the true slope of the regression of weight on length for this snake population is between 4.94 gm/cm and 9.45 gm/cm; this is a rather wide interval because the sample size is not very large. ■

Testing the Hypothesis $H_0\colon \beta_1 = 0$

In some investigations it is not a foregone conclusion that there is any linear relationship between X and Y. It then may be relevant to consider the possibility that any apparent trend in the data is illusory and reflects only sampling variability. In this situation it is natural to formulate the null hypothesis

$$H_0\colon \mu_{Y|X} \text{ does not depend on } X$$

Within the linear model, this hypothesis can be translated as

$$H_0\colon .\beta_1 = 0$$

A t test of H_0 is based on the test statistic*

$$t_s = \frac{b_1 - 0}{\mathrm{SE}_{b_1}}$$

Critical values are obtained from Student's t distribution with

$$\mathrm{df} = n - 2$$

The following example illustrates the application of this t test.

Example 12.5.3

Blood Pressure and Platelet Calcium The blood pressure and platelet calcium data from Example 12.2.3 are shown in Figure 12.5.2. Calculations from the data yield $\bar{x} = 94.50000$, $\bar{y} = 107.86840$, $s_x = 8.04968$, $s_y = 16.07780$, from which we can calculate†

$$b_0 = -2.2009 \text{ and } b_1 = 1.16475$$

The residual sum of squares is 6311.7618.

Thus,

$$s_e = \sqrt{\frac{6311.76}{38 - 2}} = 13.24 \text{ and } \mathrm{SE}_{b_1} = \frac{13.24}{8.04968\sqrt{38 - 1}} = 0.2704$$

The values of b_0, b_1, SS(resid), and SE_{b_1} are generally found using computer software. The following computer output is typical:

```
The regression equation is
Platelet Calcium = −2.2 + 1.16 Blood Pressure
Predictor          Coef   SE Coef       T       P
Constant          −2.20     25.65   −0.09   0.932
Blood Pressure   1.1648    0.2704    4.31   0.000
S = 13.2411   R − Sq = 34.0%   R − Sq(adj) = 32.2%

Analysis of Variance
Source           DF      SS      MS       F       P
Regression        1  3252.6  3252.6   18.55   0.000
Residual Error   36  6311.8   175.3
Total            37  9564.3
```

We will test the null hypothesis

$$H_0: \beta_1 = 0$$

against the nondirectional alternative

$$H_A: \beta_1 \neq 0$$

*We include the "−0" in the numerator of the test statistic to remind us that we are comparing our estimated (observed) slope, b_1, to the slope we'd expect to observe if the null hypothesis were true. In the exercises we will consider a situation for which the hypothesized slope may be a value other than zero.

†As the following values are intermediate calculations used in the regression, we include more digits than one would typically display in a summary.

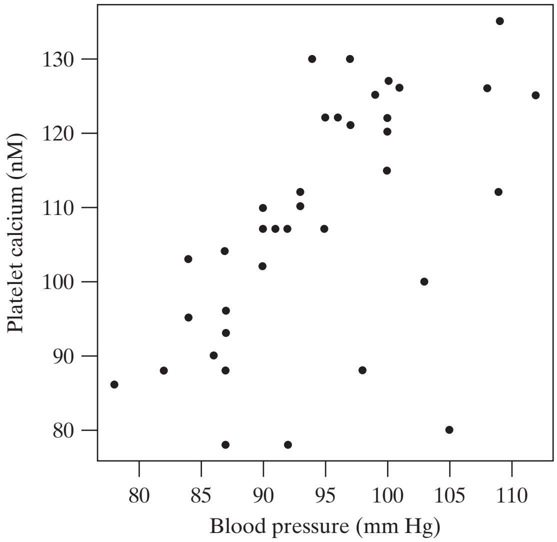


Figure 12.5.2 Blood pressure and platelet calcium for 38 persons with normal blood pressure

These hypotheses are translations, within the linear model, of the verbal hypotheses

H_0: Mean platelet calcium is not linearly related to blood pressure

H_A: Mean platelet calcium is linearly related to blood pressure

(*Note*: "Linearly related" does not necessarily refer to causal dependence as we have discussed in Section 12.2.)

Let us choose $\alpha = 0.05$. The test statistic is

$$t_s = \frac{1.16475}{0.2704} = 4.308$$

From Table 4 with df $= n - 2 = 36 \approx 40$, we find $t_{40,0.0005} = 3.551$. Thus, we find P-value < 0.001 and we reject H_0. The data provide sufficient (and very strong) evidence to conclude that the true slope of the regression of platelet calcium on blood pressure in this population is positive (that is, $\beta_1 > 0$). ■

Note that the test on β_1 does not ask *whether* the relationship between $\mu_{Y|X}$ and X is linear. Rather, the test asks whether, *assuming* that the linear model holds, we can conclude that the slope is nonzero. It is therefore necessary to be careful in phrasing the conclusion from this test. For instance, the statement "There is a significant linear trend" could easily be misunderstood.*

As is the case with other hypothesis tests, if we wish to use a directional alternative hypothesis we follow the two-step procedure of (1) checking that the specified direction is correct (which in a regression setting means checking that the slope of the regression line has the correct + or − sign) and (2) cutting the nondirectional P-value in half if this condition is met.

*There are tests that can (in some circumstances) test whether the true relationship is linear. Furthermore, there are tests that can test for a linear component of trend without assuming that the relationship is linear. These tests are beyond the scope of this book.

Exercises 12.5.1–12.5.9

12.5.1 Refer to the leucine data given in Exercise 12.3.1.

(a) Construct a 95% confidence interval for β_1.

(b) Interpret the confidence interval from part (a) in the context of this setting.

12.5.2 Refer to the body temperature data of Exercise 12.3.2. For these data, $s_e = 0.91472$. Construct a 95% confidence interval for β_1.

12.5.3 Refer to the cob weight data of Exercise 12.2.5. For these data, SS(resid) = 1,337.3.

(a) Construct a 95% confidence interval for β_1.

(b) Interpret the confidence interval from part (a) in the context of this setting.

12.5.4 Refer to the fungus growth data of Exercise 12.2.6. For these data, SS(resid) = 16.7812.

(a) Calculate the standard error of the slope, SE_{b_1}.

(b) Consider the null hypothesis that laetisaric acid has no effect on growth of the fungus. Assuming that the linear model is applicable, formulate this as a hypothesis about the true regression line, and test the hypothesis against the alternative that laetisaric acid inhibits growth of the fungus. Let $\alpha = 0.05$.

12.5.5 Refer to the energy expenditure data of Exercise 12.2.7. For these data, SS(resid) = 21,026.1.

(a) Construct a 95% confidence interval for β_1.

(b) Construct a 90% confidence interval for β_1.

12.5.6 Refer to the basal Ca pump data from Exercise 12.2.8. For these data, $s_e = 548.78$.

(a) Construct a 95% confidence interval for β_1.

(b) What do you think about a claim that that β_1 is less than −800 (nmol/mg/hr)/(μg/g)? Use your interval from part (a) to support your answer.

(c) What do you think about a claim that β_1 is less than 800 (nmol/mg/hr)/(μg/g) in magnitude? Use your interval from part (a) to support your answer.

12.5.7 Refer to the respiration data of Exercise 12.3.6. Assuming that the linear model is applicable, test the null hypothesis of no relationship against the alternative that trees from higher altitudes tend to have higher respiration rates. Let $\alpha = 0.05$.

12.5.8 The following computer output is from fitting a regression model to the snake length data of Example 12.2.2. Use this output to construct a 95% confidence interval for β_1.

```
The regression equation is
Weight = −301 + 7.19 Length

Predictor     Coef   Stdev  t-ratio      p
Constant   −301.09   60.19    −5.00  0.000
Length      7.1919  0.9531     7.55  0.000

s = 12.50 R−sq = 89.1%  R−sq(adj) = 87.5%

Analysis of Variance

SOURCE      DF  SS       MS       F      p
Regression  1   8896.3  8896.3  56.94  0.000
Error       7   1093.7   156.2
Total       8   9990.0
```

12.5.9 Refer to the peak flow data of Exercise 12.3.8. Assume that the linear model is applicable.

(a) Test the null hypothesis of no relationship against the alternative that peak flow is related to height. Use a nondirectional alternative with $\alpha = 0.10$.

(b) Repeat the test from part (a), but this time use the directional alternative that peak flow tends to increase with height. Again let $\alpha = 0.10$.

12.6 Guidelines for Interpreting Regression and Correlation

Any set of (X, Y) data can be submitted to a regression analysis and values of b_0, b_1, s_e, and r can be calculated. But these quantities require care in interpretation. In this section we discuss guidelines and cautions for interpretation of linear regression and correlation. We first consider the use of regression and correlation for purely descriptive purposes and then turn to inferential uses.

When Is Linear Regression Descriptively Inadequate?

Linear regression and correlation may provide inadequate description of a data set if any of the following features is present:

- curvilinearity
- outliers
- influential points

We briefly discuss each of these.

If the dependence of Y on X is actually curvilinear rather than linear, the application of linear regression and correlation can be very misleading. The following example shows how this can happen.

Example 12.6.1

A Curvilinear Relationship with X Figure 12.6.1 shows a set of fictitious data that obeys an exact relationship: $Y = -1 + 6X - X^2$. Nevertheless, X and Y are uncorrelated: $r = 0$ and $b_1 = 0$. The best straight line through the data would be a horizontal one, but of course the line would be a poor summary of the curvilinear relationship between X and Y. The residual SD is $s_e = 2.27$; however, since these data are nonrandom, s_e does not measure random variation, but rather measures deviation from linearity. ■

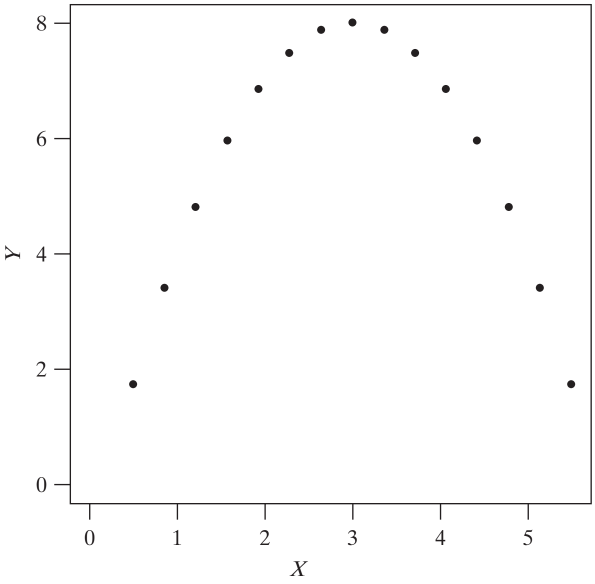


Figure 12.6.1 Data for which X and Y are uncorrelated but have a strong curvilinear relationship

Generally, the consequences of curvilinearity are that (1) the fitted line does not adequately represent the data; (2) the correlation is misleadingly small; (3) s_e is inflated. Of course, Example 12.6.1 is an extreme case of this distortion. A data set with mild, but still noticeable, curvilinearity is shown in Figure 12.6.2.

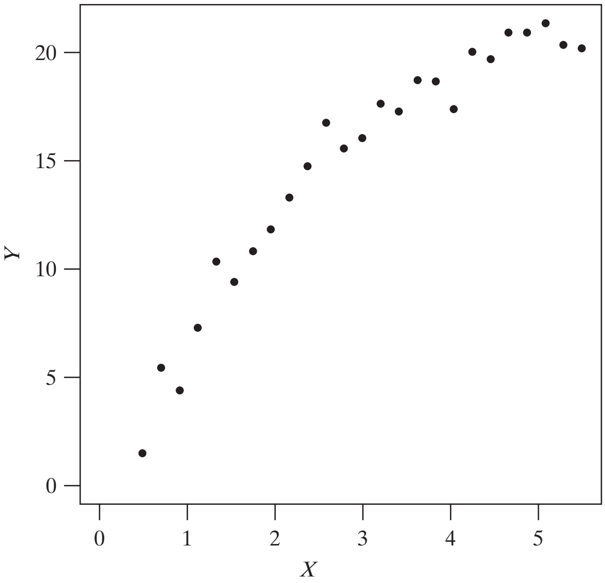


Figure 12.6.2 Data displaying mild curvilinearity

Outliers in a regression setting are data points that are unusually far from the linear trend formed by the data. Outliers can distort regression analysis in two ways: (1) by inflating s_e and reducing correlation; and (2) by unduly influencing the regression line. Note that a point can be an outlier in a scatterplot without being an outlier in either the distribution of X values or the distribution of Y values as we shall see in the following example.

Figure 12.6.3 displays a data set with a variety of outliers. Figure 12.6.3(a) displays a data set with no outliers, while (b) and (c) show data with regression outliers—they have points that fall far from the regression line. In plot (b) the outlying point does not appear to affect the slope of the regression line very much, but it does increase the residual standard deviation, s_e, and reduce correlation. The outlying point in plot (c) appears to greatly affect the slope of the estimated regression line; it also increases s_e and reduces the correlation. While the unusual point in plot (d) is an outlier with respect to the X (and Y) distribution, it is not an outlier in the regression context as it does not fall far from the regression line.

Leverage points are points that have the potential to greatly influence the slope of the fitted regression model. The further a point's X value is from the center of the X distribution, the more leverage that point has on the overall regression model. *Having* and actually *exerting* leverage are two different things, however. Figure 12.6.3 plots (c) and (d) display examples of leverage points. In plot (c) the leverage point is shown to actually exert its leverage on the line, tipping the regression from the bulk of the data. A point that has a large effect on the regression model is called an **influential point**. Plot (d) shows a leverage point (because of the extreme X value) that is not influential because the regression line does not get pulled away from the trend in the bulk of the data. Note that the outlier in plot (b) is not consid-

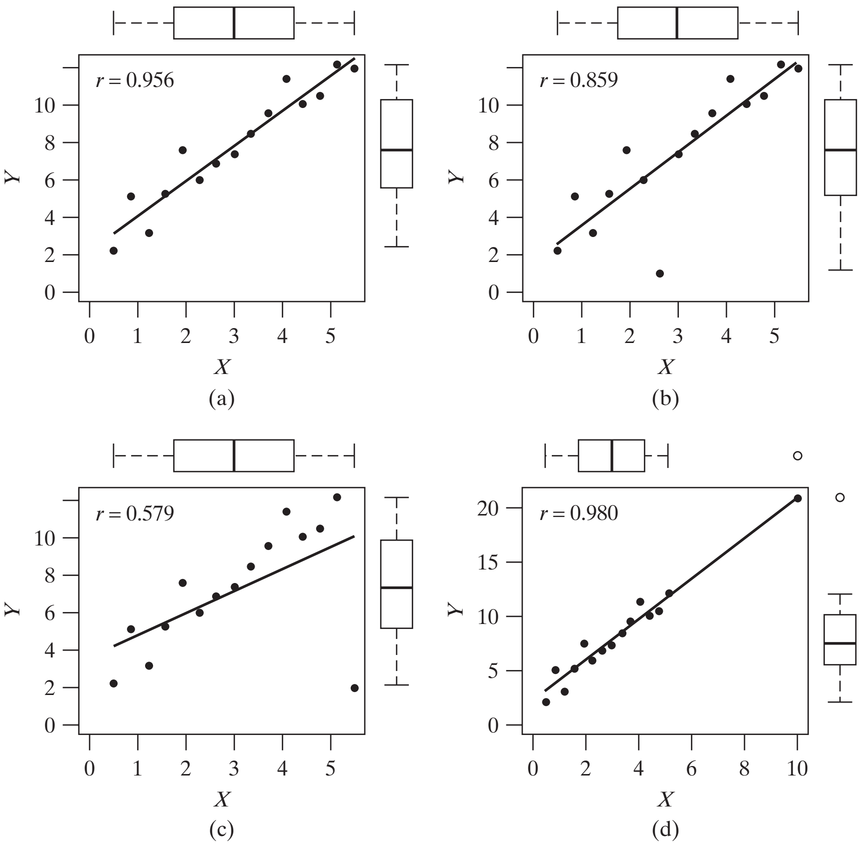


Figure 12.6.3 Different effects of outliers on the regression line. Boxplots of the X and Y data appear in the margins of each scatterplot. (a) A data set with no outliers; (b) the same data except for one outlier in the middle of the X values; (c) the same data except for one outlier at the high end of the X values (a point with leverage and influence); and (d) the same data except for one point that is an outlier with respect to the X (and Y) distribution, but not with respect to the regression line (a point with leverage, but little influence)

ered a leverage point—its ability to affect the slope of the line is weak as its *X* value is near the center of the *X* distribution.

Influential points can also greatly affect (increase or decrease) the size of the correlation coefficient. In Figure 12.6.3, the influential point in (c) lowered the correlation from $r = 0.956$ in (a) to $r = 0.579$. Example 12.6.3 shows a situation for which the correlation is increased by the presence of an influential point.

Figure 12.6.4 (a) shows a data set and a regression line. Figure 12.6.4 (b) shows the same data set, but with an influential point added. Including the influential point in the data set changes the regression line noticeably. Although the influential point is an outlier in the *X* and *Y* distributions, it is not a regression outlier since the residual for this point is not very large.

The correlation coefficient for the data in Figure 12.6.4(a) is $r = 0.053$. Adding the influential point to the data set changes the correlation to $r = 0.759$ for the data in Figure 12.6.4(b).

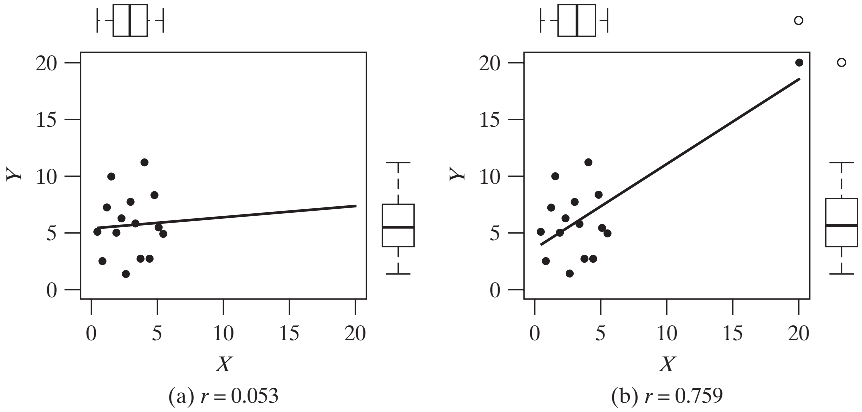


Figure 12.6.4 The effect of an influential point on the regression line. (a) A data set; (b) the same data with an influential point added

Conditions for Inference

The quantities b_0, b_1, s_e, and r can be used to describe a scatterplot that shows a linear trend. However, statistical inference based on these quantities depends on certain conditions concerning the design of the study, the parameters, and the conditional population distributions. We summarize these conditions and then discuss guidelines and cautions concerning them.

1. *Design conditions.* We have discussed two sampling models for regression and correlation:

   (a) Random subsampling model: For each observed *X*, the corresponding observed *Y* is viewed as randomly chosen from the conditional population distribution of *Y* values for that *X*.*

   (b) Bivariate random sampling model: Each observed pair (*X*, *Y*) is viewed as randomly chosen from the joint population distribution of bivariate pairs (*X*, *Y*).

   In either sampling model, each observed pair (*X*, *Y*) must be independent of the others. This means that the experimental design must not include any pairing, blocking, or hierarchical structure.

*If the *X* variable includes measurement error, then *X* in the linear model must be interpreted as the measured value of *X* rather than some underlying "true" value of *X*. A linear model involving the "true" value of *X* leads to a different kind of regression analysis.

2. *Conditions concerning parameters.* The linear model states that
   (a) $\mu_{Y|X} = \beta_0 + \beta_1 X$.
   (b) σ_e does not depend on X.
3. *Condition concerning population distributions.* The confidence interval and t test are based on the conditional population distribution of Y for each fixed X having a normal distribution.

The random subsampling model is required if b_0, b_1, and s_e are to be viewed as estimates of the parameters β_0, β_1, and σ_ε mentioned in the linear model. The bivariate random sampling model is required if r is to be viewed as an estimate of a population parameter ρ. It can be shown that if the bivariate random sampling model is applicable, then the random subsampling model is also applicable. Thus, regression parameters can always be estimated if correlation can be estimated, but not vice versa.

Guidelines Concerning the Sampling Conditions

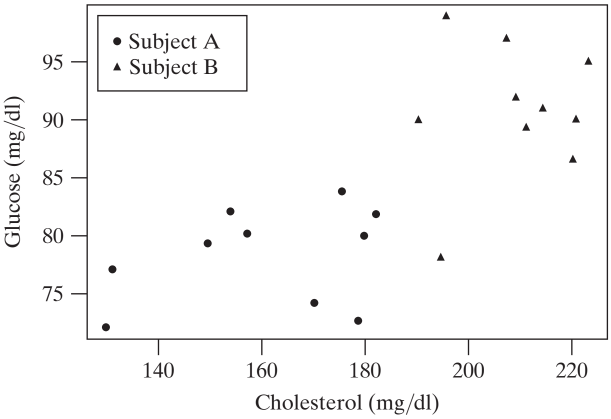
Departures from the sampling conditions not only affect the validity of formal techniques such as the confidence interval for β_1, but can also lead to faulty interpretation of the data even if no formal statistical analysis is performed. Two errors of interpretation that sometimes occur in practice are (1) failure to take into account dependency in the observations, and (2) insufficient caution in interpreting r when the X's do not represent a random sample.

The following two examples illustrate studies with dependent observations.

Example 12.6.2

Serum Cholesterol and Serum Glucose A data set consists of 20 pairs of measurements on serum cholesterol (X) and serum glucose (Y) in humans. However, the experiment included only two subjects; each subject was measured on 10 different occasions. Because of the dependency in the data, it is not correct to naively treat all 20 data points alike. Figure 12.6.5 illustrates the difficulty; the figure shows that there is no evidence of any correlation between X and Y, except for the modest fact that the subject who has larger X values happens also to have larger Y values. Clearly it would be impossible to properly interpret the scatterplot if all 20 points were plotted with the same symbol. By the same token, application of regression or correlation formulas to the 20 observations would be seriously misleading.[20] ■

Figure 12.6.5 Twenty observations of X = serum cholesterol and Y = serum glucose in humans



Example 12.6.3

Growth of Beef Steers Figure 12.6.6 shows 20 pairs of measurements on the weight (Y) of beef steers at various times (X) during a feeding trial. The data represent four animals, each weighed at five different times; observations on the same animal are joined by lines in the figure. An ordinary regression analysis on the 20 data points

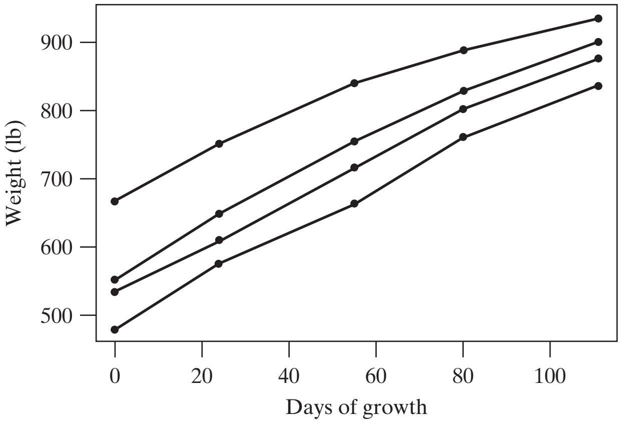


Figure 12.6.6 Twenty observations of X = days and Y = weight in steers. Data for individual animals are joined by lines

would ignore the information carried in the lines and would yield inflated SEs and weak tests. Similarly, an ordinary scatterplot (without the lines) would be an inadequate representation of the data.[21] ■

In Example 12.6.2, ignoring the dependency in the observations would lead to *overinterpretation* of the data—that is, concluding that a relationship exists when there is actually very little evidence for it. By contrast, ignoring the dependency in Example 12.6.3 would lead to *underinterpretation* of the data—that is, insufficiently extracting the "signal" from the "noise."

In interpreting the correlation coefficient r, one should recognize that r is influenced by the degree of spread in the values of X. If the regression quantities b_0, b_1, and s_e are unchanged, *more spread in the X values leads to a stronger correlation (larger magnitude of r)*. The following example shows how this happens.

Example 12.6.4

Figure 12.6.7 shows fictitious data that illustrate how r can be affected by the distribution of X. The data points in parts (a) and (b) have been plotted together in part (c). The regression line is nearly the same in all three scatterplots, but notice that X and Y appear more highly correlated in (c) than in either (a) or (b). The contrasting appearance of the scatterplots is reflected in the correlation coefficients; in fact, $r = 0.60$ for (a), $r = 0.58$ for (b), but $r = 0.85$ for (c). ■

The fact that r depends on the distribution of X does not mean that r is invalid as a descriptive statistic. But it does mean that, when the values of X cannot be viewed as a random sample, r must be interpreted cautiously. For instance, suppose two

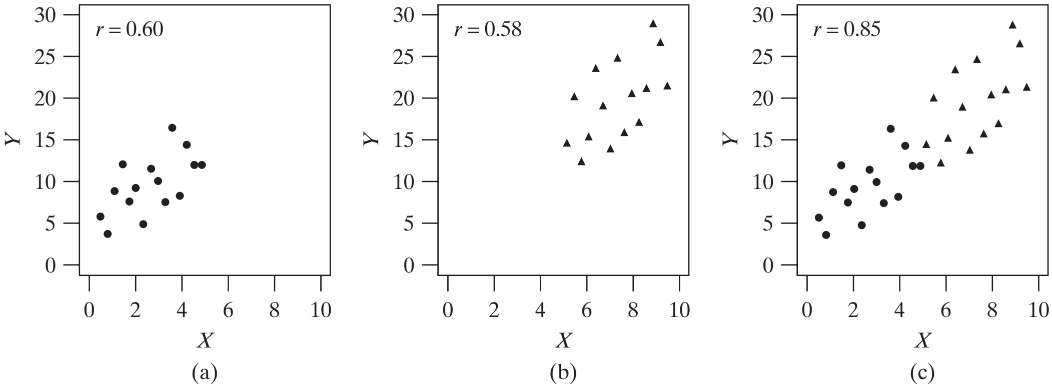


Figure 12.6.7 Dependence of r on the distribution of X. The data of (a) and (b) are plotted together in (c)

experimenters conduct separate studies of response (Y) to various doses (X) of a drug. Each of them could calculate r as a description of her or his own data, but they should *not* expect to obtain *similar* values of r unless they both use the same choice of doses (X values). By contrast, they might reasonably expect to obtain similar regression lines and similar residual standard deviations, regardless of their choice of X values, as long as the dose-response relationship remains the same throughout the range of doses used.

Labeling X and Y. If the bivariate random sampling model is applicable, then the investigator is free to decide which variable to label X and which to label Y. Of course, for calculation of r the labeling does not matter. For regression calculations, the decision depends on the purpose of the analysis. The regression of Y on X yields (within the linear model) estimates of $\mu_{Y|X}$—that is, the population mean Y value for fixed X. Similarly, the regression of X on Y is aimed at estimating $\mu_{X|Y}$—that is, the mean X value for fixed Y. These approaches do not lead to the same regression line because they are directed at answering different questions. An intuitive example follows.

Example 12.6.5

Height and Weight of Young Men For the population of young men described in Example 12.4.4, the mean weight of young men 76" (6'4") tall is 178 lb. Now consider this question: What would be the mean height of young men who weigh 178 lb? There is no reason that the answer should be 76". Intuition suggests that the answer should be less than 76"—and in fact it is about 71". ■

Guidelines Concerning the Linear Model and Normality Condition

The test and confidence interval for β_1 are based on the linear model and the condition of normality. The interpretation of these inferences can be seriously degraded if the linearity condition is not met; after all, we have seen earlier in this section that even the descriptive usefulness of regression is reduced if curvilinearity or outliers are present.

In addition to linearity, the linear model specifies that σ_ε is the same for all the observations. A common pattern of departure from this condition is a trend for larger means to be associated with larger SDs. Mild nonconstancy of the SDs does not seriously affect the interpretation of b_0, b_1, SE_{b_1}, and r (although it does invalidate the interpretation of s_e as a pooled estimate of a common SD).

Residual Plots

Formal statistical tests for curvilinearity, unequal standard deviations, nonnormality, and outliers are beyond the scope of this book. However, the single most useful instrument for detecting these features is the human eye, aided by scatterplots. For instance, notice how easily the eye detects the mild curvilinearity in Figure 12.6.2 and the outlier in Figure 12.6.3(b). Notice also in Figure 12.6.3(b) that examination of the marginal distributions of X and Y separately would not have revealed the outlier.

In addition to scatterplots of Y versus X, it is often useful to look at various displays of the residuals. A scatterplot of each residual $(y_i - \hat{y}_i)$ against $\hat{y}_i$ is called a **residual plot**. Residual plots are very useful for detecting curvature; they can also reveal trends in the conditional standard deviation. Figure 12.6.8 shows the data from Figure 12.6.2 together with a residual plot of those data.

A residual plot shows the data after the linear trend has been removed, which makes it easier to see nonlinear patterns in the data. The curvature in Figure 12.6.8(a) is apparent, but it is much more visible in the residual plot of Figure 12.6.8(b).

If the linear model holds, with no outliers, then the fitted regression line captures the trend in the data, leaving a random pattern in the residual plot. Thus, *we*

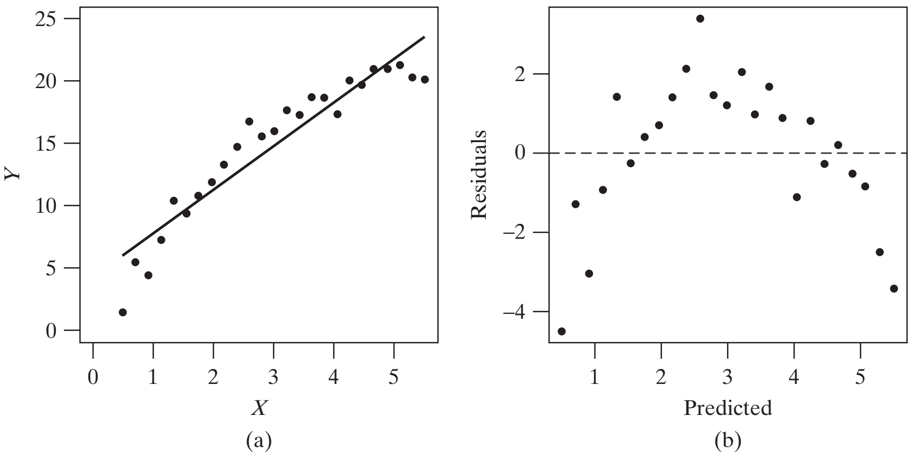


Figure 12.6.8 (a) Data displaying mild curvilinearity with linear regression line; (b) a residual plot of the data

hope to see no striking pattern in a residual plot. For example, Figure 12.6.9 shows a residual plot of the snake data of Example 12.2.1. The lack of unusual features in this plot supports the use of a regression model for these data.

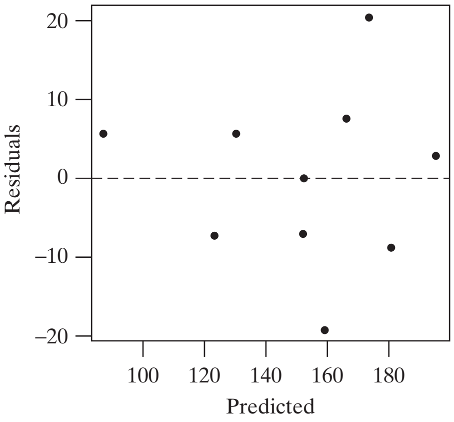


Figure 12.6.9 Residual plot of the snake data

If the condition of normality is met, then the distribution of the residuals should look roughly like a normal distribution.* A normal probability plot of the residuals provides a useful check of the normality condition. The normal probability plot of the snake data in Figure 12.6.10 is fairly linear, which supports the use of the t test and the confidence interval presented in Section 12.5.

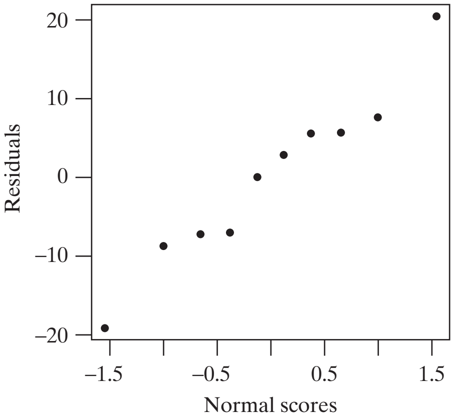


Figure 12.6.10 Normal probability plot of the snake data

*This is the basis for the 68% and 95% interpretations of s_e given in Section 12.3.

The Use of Transformations

If the conditions of linearity, constancy of standard deviation, and normality are not met, a remedy that is sometimes useful is to transform the scale of measurement of either Y, or X, or both. The following example illustrates the use of a logarithmic transformation.

Example 12.6.6

Growth of Soybeans A botanist placed 60 one-week-old soybean seedlings in individual pots. After 12 days of growth, she harvested, dried, and weighed 12 of the young soybean plants. She weighed another 12 plants after 23 days of growth, and groups of 12 plants each after 27 days, 31 days, and 34 days. Figure 12.6.11 shows the 60 plant weights plotted against days of growth; a smooth curve connects the group means. It is easy to see from Figure 12.6.11 that the relationship between mean plant weight and time is curvilinear rather than linear and that the conditional standard deviation is not constant but is strongly increasing.[22]

Figure 12.6.12 shows the logarithms (base 10) of the plant weights, plotted against days of growth together with the regression line. Notice that the logarithmic

Figure 12.6.11 Weight of soybean plants plotted against days of growth

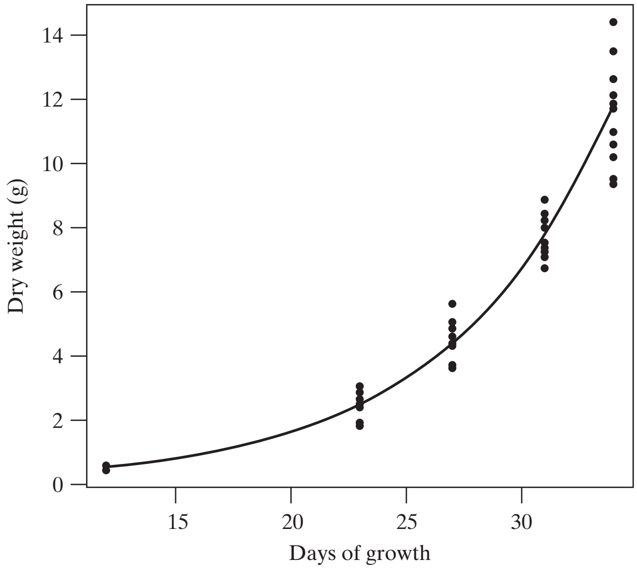
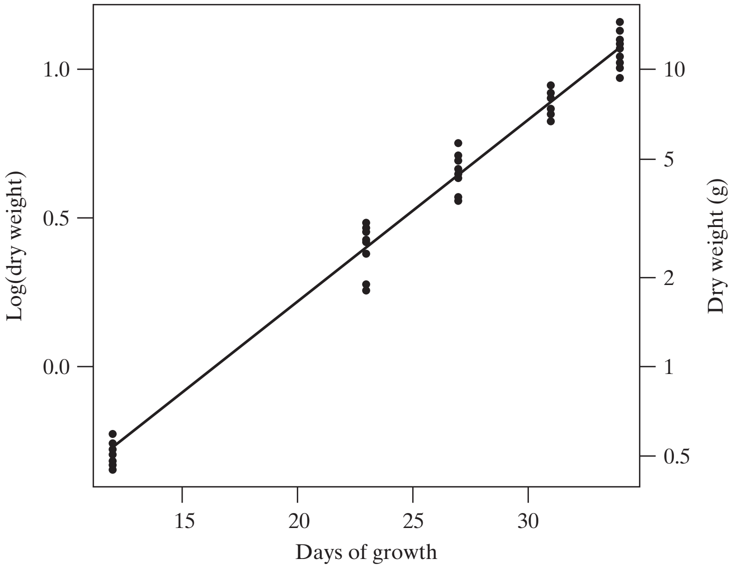


Figure 12.6.12 Log(weight) of soybean plants plotted against days of growth



transformation has simultaneously straightened the curve and more nearly equalized the standard deviations. It would not be unreasonable to assume that the linear model is valid for the variables Y = log(dry weight) and X = days of growth. Table 12.6.1 shows the means and standard deviations before and after the logarithmic transformation. Note especially the effect of the transformation on the equality of the SD*s*. ■

Table 12.6.1 Summary of soybean growth data in original scale and after log transformation

		Dry weight (gm)		Log(dry weight)	
Days of growth	Number of plants	Mean	SD	Mean	SD
12	12	0.50	0.06	−0.31	0.055
23	12	2.63	0.37	0.42	0.062
27	12	4.67	0.70	0.67	0.066
31	12	7.57	1.19	0.87	0.069
34	12	11.20	1.62	1.04	0.064

Exercises 12.6.1–12.6.9

12.6.1 In a metabolic study, four male swine were tested three times: when they weighed 30 kg, again when they weighed 60 kg, and again when they weighed 90 kg. During each test, the experimenter analyzed feed intake and fecal and urinary output for 15 days, and from these data calculated the nitrogen balance, which is defined as the amount of nitrogen incorporated into body tissue per day. The results are shown in the accompanying table.[23]

	NITROGEN BALANCE (gm/day)		
ANIMAL NUMBER	BODY WEIGHT 30 kg	60 kg	90 kg
1	15.8	21.3	16.5
2	16.4	20.8	18.2
3	17.3	23.8	17.8
4	16.4	22.1	17.5
Mean	16.48	22.00	17.50

Suppose these data are analyzed by linear regression. With X = body weight and Y = nitrogen balance, preliminary calculations yield $\bar{x} = 60$ and $\bar{y} = 18.7$. The slope is $b_1 = 0.017$, with standard error $\mathrm{SE}_{b_1} = 0.032$. The t statistic is $t_s = 0.53$, which is not significant at any reasonable significance level. According to this analysis, there is insufficient evidence to conclude that nitrogen balance depends on body weight under the conditions of this study.

The above analysis is flawed in two ways. What are they? (*Hint*: Look for ways in which the conditions for inference are not met. There may be several minor departures from the conditions, but you are asked to find two major ones. No calculation is required.)

12.6.2 For measuring the digestibility of forage plants, two methods can be used: The plant material can be fermented with digestive fluids in a glass container, or it can be fed to an animal. In either case, digestibility is expressed as the percentage of total dry matter that is digested. Two investigators conducted separate studies to compare the methods by submitting various types of forage to both methods and comparing the results. Investigator A reported a correlation of $r = 0.8$ between the digestibility values obtained by the two methods, and investigator B reported $r = 0.3$. The apparent discrepancy between these results was resolved when it was noted that one of the investigators had tested only varieties of canary grass (whose digestibilities ranged from 56% to 65%), whereas the other investigator had used a much wider spectrum of plants, with digestibilities ranging from 35% for corn stalks to 72% for timothy hay.[24]

Which investigator (A or B) used only canary grass? How does the different choice of test material explain the discrepancy between the correlation coefficients?

12.6.3 Refer to the energy expenditure data of Exercise 12.2.7. Each subject's expenditure value (Y) is the average of two measurements made on different occasions. It might be proposed that it would be better to use the two measurements as separate data points, thus yielding 14 observations rather than 7. If this proposed approach were used, one of the conditions for inference would be highly doubtful. Which one, and why?

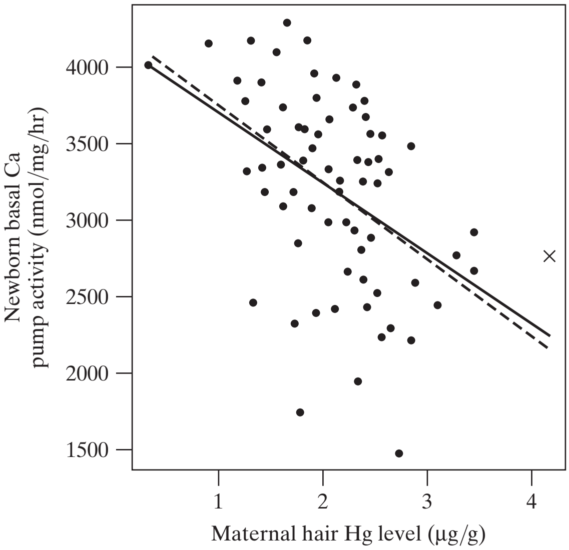
12.6.4 Refer to the fungus growth data of Exercise 12.2.6. In that exercise the investigator found $r = -0.98754$. Suppose a second investigator were to replicate the experiment, using concentrations of 0, 2, 4, 6, 8, and 10 mg, with two petri dishes at each concentration. Would you predict that the value of r calculated by this second investigator would be about the same as that found in Exercise 12.2.6, smaller in magnitude, or larger in magnitude? Explain.

12.6.5 In the following scatterplot of the Ca pump data of Exercise 12.2.8, one of the points is marked with an "×." In addition, there are two regression lines on the plot: The solid line includes all of the data and the dashed line omits the point marked "×."

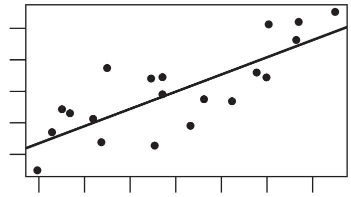
(a) Would we consider the point marked "×" an outlier? Explain.

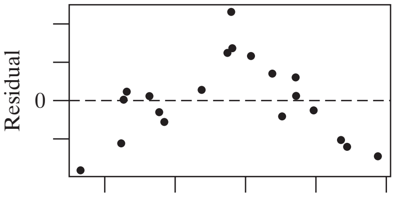
(b) Would we consider the point marked "×" a leverage point? Explain.

(c) Noting the very small change in the slopes of the dashed and solid lines, would we consider the point marked "×" an influential observation? Explain.

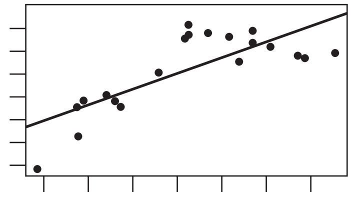


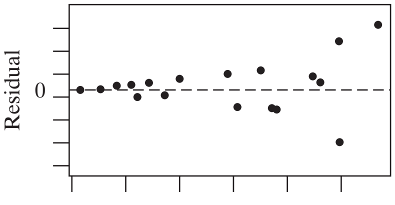
12.6.6 The following three residual plots, (i), (ii), and (iii), were generated after fitting regression lines to the following three scatterplots, (a), (b), and (c). Which residual plot goes with which scatterplot? How do you know?


(a)

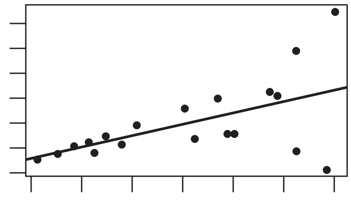


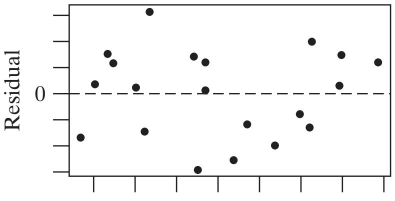
(i)


(b)



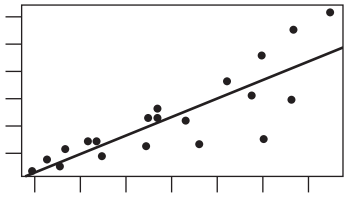
(ii)

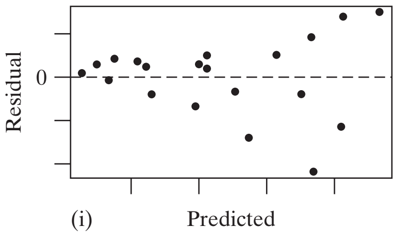

(c)



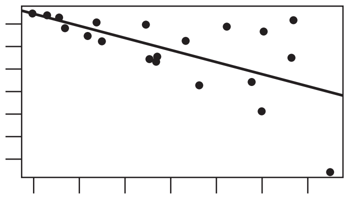
(iii)

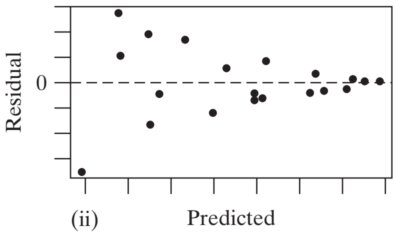
12.6.7 The following two residual plots, (i), and (ii), were generated after fitting regression lines to the two scatterplots (a) and (b). Which residual plot goes with which scatterplot? How do you know?


(a)



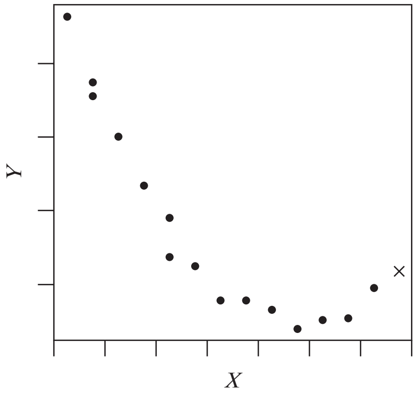
(i)


(b)



(ii)

12.6.8 Sketch the residual plot that would be produced by fitting a regression line to the following scatterplot. One of the points is plotted with an "×." Indicate this point on the residual plot.



12.6.9 (Computer exercise) Researchers measured the diameters of 20 trees in a central Amazon rain forest and used ^{14}C-dating to determine the ages of these trees. The data are given in the following table.[25] Consider the use of diameter, X, as a predictor of age, Y.

DIAMETER (cm)	AGE (yr)	DIAMETER (cm)	AGE (yr)
180	1372	115	512
120	1167	140	512
100	895	180	455
225	842	112	352
140	722	100	352
142	657	118	249
139	582	82	249
150	562	130	227
110	562	97	227
150	552	110	172

(a) Make a scatterplot of Y = age versus X = diameter and fit a regression line to the data.

(b) Make a residual plot from the regression in part (a). Then make a normal probability plot of the residuals. How do these plots call into question the use of a linear model and regression inference procedures?

(c) Take the logarithm of each value of age. Make a scatterplot of $Y = \log(\text{age})$ versus X = diameter and fit a regression line to the data.

(d) Make a residual plot from the regression in part (c). Next, make a normal probability plot of the residuals. Based on these plots, does a regression model using a log scale, from part (c), seem appropriate?

12.7 Precision in Prediction (Optional)

In Section 12.4 we learned that one very practical use of regression is prediction. In this section we shall distinguish between the prediction of the *mean* Y value for a particular X value and the prediction of a *single* Y value for a particular X value. In particular, we will compare the precisions of these two very different types of predictions.

Confidence and Prediction Intervals

In Example 12.4.6 we used a regression line to make a prediction: $\hat{y} = 197.17 - 2.51x$. Using this line again we could predict the *mean* arsenic concentration in rice from plants with straw silicon concentrations of 40 g/kg to be $\hat{y} = 197.17 - 2.51(40) = 96.77$ μg/kg. What if instead of estimating the mean arsenic concentration of all plants with this silicon concentration, we wanted to predict *the* arsenic concentration of *a* particular plant whose straw silicon concentration was 40 g/kg? Our estimate would still be the same, $\hat{y} = 96.77$ μg/kg. That is, whether we are estimating the mean Y value or a single Y value for a particular value of X, we use the regression line in the same manner. However, the precisions of these estimates are very different.

Predicting a single Y value is much less precise than predicting the mean Y value because in addition to the uncertainty in the regression line (e.g., uncertainty in our estimates of the slope and intercept of the line), there is also uncertainty due to the inherent variability in Y values that have the same value of X. For example, there is variability among the rice arsenic concentrations for all plants with straw silicon concentrations of 40 g/kg (in fact we estimate this variability to be s_e). The two graphs in Figure 12.7.1 illustrate the differences in our prediction precisions for the two types of estimates.

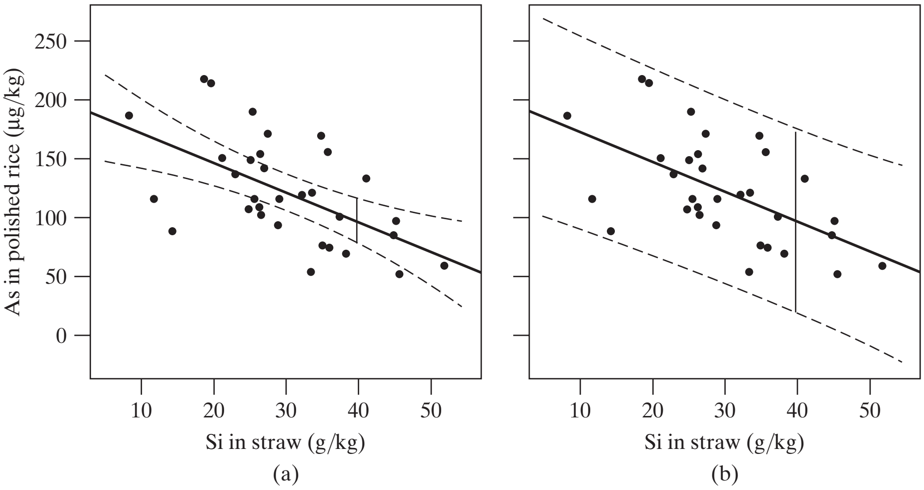


Figure 12.7.1 95% confidence and prediction bands for arsenic concentrations of rice. Plot (a) shows a 95% confidence band for the predicted mean arsenic concentrations and the 95% confidence interval for the predicted mean arsenic concentration when straw silicon is 40 g/kg. Plot (b) shows a 95% prediction band for predicted arsenic concentrations and the 95% prediction interval for the predicted arsenic concentration when straw silicon is 40 g/kg.

Figure 12.7.1 (a) displays a band representing all 95% confidence intervals for predicting mean arsenic levels as well as the specific interval for $X = 40$ g/kg marked by the vertical line. The confidence band reflects the uncertainty associated with estimating the slope and intercept of the regression line. Notice that the intervals are narrower (more precise) for straw silicon concentrations near the center of the data set and much wider near the extreme X values. We are 95% confident that the population regression line $\beta_0 + \beta_1 x$ lies within this band. The widening of the intervals on the end is a reflection of our uncertainty in our estimate of the slope of the regression line. The width of the band in the middle expresses our uncertainty of the overall height of the regression line (vis-à-vis b_0).

In contrast, Figure 12.7.1 (b) displays a band representing all 95% prediction intervals for predicting individual arsenic levels. The specific prediction interval for $X = 40$ is marked by the vertical line. Note how much wider this band is in (b) than in (a). Example 12.7.1 illustrates the use of confidence and prediction intervals for prediction in regression.

Example 12.7.1

Arsenic Concentrations in Rice Figure 12.7.1 shows that for rice with straw silicon concentrations of 40 g/kg, the 95% confidence interval for the mean arsenic concentration is about 75 to 125 μg/kg. In other words, we are 95% confident that the mean arsenic concentration of rice from plants with straw silicon concentrations of 40 g/kg is 75 to 125 μg/kg. On the other hand, using the prediction interval we estimate that 95% of plants with straw silicon of 40 g/kg will have rice arsenic concentrations roughly between 25 and 175 μg/kg. ■

Recall that the regression line can be interpreted as a "line of averages," and individuals will necessarily fall from this average. These graphs show us that we are much less certain about saying, "rice from plants with X amount of straw silicon will have Y amount of arsenic" than we are about saying "rice from plants with X amount of straw silicon will, *on average*, have Y amount of arsenic."

Computing the Intervals

Consider predicting $\mu_{Y|X=x^*}$ or $Y|X = x^*$; that is, predicting the mean or actual Y value when $X = x^*$. A 95% confidence interval for $\mu_{Y|X=x^*}$ is given by

$$\hat{y} \pm t_{0.025}s_e\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

and a 95% prediction interval for $Y|X = x^*$ is given by

$$\hat{y} \pm t_{0.025}s_e\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

with the critical value $t_{0.025}$ determined from Student's t distribution with $\mathrm{df} = n - 2$.

While these two formulas are very similar, note the extra "1" under the radical sign in the prediction interval formula. This "1" factors in the added variability associated with trying to make a prediction for an individual rather than for a population mean.

As we have seen in Figure 12.7.1, both confidence and prediction intervals are wider when we are making predictions far from the center of our data. Both formulas account for this additional uncertainty through the term $\frac{(x^* - \bar{x})^2}{(n-1)s_x^2}$. This term will be large when x^* is far from $\bar{x}$ and thus increase the width of the interval. Note that when $x^* = \bar{x}$ the confidence interval formula can reduce to a very familiar form: $\hat{y} \pm t_{0.025}\left(\frac{s_e}{\sqrt{n}}\right)$, which looks very similar to the formula for a confidence interval for a population mean from Chapter 6.

Most statistical software can compute and display confidence and prediction bands quite easily.

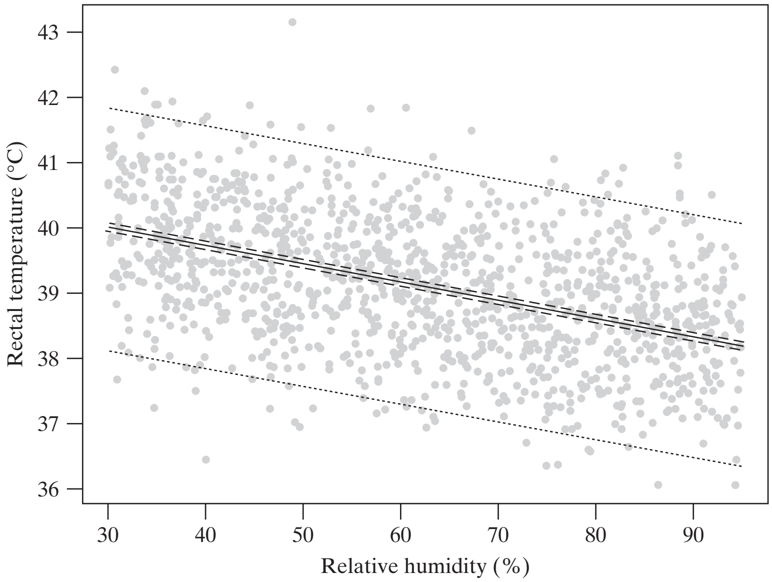
Exercises 12.7.1–12.7.3

12.7.1 In a study of heat stress on cows, researchers measured the rectal temperature (°C) (Y) and relative humidity (%) (X) for 1,280 lactating cows.[26] The following graph displays the data and regression line (solid line). There are two other pairs of lines on this graph: dashed and dotted. One pair of lines shows the 95% confidence band and the other shows the 95% prediction band.

(a) Which pair of lines shows the confidence band? What does this band tell us?

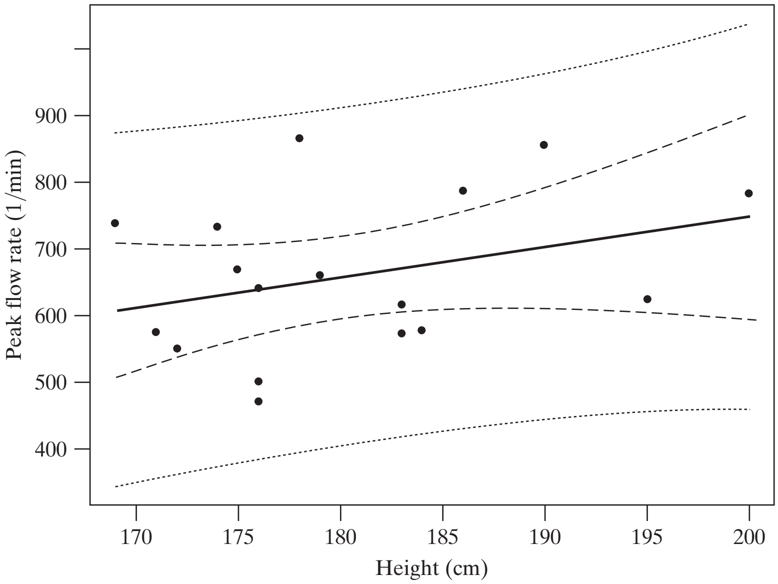
(b) Which pair of lines shows us the prediction band? What does this band tell us?

(c) If the data set were smaller, describe what would happen to these bands. Would we have narrower or wider bands around the regression line?



12.7.2 (*Continuation of 12.7.1*) Suppose 5,000 additional cows were included in the sample and a similar plot of the data, regression line, confidence and prediction bands were made of this new larger sample. Would the prediction band get much narrower? Explain your reasoning.

12.7.3 The following graph displays the regression line and 95% confidence and prediction bands for the peak respiration flow data from Exercise 12.3.8.



(a) Using the graph to justify your answer, would it be very surprising to find a 195-cm-tall individual with a peak flow rate above 900 l/min?

(b) Using the graph to justify your answer, would it be surprising to find a large group of 195-cm-tall individuals to have a mean peak flow rate above 900 l/min?

12.8 Perspective

To put the methods of Chapter 12 in perspective, we will discuss their relationship to methods described in earlier chapters, and to methods that might be included in a second statistics course. We begin by relating regression to the methods of Chapters 7 and 11.

Regression and the *t* Test

When there are several Y values for each of two values of X, one could analyze the data with a two-sample t test or with a regression analysis. Each approach uses the data to estimate the conditional mean of Y for each fixed X; these parameters are estimated by the fitted line $b_0 + b_1 x$ in the regression approach and by the individual sample means $\overline{Y}$ in the t test approach. To test the null hypothesis of no dependence of Y on X, each approach translates the null hypothesis into its own terms. The following example illustrates the approaches.

Example 12.8.1

Toluene and the Brain In Chapter 7 we analyzed data on norepinephrine (NE) concentrations in the brains of six rats exposed to toluene and of five control rats. The data are reproduced in Table 12.8.1.

Table 12.8.1 NE concentrations (ng/gm)

	Toluene	Control
	543	535
	523	385
	431	502
	635	412
	564	387
	549	
n	6	5
$\bar{y}$	540.83	444.20
s	66.12	69.64

In Chapter 7 the null hypothesis

$$H_0: \mu_1 - \mu_2 = 0$$

was tested using the (unpooled) two-sample t test. The test statistic was

$$t_s = \frac{(540.83 - 444.20) - 0}{41.195} = 2.346$$

These data could be analyzed using a pooled t test (or, equivalently, with analysis of variance). The pooled variance is

$$s^2_{\text{pooled}} = \frac{(6 - 1)66.12^2 + (5 - 1)69.64^2}{(6 + 5 - 2)} = 4584.24 = 67.71^2$$

and the pooled SE is

$$\text{SE}_{\text{pooled}} = 67.71\sqrt{\frac{1}{6} + \frac{1}{5}} = 41.00$$

This leads to a test statistic of

$$t_s = \frac{(540.83 - 444.20) - 0}{41.00} = 2.357$$

which is not much different than the unpooled t test result.

These data can also be analyzed with a regression model. To use regression, we define an **indicator variable**—a variable that indicates group membership—as follows. Let $X = 0$ for observations in the control group and let $X = 1$ for observations in the toluene group. Then we can present the data graphically with a scatterplot, as in Figure 12.8.1.

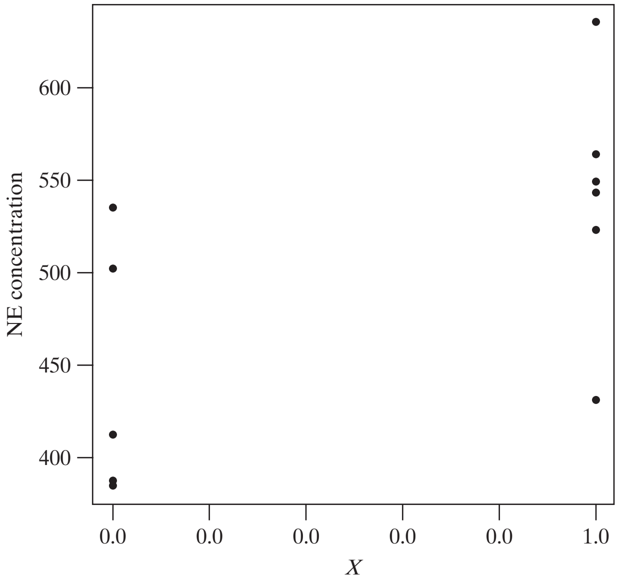


Figure 12.8.1 NE concentration data. $X = 0$ represents the control group; $X = 1$ represents the toluene group

We can analyze the data in the scatterplot with the linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

which states that $\mu_{Y|X} = \beta_0 + \beta_1 X$.

The linear model states that for rats in the control group, the (population) mean NE concentration is given by

$$\mu_{Y|X=0} = \beta_0 + \beta_1(0) = \beta_0$$

And, for rats in the toluene group, NE concentration is given by

$$\mu_{Y|X=1} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$

The difference between the two group means is β_1. Thus, the null hypothesis

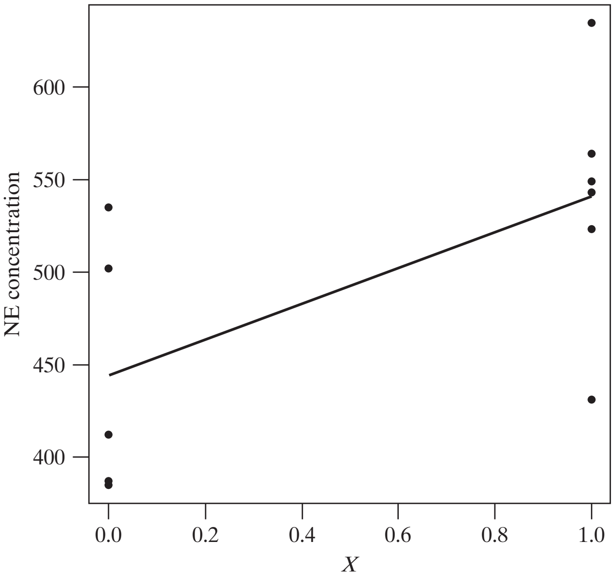
$$H_0\colon \mu_{Y|X=0} - \mu_{Y|X=1} = 0$$

is equivalent to the null hypothesis

$$H_0\colon \beta_1 = 0$$

The fitted regression line is $\hat{y} = 444.2 + 96.63x$. Note that when $X = 0$, the fitted regression line gives a value of $\hat{y} = 444.2$, which is the sample mean of the control group. When $X = 1$, the fitted regression line gives a value of $\hat{y} = 444.2 + 96.63 = 540.83$, which is the sample mean of the toluene group. That is, the sample value of the slope is equal to the change in the sample means when going from the control group ($X = 0$) to the toluene group ($X = 1$), as shown in Figure 12.8.2.

Figure 12.8.2 NE concentration data with regression line added



The test statistic for testing the hypothesis $H_0\colon \beta_1 = 0$ is

$$t_s = \frac{96.63}{41.0} = 2.36$$

This is identical to the previous pooled two-sample t test statistic. (Note that the regression analysis assumes that $\sigma_{Y|X} = \sigma_\varepsilon$ is constant. Thus, regression is similar to the pooled t test, rather than the unpooled t test.) The following computer output shows the coefficients for the fitted regression line as well as the t statistic.

```
The regression equation is
NE = 444 + 96.6X

Predictor     Coef   SE Coef       T      P

Constant    444.20     30.28   14.67  0.000
X            96.63     41.00    2.36  0.043

S = 67.7049  R-Sq = 38.2%  R-Sq(adj) = 31.3%

Analysis of Variance

Source           DF      SS      MS     F       P

Regression        1   25467   25467  5.56   0.043
Residual Error    9   41256    4584
Total            10   66723
```

■

The following example compares the regression approach and the two-sample approach to a data set for which (unlike Example 12.8.1) X varies within as well as between the samples.

Example 12.8.2

Blood Pressure and Platelet Calcium In Example 12.5.3 we described blood pressure (X) and platelet calcium (Y) measurements on 38 subjects. Actually, the study included two groups of subjects: 38 volunteers with normal blood pressure, selected from hospital lab personnel and other nonpatients, and 45 patients with a diagnosis of high blood pressure. Table 12.8.2 summarizes the platelet calcium measurements in the two groups and Figure 12.8.3 shows the blood pressure and calcium measurements for all 83 subjects.[4]

Two ways to analyze the data are (1) as two independent samples and (2) by regression analysis. To test for a relationship between blood pressure and platelet calcium (1) a two-sample t test of $H_0: \mu_1 = \mu_2$ can be applied to Table 12.8.2; (2) a regression t test of $H_0: \beta_1 = 0$ can be applied to the data in Figure 12.8.3. The two-sample t statistic (unpooled) is $t_s = 11.2$ and the regression t statistic is $t_s = 20.8$. Both of these are highly significant, but the latter is more so because the regression analysis extracts more information from the data.

For these data, the regression approach is more enlightening and convincing than the two-sample approach. Figure 12.8.3 suggests that platelet calcium is correlated with blood pressure, not only between, but also within the two groups. Relevant regression analyses would include (1) testing for a correlation within each group separately (as in Examples 12.2.3 and 12.5.3); (2) testing for an overall correlation (as in the previous paragraph); (3) testing whether the regression lines in the two groups are identical (using methods not described in this book).

Table 12.8.2 Platelet calcium (nM) in two groups of subjects

	Normal blood pressure	High blood pressure
$\bar{y}$	107.9	168.2
s	16.1	31.7
n	38	45

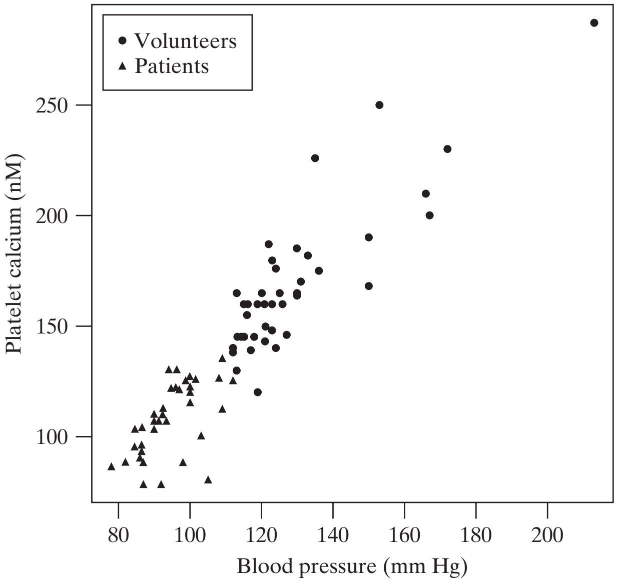


Figure 12.8.3 Blood pressure and platelet calcium for 83 subjects

Formal testing aside, notice the advantage of the scatterplot as a tool for understanding the data and for communicating the results. Figure 12.8.3 provides eloquent testimony to the reality of the relationship between blood pressure and platelet calcium. (We emphasize once again, however, that a "real" relationship is not necessarily a causal relationship. Further, even if the relationship is causal, the data do not indicate the direction of causality—that is, whether high calcium causes high blood pressure or vice versa.*) ■

Example 12.8.2 illustrates a general principle: If quantitative information on a variable X is available, it is usually better to use that information than to ignore it.

Extensions of Least Squares

We have seen that the classical method of fitting a straight line to data is based on the least-squares criterion. This versatile criterion can be applied to many other statistical problems. For instance, in **curvilinear regression**, the least-squares criterion is used to fit curvilinear relationships such as

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Another application is **multiple regression and correlation**, in which the least-squares criterion is used to fit an equation relating Y to several X variables—X_1, X_2, and so on; for instance,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

The following example illustrates both curvilinear and multiple regression.

*In fact, the authors of the study remark that "It remains possible . . . that an increased intracellular calcium concentration is a consequence rather than a cause of elevated blood pressure."

Example 12.8.3

Serum Cholesterol and Blood Pressure As part of a large health study, various measurements of blood pressure, blood chemistry, and physique were made on 2,599 men.[27] The researchers found a positive correlation between blood pressure and serum cholesterol ($r = 0.23$ for systolic blood pressure). But blood pressure and serum cholesterol are also related to age and physique. To untangle the relationships, the researchers used the method of least squares to fit the following equation:

$$Y = b_0 + b_1X_1 + c_1X_1^2 + b_2X_2 + b_3X_3 + b_4X_4$$

where

$$\begin{aligned} Y &= \text{Systolic blood pressure} \\ X_1 &= \text{Age} \\ X_2 &= \text{Serum cholesterol} \\ X_3 &= \text{Blood glucose} \\ X_4 &= \text{Ponderal index (height divided by the cube root of weight)} \end{aligned}$$

Note that the regression is curvilinear with respect to age (X_1) and linear in the other X variables.

By applying multiple regression and correlation analysis, the investigators determined that there is little or no correlation between blood pressure and serum cholesterol, after accounting for any relationship between blood pressure and age and ponderal index. They concluded that the observed correlation between serum cholesterol and blood pressure was an indirect consequence of the correlation of each of these with age and physique. ■

Nonparametric and Robust Regression and Correlation

We have discussed the classical least-squares methods for regression and correlation analysis. There are also many excellent modern methods that are not based on the least-squares criterion. Some of these methods are *robust*—that is, they work well even if the conditional distributions of Y given X have long straggly tails or outliers. The nonparametric methods assume little or nothing about the form of dependence—linear or curvilinear—of Y on X, or about the form of the conditional distributions.

Analysis of Covariance

Sometimes regression ideas can add greatly to the power of a data analysis, even if the relationship between X and Y is not of primary interest. The following is an example.

Example 12.8.4

Caterpillar Head Size Can diet affect the size of a caterpillar's head? Such an effect is plausible, because a caterpillar's chewing muscles occupy a large part of the head. To study the effect of diet, a biologist raised caterpillars (*Pseudaletia unipuncta*) on three different diets: diet 1, an artificial soft diet; diet 2, soft grasses; and diet 3, hard grasses. He measured the weight of the bead and of the entire body in the final stage of larval development. The results are shown in Figure 12.8.4, where $Y = \ln(\text{head weight})$ is plotted against $X = \ln(\text{body weight})$, with different symbols for the three diets.[28] Note that the effect of diet is striking; there is virtually no overlap between the three groups of points. But if we were to ignore X and consider Y only, as displayed in Figure 12.8.5, the effect of diet would be much less pronounced. ■

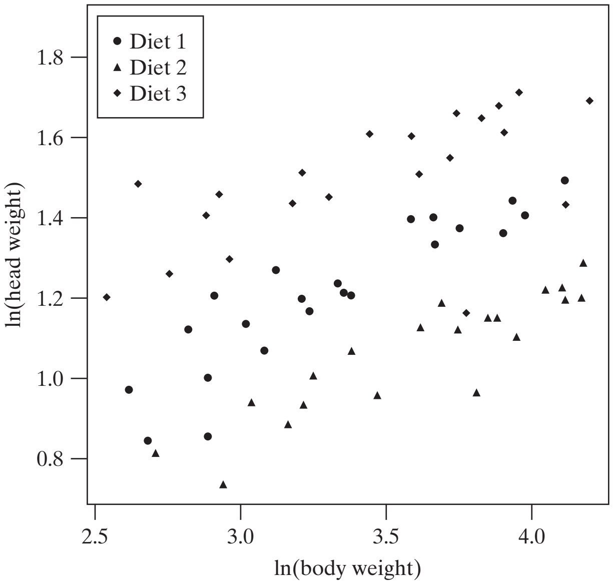


Figure 12.8.4 Head weight versus body weight (on logarithmic scales) for caterpillars on three different diets

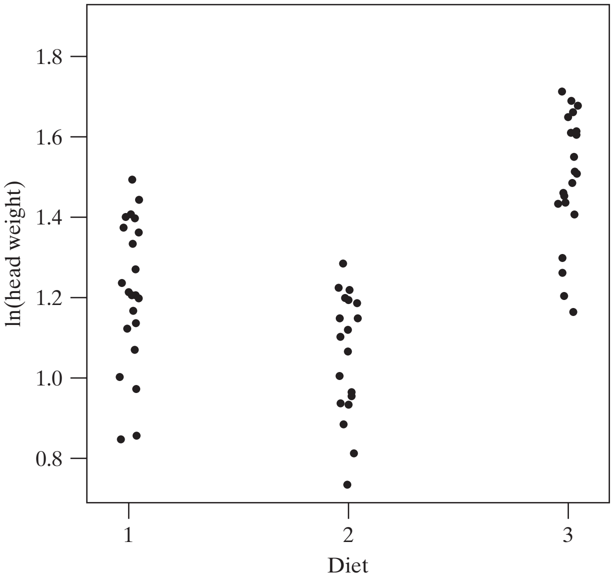


Figure 12.8.5 Head weight (on a logarithmic scale) for caterpillars on three different diets

Example 12.8.4 shows how comparison of several groups with respect to a variable Y can be strengthened by using information on an auxiliary variable X that is correlated with Y. A classical method of statistical analysis for such data is **analysis of covariance**, which proceeds by fitting regression lines to the (X, Y) data. But even without this formal technique, an investigator can often clarify the interpretation of data simply by constructing a scatterplot like Figure 12.8.4. Plotting the data against

X has the visual effect of removing that part of the variability in Y which is accounted for by X, causing the treatment effect to stand out more clearly against the residual background variation.

Logistic Regression

Regression and correlation are used to analyze the relationship between two quantitative variables, X and Y. Sometimes data arise in which a quantitative variable X is used to predict the response of a categorical variable Y. For example, we might wish to use X = cholesterol level as a predictor of whether or not a person has heart disease. Here we could define a variable Y as 1 if a person has heart disease and 0 otherwise. We could then study how Y depends on X. When the response variable is dichotomous, as in this case, a technique known as **logistic regression** can be used to model the relationship. For example, logistic regression could be used to model how the probability of heart disease depends on blood pressure.

Example 12.8.5 provides a more detailed look at the use of logistic regression.

Example 12.8.5

Esophageal Cancer Esophageal cancer is a serious and very aggressive disease. Scientists conducted a study of 31 patients with esophageal cancer in which they studied the relationship between the size of the tumor that a patient had and whether or not the cancer had spread (metastasized) to the lymph nodes of the patient. In this study the response variable is dichotomous: $Y = 1$ if the cancer had spread to the lymph nodes and $Y = 0$ if not. The predictor variable is the size (recorded as the maximum dimension, in cm) of the tumor found in the esophagus. The data are given in Table 12.8.3 and plotted in Figure 12.8.6.[29]

Table 12.8.3 Esophageal cancer data

Patient number	Tumor size (cm), X	Lymph node metastasis, Y	Patient number	Tumor size (cm), X	Lymph node metastasis, Y
1	6.5	1	17	6.2	1
2	6.3	0	18	2.0	0
3	3.8	1	19	9.0	1
4	7.5	1	20	4.0	0
5	4.5	1	21	3.0	1
6	3.5	1	22	6.0	1
7	4.0	0	23	4.0	0
8	3.7	0	24	4.0	0
9	6.3	1	25	4.0	0
10	4.2	1	26	5.0	1
11	8.0	0	27	9.0	1
12	5.2	1	28	4.5	1
13	5.0	1	29	3.0	0
14	2.5	0	30	3.0	1
15	7.0	1	31	1.7	0
16	5.3	0			

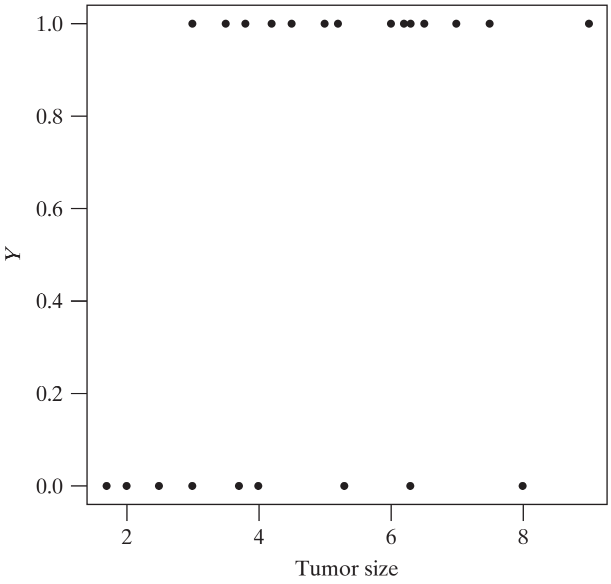


Figure 12.8.6 Lymph node metastasis, Y, as a function of tumor size, X

The idea of logistic regression is to model the relationship between X and Y by fitting a response curve that is always between 0 and 1. With values bound between 0 and 1, the logistic regression model can be used to estimate the probability $Y = 1$ (e.g., metastasis) for a given value of X (e.g., tumor size). Thus, unlike linear regression, in which we model Y as a linear function of X (which does not remain between 0 and 1), with logistic regression we model the relationship between X and Y as having an "S" shape, as shown in Figure 12.8.7.

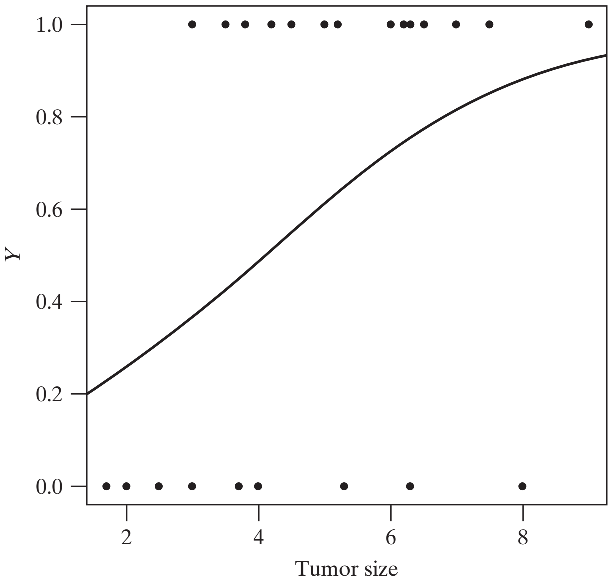


Figure 12.8.7 Lymph node metastasis, Y, as a function of tumor size, X, with smooth curve added

One way to begin understanding the data is to form groups on the basis of size, X, and calculate for each group the proportion of the Y values that are 1's. (This is somewhat analogous to finding the graph of averages described in Section 12.3, except that here we group together data points with differing X values.) Table 12.8.4 provides such a summary, which is shown graphically in Figure 12.8.8. Note that the

Table 12.8.4 Esophageal cancer data in groups

Size range	Points with $Y = 1$	Points with $Y = 0$	Fraction $Y = 1$	Proportion $Y = 1$
(1.5, 3.0]	2	4	2/6	0.33
(3.0, 4.5]	5	6	5/11	0.45
(4.5, 6.0]	4	1	4/5	0.80
(6.0, 7.5]	5	1	5/6	0.83
(7.5, 9.0]	2	1	2/3	0.67

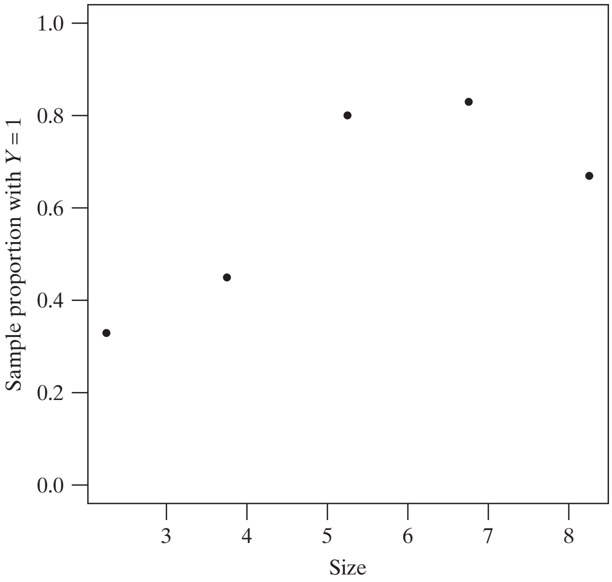


Figure 12.8.8 Sample proportion of patients with lymph node metastasis ($Y = 1$) for patients grouped by tumor size, X

proportion of 1's (that is, the proportion of patients for whom the cancer has metastasized) increases as tumor size increases (except for the last category of (7.5, 9], which has only three cases).

We can fit a smooth, continuous function to the data, to smooth out the percentages in the last column of Table 12.8.4. We can also impose the condition that the function be monotonically increasing, meaning that the probability of metastatis ($Y = 1$) strictly increases as tumor size increases. To do this, we use a computer to fit a **logistic response function**.* The fitted logistic response function for the esophageal cancer data is

$$\Pr\{Y = 1\} = \frac{e^{-2.086 + 0.5117 \times \text{size}}}{1 + e^{-2.086 + 0.5117 \times \text{size}}}$$

For example, suppose the size of a tumor is 4.0 cm. Then the predicted probability that the cancer has metastasized is

$$\frac{e^{-2.086 + 0.5117(4)}}{1 + e^{-2.086 + 0.5117(4)}} = \frac{e^{-0.0392}}{1 + e^{-0.0392}} = \frac{0.96156}{1 + 0.96156} = 0.49$$

*Fitting a logistic model is quite a bit more complicated than is fitting a linear regression model. A technique known as maximum likelihood estimation is commonly used, with the help of a computer.

On the other hand, suppose the size of a tumor is 8.0 cm. Then the predicted probability that the cancer has metastasized is

$$\frac{e^{-2.086 + 0.5117 \times 8}}{1 + e^{-2.086 + 0.5117 \times 8}} = \frac{e^{2.0076}}{1 + e^{2.0076}} = \frac{7.4454}{1 + 7.4454} = 0.88$$

We can calculate a predicted probability that $Y = 1$ for each value of X. Figure 12.8.9 shows a graph of such predictions, which have, generally speaking, an S shape. ■

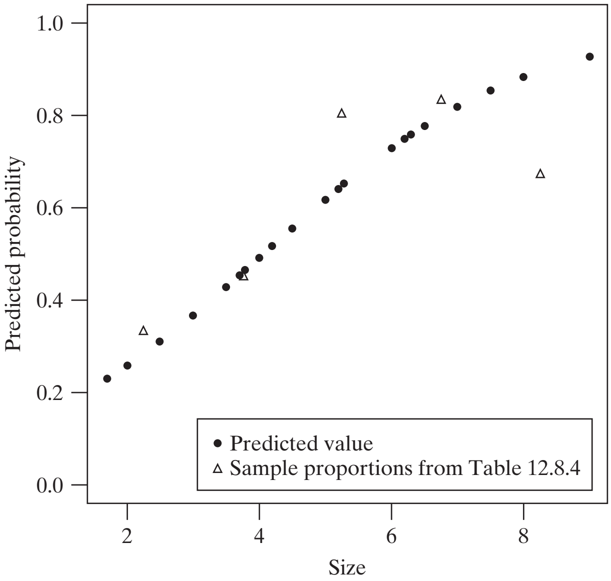


Figure 12.8.9 Predicted probability that $Y = 1$ as a function of tumor size, X with sample proportions from Table 12.8.4

The S shape of the logistic curve is easier to see if we extend the range of X, as shown in Figure 12.8.10. As X grows, the logistic curve approaches, but never exceeds, 1. Likewise, if we were to extend the curve into the region where X is less than zero we would see that as X gets smaller and smaller, the logistic curve approaches,

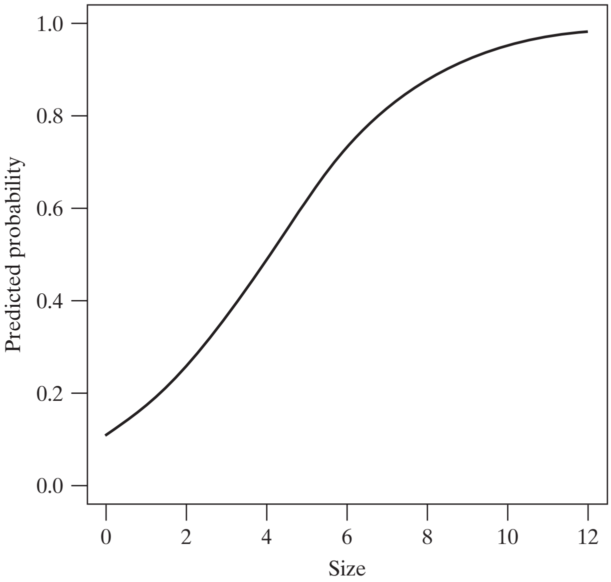


Figure 12.8.10 Logistic response function for the cancer data, shown over a larger range

but never drops below, 0. (Of course, in the setting of Example 12.8.5 it does not make sense to talk about tumor sizes that are negative. Thus, we only show the logistic curve for positive values of X.)

In general, if we have a logistic response function

$$\Pr\{Y = 1\} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

with b_1 positive, then as X grows, $\Pr\{Y = 1\}$ approaches one and as X gets smaller, $\Pr\{Y = 1\}$ approaches zero. Thus, unlike a linear regression model, a logistic curve stays between zero and one, which makes it appropriate for modeling a response probability.

12.9 Summary of Formulas

For convenient reference, we summarize the formulas presented in Chapter 12.

Correlation Coefficient

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

Fact 12.3.1:

$$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Fitted Regression Line

$$\hat{y} = b_0 + b_1 x$$

where

$$b_1 = r \times \left(\frac{s_y}{s_x}\right)$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Residuals:

$$y_i - \hat{y}_i \quad \text{where} \quad \hat{y}_i = b_0 + b_1 x_i$$

Residual Sum of Squares:

$$\text{SS(resid)} = \sum (y_i - \hat{y}_i)^2$$

Residual Standard Deviation:

$$s_e = \sqrt{\frac{\text{SS(resid)}}{n-2}}$$

Inference

Standard Error of b_1:

$$\mathrm{SE}_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$$

95% confidence interval for β_1:

$$b_1 \pm t_{0.025}\mathrm{SE}_{b_1}$$

Test of $H_0{:}\,\beta_1 = 0$ or $H_0{:}\,\rho = 0$:

$$t_s = \frac{b_1}{\mathrm{SE}_{b_1}} = r\sqrt{\frac{n-1}{1-r^2}}$$

Critical values for the test and confidence interval are determined from Student's t distribution with df $= n - 2$.

Prediction

A 95% confidence interval for $\mu_{Y|X=x^*}$ is given by

$$\hat{y} \pm t_{0.025}s_e\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

A 95% prediction interval for $Y|X = x^*$ is given by

$$\hat{y} \pm t_{0.025}s_e\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

Critical values for intervals are determined from Student's t distribution with df $= n - 2$.

Exercises 12.S.1–12.S.22

12.S.1 In a study of the Mormon cricket (*Anabrus simplex*), the correlation between female body weight and ovary weight was found to be $r = 0.836$. The standard deviation of the ovary weights of the crickets was 0.429 g. Assuming that the linear model is applicable, estimate the standard deviation of ovary weights of crickets whose body weight is 4 g.[30]

12.S.2 In a study of crop losses due to air pollution, plots of Blue Lake snap beans were grown in open-top field chambers, which were fumigated with various concentrations of sulfur dioxide. After a month of fumigation, the plants were harvested and the total yield of bean pods was recorded for each chamber. The results are shown in the table.[31]

	X = SULFUR DIOXIDE CONCENTRATION (ppm)			
	0	0.06	0.12	0.30
Y = yield (kg)	1.15	1.19	1.21	0.65
	1.30	1.64	1.00	0.76
	1.57	1.13	1.11	0.69
Mean	1.34	1.32	1.11	0.70

Preliminary calculations yield the following results.

$$\bar{x} = 0.12 \qquad \bar{y} = 1.117$$
$$s_X = 0.11724 \qquad s_Y = 0.31175$$
$$r = -0.8506 \qquad \mathrm{SS(resid)} = 0.2955$$

(a) Calculate the linear regression of Y on X.

(b) Plot the data and draw the regression line on your graph.

(c) Calculate s_e. What are the units of s_e?

12.S.3 Refer to Exercise 12.S.2.

(a) Assuming that the linear model is applicable, find estimates of the mean and the standard deviation of yields of beans exposed to 0.24 ppm of sulfur dioxide.

(b) Which condition of the linear model appears doubtful for the snap bean data?

12.S.4 Refer to Exercise 12.S.2. Consider the null hypothesis that sulfur dioxide concentration has no effect on yield. Assuming that the linear model holds, formulate this as a hypothesis about the true regression line. Use the data to test the hypothesis against a directional alternative. Let $\alpha = 0.05$.

12.S.5 Another way to analyze the data of Exercise 12.S.2 is to take each treatment mean as the observation Y; then the data would be summarized as in the accompanying table.

	SULFUR DIOXIDE X (ppm)	MEAN YIELD Y (kg)
	0.00	1.34
	0.06	1.32
	0.12	1.11
	0.30	0.70
Mean	0.1200	1.1175
SD	0.12961	0.29714

$r = -0.98666$
SS(resid) $= 0.007018$

(a) For the regression of mean yield on X, calculate the regression line and the residual standard deviation, and compare with the results of Exercise 12.S.2. Explain why the discrepancy is not surprising.

(b) What proportion of the variability in mean yield is explained by the linear relationship between mean yield and sulfur dioxide? Using the data in Exercise 12.S.5, what proportion of the variability in individual chamber yield is explained by the linear relationship between individual chamber yield and sulfur dioxide? Explain why the discrepancy is not surprising.

12.S.6 In a study of the tufted titmouse (*Parus bicolor*), an ecologist captured seven male birds, measured their wing lengths and other characteristics, and then marked and released them. During the ensuing winter, he repeatedly observed the marked birds as they foraged for insects and seeds on tree branches. He noted the branch diameter on each occasion, and calculated (from 50 observations) the average branch diameter for each bird. The results are shown in the table.[32]

BIRD	WING LENGTH X (mm)	BRANCH DIAMETER Y (cm)
1	79.0	1.02
2	80.0	1.04
3	81.5	1.20
4	84.0	1.51
5	79.5	1.21
6	82.5	1.56
7	83.5	1.29
Mean	81.429	1.2614
SD	1.98806	0.21035

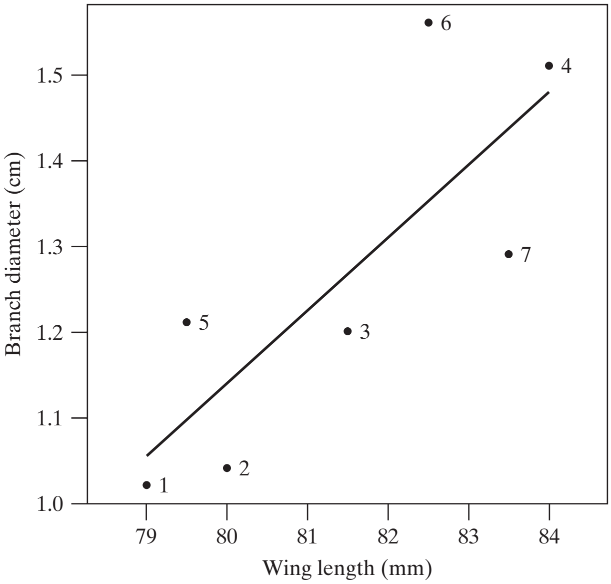
$r = 0.80335$
SS(resid) $= 0.09415$

(a) Calculate s_e and specify the units. Verify the approximate relationship between s_Y and s_e, and r.

(b) Do the data provide sufficient evidence to conclude that the diameter of the forage branches chosen by male titmice is correlated with their wing length? Test an appropriate hypothesis against a nondirectional alternative. Let $\alpha = 0.05$.

(c) The test in part (a) was based on 7 observations, but each branch diameter value was the mean of 50 observations. If we were to test the hypothesis of part (a) using the raw numbers, we would have 350 observations rather than only 7. Why would this approach not be valid?

12.S.7 (*Continuation of 12.S.6*) A scatterplot and fitted regression line of the data from Exercise 12.S.6 follow. The individual birds are labeled in the plot.



(a) Which bird/point has the largest regression residual?

(b) Which bird(s)/points(s) have the most leverage?

(c) Are there any birds/points that are influential?

(d) Invent your own bird observation of x = wing length and y = branch diameter that would be an example of a regression outlier.

(e) Invent your own bird observation of x = wing length and y = branch diameter that would be an example of a leverage point.

12.S.8 Exericise 12.3.7 deals with data on the relationship between body length and jumping distance of bullfrogs. A third variable that was measured in that study was the mass of each bullfrog. The following table shows these data.[16]

BULLFROG	LENGTH X (mm)	MASS Y (g)
1	155	404
2	127	240
3	136	296
4	135	303
5	158	422
6	145	308
7	136	252
8	172	533.8
9	158	470
10	162	522.9
11	162	356
Mean	149.636	373.427
SD	14.4725	104.2922

Preliminary calculations yield the following results:

$$r = 0.90521 \quad \text{SS(resid)} = 19642$$

(a) Calculate the linear regression of Y on X.

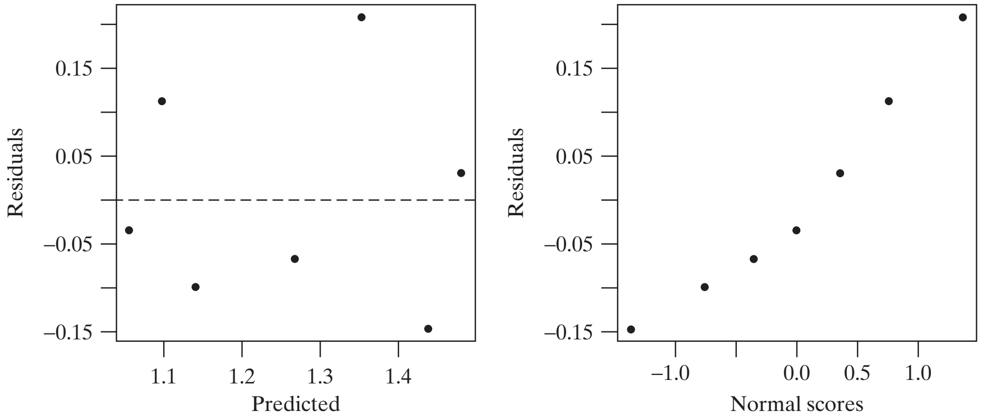
(b) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

(c) Calculate and interpret the value of s_e in the context of this setting.

(d) Calculate and interpret the value of r^2 in the context of this problem.

12.S.9 (*Continuation of 12.S.8*). A residual plot and normal probability plot from the linear regression of Y on X based on the bullfrog mass data in Exercise 12.S.8 follow.

Use these plots to comment on the required conditions for inference in regression. Is there any reason to substantially doubt that these conditions are met?

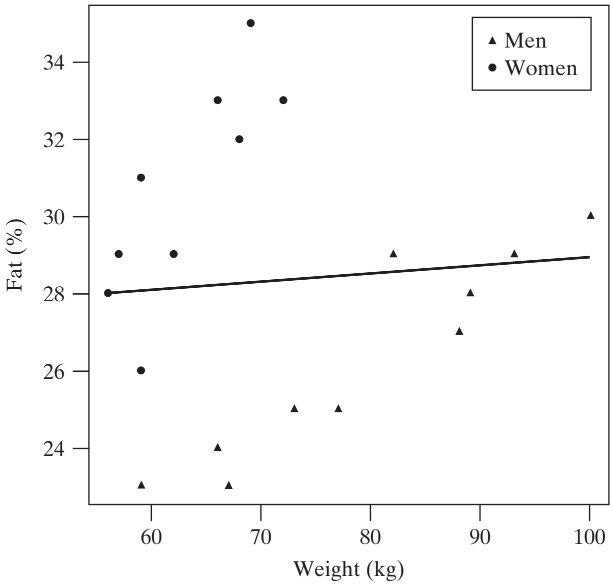


12.S.10 An exercise physiologist used skinfold measurements to estimate the total body fat, expressed as a percentage of body weight, for 19 participants in a physical fitness program. The body fat percentages and the body weights are shown in the table.[33]

	WEIGHT	FAT		WEIGHT	FAT
PARTICIPANT	X (kg)	Y (%)	PARTICIPANT	X (kg)	Y (%)
1	89	28	11	57	29
2	88	27	12	68	32
3	66	24	13	69	35
4	59	23	14	59	31
5	93	29	15	62	29
6	73	25	16	59	26
7	82	29	17	56	28
8	77	25	18	66	33
9	100	30	19	72	33
10	67	23			

Actually, participants 1 to 10 are men, and participants 11 to 19 are women. A summary and graph of the data for men, women, and both sexes combined into a single sample follow.

MEN (n = 10)	WOMEN (n = 9)	BOTH SEXES (n = 19)
$\bar{x} = 79.40$ $\bar{y} = 26.30$	$\bar{x} = 63.1$ $\bar{y} = 30.67$	$\bar{x} = 71.68$ $\bar{y} = 28.37$
$s_X = 13.2430$ $s_Y = 2.6269$	$s_X = 5.7975$ $s_Y = 2.8723$	$s_X = 13.1320$ $s_Y = 3.4835$
$r = 0.9352$	$r = 0.8132$	$r = 0.0780$



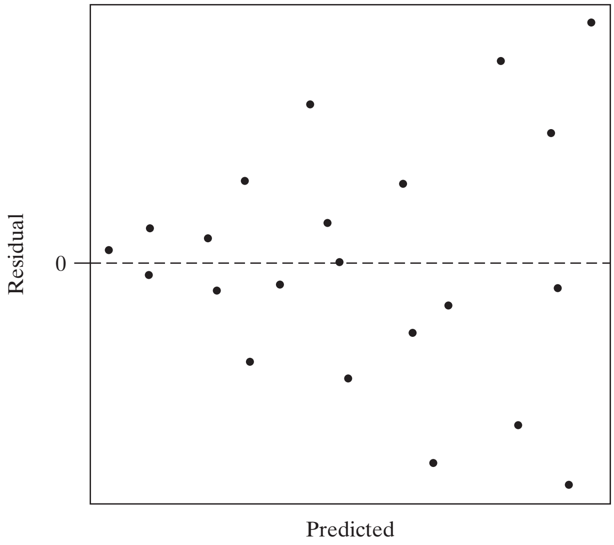
(a) Compute the regression equations for the males and females separately.

(b) The equation to the fitted regression line for both sexes combined, which is shown on the plot, is $\hat{y} = 26.88 + 0.021x$. How does the slope of this line compare to the slopes you computed in part (a)? Can you explain the discrepancy?

(c) Examine the correlation coefficients for (i) the males, (ii) the females, and (iii) both sexes combined. Do these values agree with your reasoning provided in part (b)?

12.S.11 Refer to the respiration rate data of Exercise 12.3.6. Construct a 95% confidence interval for β_1.

12.S.12 The following plot is a residual plot from fitting a regression model to some data. Make a sketch of the scatterplot of the data that led to this residual plot. (*Note*: There are two possible scatterplots—one in which b_1 is positive and one in which b_1 is negative.)



12.S.13 Biologists studied the relationship between embryonic heart rate and egg mass for 20 species of birds. They found that heart rate, Y, has a linear relationship with the logarithm of egg mass, X. The data are given in the following table.[34]

SPECIES	EGG MASS (g)	LOG-(EGG MASS) X	HEART RATE Y (beats/min)
Zebra finch	0.96	−0.018	335
Bengalese finch	1.10	0.041	404
Marsh tit	1.39	0.143	363
Bank swallow	1.42	0.152	298
Great tit	1.59	0.201	348
Varied tit	1.69	0.228	356
Tree sparrow	2.09	0.320	335
Budgerigar	2.19	0.340	314
House martin	2.25	0.352	357
Japenese bunting	2.56	0.408	370
Red-cheeked starling	4.14	0.617	358
Cockatiel	5.08	0.706	300
Brown-eared bulbul	6.40	0.806	333
Domestic pigeon	17.10	1.233	247
Fantail pigeon	19.70	1.294	267
Homing pigeon	19.80	1.297	230
Barn owl	20.10	1.303	219
Crow	20.50	1.312	297
Cattle egret	27.50	1.439	251
Lanner falcon	41.20	1.615	242
Mean	9.94	0.690	311

For these data the fitted regression equation is

$$\hat{y} = 368.06 - 82.452x$$

and

$$\text{SS(resid)} = 15748.6$$

(a) Interpret the value of the intercept of the regression line, b_0, in the context of this setting.

(b) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

(c) Calculate s_e and specify the units.

(d) Interpret the value of s_e in the context of this setting.

12.S.14 (*Computer exercise*) The accompanying table gives two data sets: (A) and (B). The values of X are the same for both data sets and are given only once.

	(A)	(B)
X	Y	Y
0.61	0.88	0.96
0.93	1.02	0.97
1.02	1.12	0.07
1.27	1.10	2.54
1.47	1.44	1.41
1.71	1.45	0.84
1.91	1.41	0.32
2.00	1.59	1.46
2.27	1.58	2.29
2.33	1.66	2.51
2.56	1.97	1.20
2.74	2.02	3.59
3.04	2.26	3.09
3.13	2.27	1.55
3.45	2.43	0.71
3.48	2.57	3.05
3.79	2.53	2.54
3.96	2.73	3.33
4.12	2.92	2.38
4.21	2.96	3.08

(a) Generate scatterplots of the two data sets.

(b) For each data set (i) estimate r visually and (ii) calculate r.

(c) For data set (a), multiply the values of X by 10, and multiply the values of Y by 3 and add 5. Recalculate r and compare with the value before the transformation. How is r affected by the linear transformation?

(d) Find the equations of the regression lines and verify that the regression lines for the two data sets are virtually identical (even though the correlation coefficients are very different).

(e) Draw the regression line on each scatterplot.

(f) Construct a scatterplot in which the two data sets are superimposed, using different plotting symbols for each data set.

12.S.15 (*Computer exercise*) This exercise shows the power of scatterplots to reveal features of the data that may not be apparent from the ordinary linear regression calculations. The accompanying table gives three fictitious data sets, A, B, and C. The values of X are the same for each data set, but the values of Y are different.[35]

DATA SET:	A	B	C
X	Y	Y	Y
10	8.04	9.14	7.46
8	6.95	8.14	6.77
13	7.58	8.74	12.74
9	8.81	8.77	7.11
11	8.33	9.26	7.81
14	9.96	8.10	8.84
6	7.24	6.13	6.08
4	4.26	3.10	5.39
12	10.84	9.13	8.15
7	4.82	7.26	6.42
5	5.68	4.74	5.73

(a) Verify that the fitted regression line is almost exactly the same for all three data sets. Are the residual standard deviations the same? Are the values of r the same?

(b) Construct a scatterplot for each of the data sets. What does each plot tell you about the appropriateness of linear regression for the data set?

(c) Plot the fitted regression line on each of the scatterplots.

12.S.16 (*Computer exercise*) In a pharmacological study, 12 rats were randomly allocated to receive an injection of amphetamine at one of two dosage levels or an injection of saline. Shown in the table is the water consumption of each animal (ml water per kg body weight) during the 24 hours following injection.[36]

DOSE OF AMPHETAMINE (ml/kg)		
0	1.25	2.5
122.9	118.4	134.5
162.1	124.4	65.1
184.1	169.4	99.6
154.9	105.3	89.0

(a) Calculate the regression line of water consumption on dose of amphetamine, and calculate the residual standard deviation.

(b) Construct a scatterplot of water consumption against dose.

(c) Draw the regression line on the scatterplot.

(d) Use linear regression to test the hypothesis that amphetamine has no effect on water consumption against the alternative that amphetamine tends to reduce water consumption. (Use $\alpha = 0.05$.)

(e) Use analysis of variance to test the hypothesis that amphetamine has no effect on water consumption. (Use $\alpha = 0.05$.) Compare with the result of part (d).

(f) What conditions are necessary for the validity of the test in part (d) but not for the test in part (e)?

(g) Calculate the pooled standard deviation from the ANOVA, and compare it with the residual standard deviation calculated in part (a).

12.S.17 (*Computer exercise*) Consider the Amazon tree data from Exercise 12.6.9. The researchers in this study were interested in how age, Y, is related to X = "growth rate," where growth rate is defined as diameter/age (i.e., cm of growth per year).

(a) Create the variable "growth rate" by dividing each diameter by the corresponding tree age.

(b) Make a scatterplot of Y = age versus X = growth rate and fit a regression line to the data.

(c) Make a residual plot from the regression in part (b). Then make a normal probability plot of the residuals. How do these plots call into question the use of a linear model and regression inference procedures?

(d) Take the logarithm of each value of age and of each value of growth rate. Make a scatterplot of $Y = \log(\text{age})$ versus $X = \log(\text{growth rate})$ and fit a regression line to the data.

(e) Make a residual plot from the regression in part (d). Then make a normal probability plot of the residuals. Based on these plots, does a regression model in log scale, from part (d), seem appropriate?

12.S.18 (*Computer exercise*) Researchers measured the blood pressures of 22 students in two situations: when the students were relaxed and when the students were taking an important examination. The table lists the systolic and diastolic pressures for each student in each situation.[37]

DURING EXAM		RELAXED	
SYSTOLIC PRESSURE (mm Hg)	DIASTOLIC PRESSURE (mm Hg)	SYSTOLIC PRESSURE (mm Hg)	DIASTOLIC PRESSURE (mm Hg)
132	75	110	70
124	170	90	75
110	65	90	65
110	65	110	80
125	65	100	55
105	70	90	60
120	70	120	80
125	80	110	60
135	80	110	70
105	80	110	70
110	70	85	65
110	70	100	60
110	70	120	80
130	75	105	75
130	70	110	70
130	70	120	80
120	75	95	60
130	70	110	65
120	70	100	65
120	80	95	65
120	70	90	60
130	80	120	70

(a) Compute the change in systolic pressure by subtracting systolic pressure when relaxed from systolic pressure during the exam; call this variable X.

(b) Repeat part (a) for diastolic pressure. Call the resulting variable Y.

(c) Make a scatterplot of Y versus X and fit a regression line to the data.

(d) Make a residual plot from the regression in part (c).

(e) Note the outlier in the residual plot [and on the scatterplot from part (c)]. Delete the outlier from the data set. Then repeat parts (c) and (d).

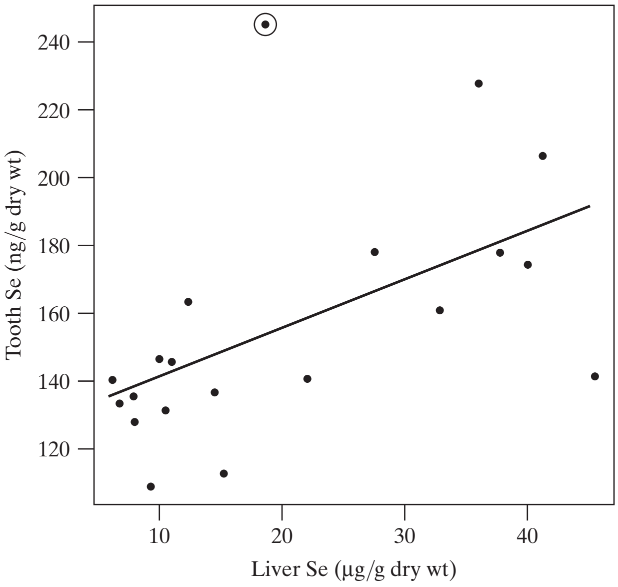
(f) What is the fitted regression model (after the outlier has been removed)?

12.S.19 (*Continuation of 12.S.18*) Consider the data from Exercise 12.S.18, part (f).

(a) Construct a 95% confidence interval for β_1.

(b) Interpret the confidence interval from part (a) in the context of this setting.

12.S.20 Selenium (Se) is an essential element which has been shown to play an important role in protecting marine mammals against the toxic effects of mercury (Hg) and other metals. It has been suggested that metal concentrations in marine mammal teeth can potentially be used as bioindicators for body burden. Twenty Belugas (*Delphinapterus leucas*) were harvested from the Mackenzie Delta, Northwest Territories, in 1996 and 2002, as part of an annual traditional Inuit hunt. Tooth and liver Se concentrations are reported in the table, summarized, and graphed.[38]



(a) Can we regard the sample correlation between Tooth (Y) and Liver (X) selenium, $r = 0.53726$, as an estimate of the population correlation coefficient? Briefly explain.

(b) If the circled point were removed from the data set, would the sample correlation listed in part (a) increase, decrease, or stay about the same?

WHALE	LIVER SE (μg/g)	TOOTH SE (ng/g)	WHALE	LIVER SE (μg/g)	TOOTH SE (ng/g)
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

(c) If the roles of X and Y were reversed (i.e., Y = Liver and X = Tooth Selenium), would the sample correlation listed in part (a) increase, decrease, or stay about the same?

(d) Is the circled point on the plot a leverage and/or influential point? Explain briefly.

(e) Is the circled point on the plot an outlier?

12.S.21 (*Continuation of 12.S.20*) The following are summary statistics for the Selenium data in Exercise 12.S.20.

$$\bar{x} = 20.684 \qquad \bar{y} = 156.599$$
$$s_X = 13.4489 \qquad s_Y = 36.0586$$
$$r = 0.53726 \qquad \text{SS(resid)} = 17{,}573.3$$

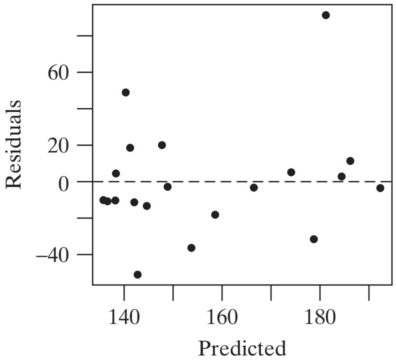
(a) Calculate the regression line of Tooth Selenium on Liver Selenium.

(b) Compute a 95% confidence interval for the slope of the regression line.

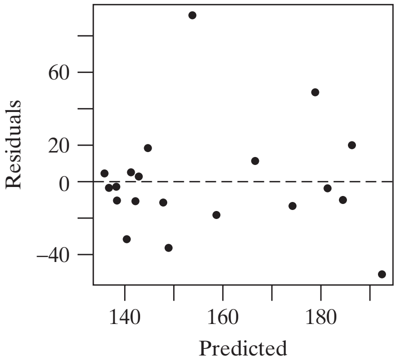
(c) Interpret the interval computed in part (b) in the context of the problem.

(d) Using the interval computed in part (b), is it reasonable to believe that the slope is as small as 0.25 (ng/g)/(μg/g)?

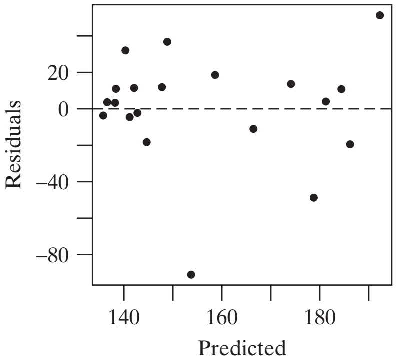
12.S.22 (*Continuation of 12.S.20 and 12.S.21*) Referring to the data plotted in Exercise 12.S.20, which of the following is a residual plot resulting from fitting the regression line in Exercise 12.S.21, part (a)? Justify your choice.



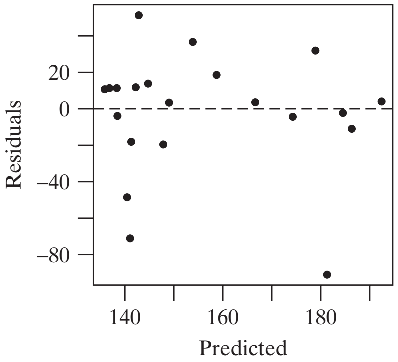
(a)



(b)



(c)



(d)

12.S.23 (*Continuation of 12.S.20*) The whales observed in this study were harvested during a traditional Inuit hunt in two particular years. What are we assuming about the captured whales to justify our analyses of these data in the preceding problems?