9

REGRESSION

In the Black Spruce Seedings Case Study in Section 1.9, the biologist was interested in how much the seedlings grew over the course of the study. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_{72}, y_{72})$ denote the height and diameter change, respectively, for each of the 72 seedlings. In Figure 9.1, we see that there is a strong, positive, and linear relationship between height and diameter changes.

In this chapter, we will describe a method to model this relationship, that is, we will find a mathematical equation that best explains the linear relationship between the change in height and the change in diameter.

9.1 COVARIANCE

In Chapter 2, we introduced the scatter plot as a graphical tool to explore the relationship between two numeric variables. For example, referring to Figure 9.2a, we might describe the relationship here between the two variables as positive, linear, and moderate to moderately strong.

Now, consider the graph in Figure 9.2b. How would you describe the relationship here? This relationship would be described as linear, positive, and strong.

In fact, these two graphs are of the same two variables! The difference in the two impressions are due to the y-axis scaling. In the first graph, the range of the y-axis is roughly -2.5 to 2.5; in the second graph, the y-axis range is roughly -4.5 to 4.5. Graphs are excellent tools for exploring data, but issues such as scaling can distort

Mathematical Statistics with Resampling and R, First Edition. By Laura Chihara and Tim Hesterberg.
© 2011 John Wiley & Sons, Inc. Published 2011 by John Wiley & Sons, Inc.

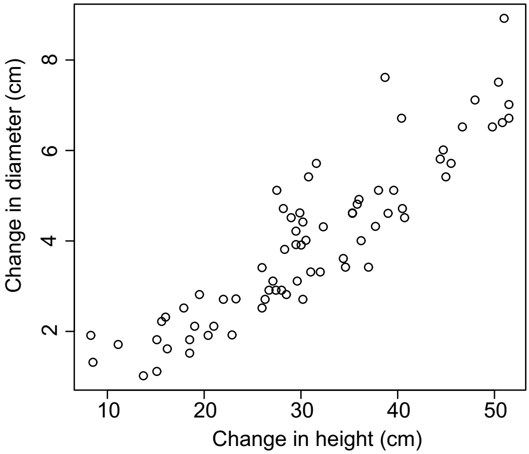

FIGURE 9.1 Change in diameter against change in height of seedlings over a 5-year period.

our perception of underlying properties and relationships. Thus, we will consider a numeric measure that indicates the strength of a linear relationship between the two variables.

We return to the Black Spruce data and recreate the scatter plot of diameter change against height change, adding a vertical line to mark the mean of the height changes ($\bar{x}$) and a horizontal line to mark the mean of the diameter changes ($\bar{y}$) (Figure 9.3).

For points in quadrant I, both $x_i - \bar{x}$ and $y_i - \bar{y}$ are positive, so $(x_i - \bar{x})(y_i - \bar{y})$ is positive. Similarly, in quadrant III, $(x_i - \bar{x})(y_i - \bar{y})$ is positive since each of the factors is negative. In quadrants II and IV, $(x_i - \bar{x})(y_i - \bar{y})$ is negative since the factors have opposite signs. For the Spruce data, on average, $(x_i - \bar{x})(y_i - \bar{y})$ is positive since most of the points are in quadrants I and III.

This motivates the following definition, a measure of how X and Y are related.

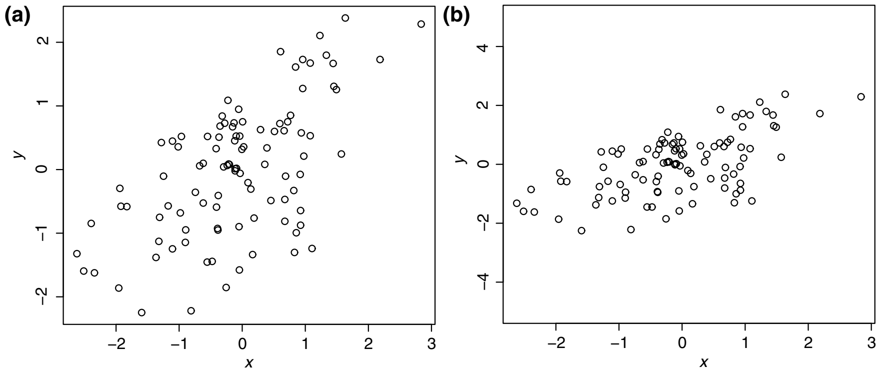

FIGURE 9.2 Two scatter plots of two numeric variables.

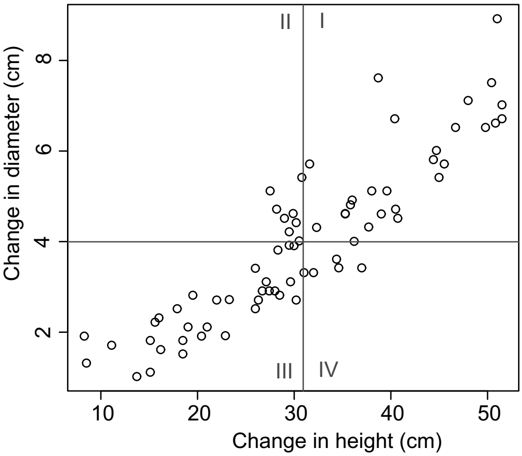

FIGURE 9.3 Change in diameter against change in height of seedlings over a 5-year period with vertical and horizontal lines at the mean values.

Definition 9.1 The *covariance of X and Y* is

$$\mathrm{Cov}[X, Y] = \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right]. \qquad ||$$

If $\mathrm{Cov}[X, Y] > 0$, then we say that X and Y are *positively correlated*: on average, X and Y are both either greater or less than their respective means. If $\mathrm{Cov}[X, Y] < 0$, then we say that X and Y are *negatively correlated*: on average, one of X or Y is greater than and one less than their mean.

Note that the definition implies that

$$\mathrm{Cov}[X, X] = \mathrm{E}\left[(X - \mu_X)^2\right] = \mathrm{Var}[X].$$

Proposition 9.1 $\mathrm{Cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$.

Proof

$$\begin{aligned}
\mathrm{Cov}[X, Y] &= \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \\
&= \mathrm{E}\left[(XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y)\right] \\
&= \mathrm{E}[XY] - \mu_Y \mathrm{E}[X] - \mu_X \mathrm{E}[Y] + \mu_X\mu_Y \\
&= \mathrm{E}[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]. \qquad \square
\end{aligned}$$

Example 9.1 Let X and Y have the joint distribution:

$$f(x, y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find the covariance of X and Y.

Solution The marginal distributions of X and Y are respectively as follows:

$$f_X(x) = \int_0^1 \frac{3}{2}(x^2 + y^2)\, dy = \frac{3}{2}\left(x^2 + \frac{1}{3}\right),$$

$$f_Y(y) = \int_0^1 \frac{3}{2}(x^2 + y^2)\, dx = \frac{3}{2}\left(y^2 + \frac{1}{3}\right).$$

Thus,

$$\mathrm{E}\,[X] = \int_0^1 x \cdot \frac{3}{2}\left(x^2 + \frac{1}{3}\right)\, dx = \frac{5}{8}$$

and, similarly, $\mathrm{E}\,[Y] = 5/8$. In addition,

$$\mathrm{E}\,[XY] = \int_0^1 \int_0^1 xy \cdot \frac{3}{2}(x^2 + y^2)\, dxdy = \frac{3}{8}.$$

Thus,

$$\mathrm{Cov}[X, Y] = \frac{3}{8} - \frac{5}{8} \cdot \frac{5}{8} = -\frac{1}{64}.$$

□

Corollary 9.1 *If X and Y are independent, then* $\mathrm{Cov}[X, Y] = 0$.

Remark The converse is false: $\mathrm{Cov}[X, Y] = 0$ does not imply that X and Y are independent. For example, if X has a symmetric distribution and $Y = (X - \mu_X)^2$, then Y completely depends on X, but the covariance is zero. ||

Covariances of sums of random variables add up; this yields a useful expression for the variance of a sum of random variables.

Theorem 9.1

$$\mathrm{Cov}\left[\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}[X_i, Y_j].$$

Proof

$$\begin{aligned}\mathrm{Cov}\left[\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{n} X_i - \mathrm{E}\left[\sum_{i=1}^{n} X_i\right]\right)\left(\sum_{j=1}^{m} Y_j - \mathrm{E}\left[\sum_{i=1}^{n} Y_j\right]\right)\right] \\ &= \mathrm{E}\left[\sum_{i=1}^{n}\left(X_i - \mu_{X_i}\right)\sum_{j=1}^{m}(Y_j - \mu_{Y_j})\right]\end{aligned}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{E}\left[(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})\right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{Cov}[X_i, Y_j]. \qquad \square$$

Corollary 9.2 *If $X_1, X_2, \ldots, X_n$ are random variables, then*

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}[X_i, X_j] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}[X_i, X_j].$$

In particular, $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y]$.

Thus, from the Corollary 9.1, we have Corollary 9.3.

Corollary 9.3 *If X and Y are independent, then*

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

9.2 CORRELATION

In the Black Spruce Case Study, the biologist made his measurements using centimeters. Now, suppose he decides to convert his measurements to inches.

Let X, Y denote the heights and diameters in centimeters, while X', Y' denote the measurements in inches. There are 0.3937 in. to the centimeter, so

$$\begin{aligned} \mathrm{Cov}[X', Y'] &= \mathrm{Cov}[0.3937X,\ 0.3937Y] \\ &= \mathrm{E}\,[(0.3937X)(0.3937Y)] - \mathrm{E}\,[0.3937X]\,\mathrm{E}\,[0.3937Y] \\ &= 0.3937^2\,\mathrm{Cov}[X, Y]. \end{aligned}$$

Thus, the covariance decreases by a factor of 0.155.

But changing the measurement units really does not affect how strongly the variables are related. We could even do a scatter plot that would look exactly the same, except for axis labels. We need a measure of the relationship that is unitless.

Definition 9.2 The *correlation coefficient* of random variables X and Y is

$$\rho(X, Y) = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

||

Correlation is not affected by adding constants or multiplying by positive constants.

Proposition 9.2 *Let $X' = a + bX$ and $Y' = c + dY$ for constants a, $b \geq 0$, and c, $d \geq 0$. Then*

$$\rho(X', Y') = \rho(a + bX, c + dY) = \rho(X, Y).$$

Proof Exercise. □

Since correlation is not affected by linear transformations, the correlation may be expressed in terms of the correlation of standardized variables, which in turn equals the covariance of the standardized variables.

Corollary 9.4

$$\rho(X, Y) = \rho\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) = \mathrm{Cov}\left[\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right].$$

The correlation is bounded by −1 and 1.

Proposition 9.3 $|\rho(X, Y)| \leq 1$.

Proof Let $Z_X = (X - \mu_X)/\sigma_X$ and $Z_Y = (Y - \mu_Y)/\sigma_Y$, so $\rho(X, Y) = \mathrm{Cov}[Z_X, Z_Y]$.

$$\begin{aligned}\mathrm{Var}[Z_X \pm Z_Y] &= \mathrm{Var}[Z_X] + \mathrm{Var}[Z_Y] \pm 2\mathrm{Cov}[Z_X, Z_Y] \\ &= 2 \pm 2\rho(X, Y).\end{aligned}$$

But, variances are always nonnegative, so

$$\begin{aligned}2 \pm 2\rho(X, Y) &\geq 0 \\ \pm\rho(X, Y) &\geq -1 \\ |\rho(X, Y)| &\leq 1.\end{aligned}$$

□

Proposition 9.4 $|\rho(X, Y)| = 1$ *if and only if $Y = a + bX$ for some real numbers a and b.*

Proof From the proof of the previous proposition, if $\rho(X, Y) = 1$, then $\mathrm{Var}[Z_X - Z_Y] = 0$.

Thus, $Z_X - Z_Y = C$ for some constant C.

$$\begin{aligned}Z_Y &= Z_X - C, \\ \frac{Y - \mu_Y}{\sigma_Y} &= \frac{X - \mu_X}{\sigma_X} - C, \\ Y &= \frac{\sigma_Y}{\sigma_X}X - \sigma_Y C + \mu_Y - \mu_X\frac{\sigma_Y}{\sigma_X}, \\ Y &= aX + B.\end{aligned}$$

Similarly, for $\rho(X, Y) = -1$.

We leave the converse as an exercise. □

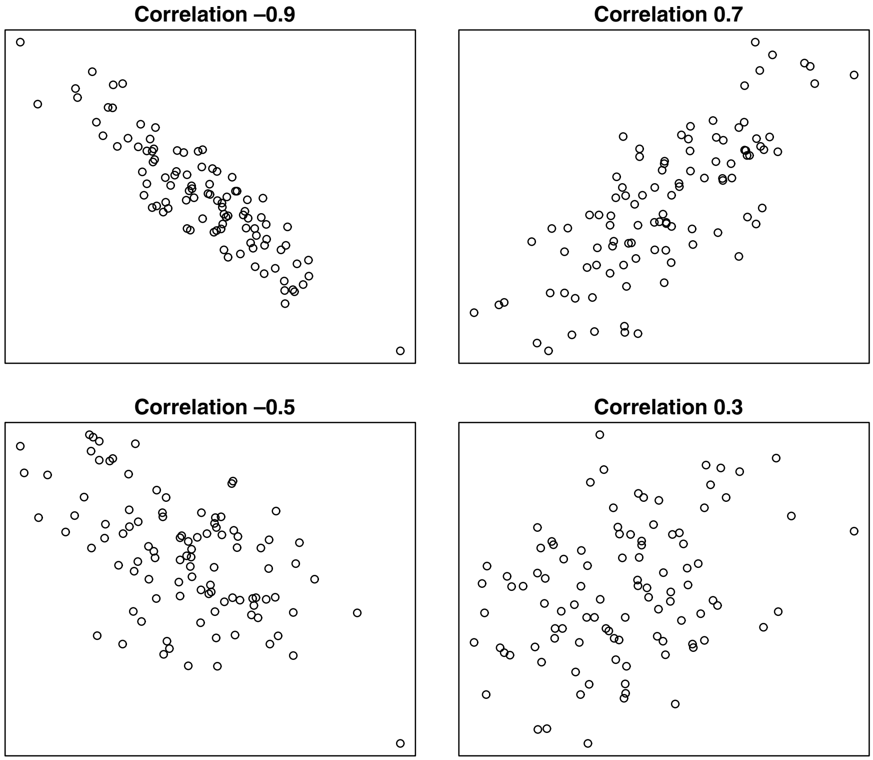

FIGURE 9.4 Examples of correlation.

The *sample correlation* for data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is obtained by plugging in sample moments for population moments (Figure 9.4). The population correlation is

$$\rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

The sample correlation is

$$\begin{aligned} r &= \frac{(1/n)\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(1/n)\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{(1/n)\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{(1/n)\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(1/(n-1))\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{(1/(n-1))\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \qquad (9.1) \end{aligned}$$

It does not matter whether you use a divisor of n or $n - 1$, provided you are consistent.

Remark Many textbooks give the following algebraically equivalent form of the correlation as a calculation aid:

$$r = \frac{(1/n)\sum_{i=1}^{n} x_i y_i) - \bar{x}\bar{y}}{\sqrt{((1/n)\sum_{i=1}^{n} x_i^2) - \bar{x}^2}\sqrt{((1/n)\sum_{i=1}^{n} y_i^2) - \bar{y}^2}}.$$

We advise against using this version since it is inaccurate due to roundoff error—the variances in the denominator can end up negative! Try looking up "software numerical accuracy" in your favorite web search engine for some background or see McCullough (2000). ||

R Note:

The command `cor` computes the correlation between two variables.

```
> plot(Spruce$Di.change, Spruce$Ht.change)  # x first, then y
> cor(Spruce$Di.change, Spruce$Ht.change)
[1]  0.9021
```

9.3 LEAST-SQUARES REGRESSION

We introduced correlation as a numeric measure of the strength of the linear relationship between two numeric variables. We now characterize a linear relationship between two variables by determining the "best" line that describes the relationship.

What do we mean by "best"? In most statistical applications, we pick the line $y = a + bx$ to make the vertical distances from observations to the line small, as shown in Figure 9.5. The reason using vertical distances is that we typically use one

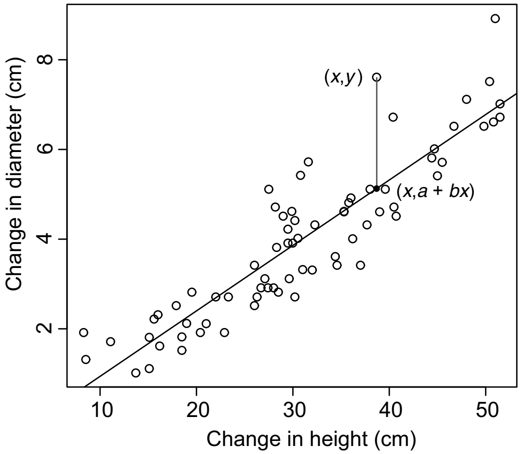

FIGURE 9.5 "Best fit" line.

variable, the x variable, to predict or explain the other (y) and we try to make the prediction errors as small as possible.

Next, we need some way to measure the overall error, taking into account all vertical distances. The most natural choice would be to add the distances,

$$\sum_i |y_i - (a + bx_i)|, \tag{9.2}$$

and in some applications this is a good choice. But more common is to choose a and b to minimize the *sum of squared distances*

$$g(a, b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2. \tag{9.3}$$

To minimize, we set the partial derivatives equal to zero:

$$\begin{aligned} \frac{\partial g}{\partial a} &= 2\sum_{i=1}^{n}(y_i - a - bx_i)(-1) = 0, \\ \frac{\partial g}{\partial b} &= 2\sum_{i=1}^{n}(y_i - a - bx_i)(-x_i) = 0, \end{aligned}$$

and solve for a and b; this simplifies to

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{9.4}$$

$$a = \bar{y} - b\bar{x}. \tag{9.5}$$

The line $\hat{y} = a + bx$ is called the *least-squares regression* line. In practice, we use statistical software to calculate the coefficients.

Example 9.2 For the Spruce data, let x denote height change and y denote the diameter change. Then, the least-squares line is

$$\hat{y} = -0.519 + 0.146x.$$

For every centimeter increase in the change in height, there is an associated increase of 0.146 cm in the change in diameter.

None of the seedlings in the study grew 25 cm over the course of 5 years, so we compute $\hat{y} = -0.519 + 0.149 \times 25 = 3.206$; that is, we predict that a seedling growing 25 cm in height would grow 3.206 cm in diameter. □

Definition 9.3 For any x, let $\hat{y} = a + bx$, then $\hat{y}$ is called a *predicted value* or a *fitted value*. ||

Note that Equation 9.5 can be written as $\bar{y} = a + b\bar{x}$, which implies that $(\bar{x}, \bar{y})$ lies on the least-squares line $y = a + bx$.

Proposition 9.5 $(1/n)\sum_{i=1}^{n}\hat{y}_i = \bar{y}$. *That is,* $\bar{\hat{y}} = \bar{y}$. *The mean of the predicted y's is the mean of the observed y's.*

Proof Exercise. □

Next, we see that there is a relationship between correlation and least-squares regression: in particular, the slope of the least-squares line is proportional to the correlation.

Let

$$\mathrm{ss}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{9.6}$$

$$\mathrm{ss}_y = \sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{9.7}$$

$$\mathrm{ss}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{9.8}$$

Then, from Equation 9.4, we re-express the estimated slope as

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{ss}_{xy}}{\mathrm{ss}_x}.$$

We can also re-express the correlation (page 253) as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{ss_{xy}}{\sqrt{ss_x ss_y}}$$

Therefore, $\mathrm{ss}_{xy} = r\sqrt{\mathrm{ss}_x}\sqrt{\mathrm{ss}_y}$, so we have

$$\begin{aligned} b &= r\frac{\sqrt{ss_x}\sqrt{ss_y}}{ss_x} \\ &= r\frac{\sqrt{ss_y}}{\sqrt{ss_x}} \\ &= r\frac{(1/(n-1))\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sqrt{(1/(n-1))\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\ &= r\frac{s_y}{s_x} \end{aligned}$$

where s_x and s_y denote the sample standard deviations for x and y.

Example 9.3 The correlation between the diameter change and height change is 0.9021. The standard deviation of the diameter changes and height changes are 1.7877 and 11.0495, respectively. Thus, $b = 0.9021 \times 1.7877/11.0495 = 0.146$, which agrees with what we obtained on page 255. □

LEAST-SQUARES REGRESSION

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n observations. The least-squares regression line is $\hat{y} = a + bx$, where

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{9.9}$$

$$a = \bar{y} - b\bar{x}. \tag{9.10}$$

In addition,

$$b = r\,\frac{s_y}{s_x}, \tag{9.11}$$

where r is the correlation and s_x and s_y are the standard deviations of the x_i's and y_i's, respectively. The variable being predicted, y, is called the "outcome", "response", or fitted variable. The variable used for predicting is called the "predictor" or "explanatory" variable.

Some authors refer to the y and x variables as "dependent" and "independent," respectively. However, this can be confused with independence and dependence of random variables.

R Note:

```
> spruce.lm <- lm(Di.change~Ht.change, data=Spruce)
> spruce.lm
...
Coefficients:
(Intercept)    Ht.change
    -0.5189       0.1459
> plot(Spruce$Ht.change, Spruce$Di.change)
> abline(spruce.lm)
```

To obtain the predicted values,

```
> fitted(spruce.lm) # or predict(spruce.lm)

        1         2         3         4         5         6
6.0488081 4.7644555 3.8595707 4.7060758 4.6331013 5.9612386
...
```

To calculate sums of squares in R, a shortcut is to use the relationship $\mathrm{ss}_x = (n-1)s_x^2$, where the sample variance s_X^2 is calculated in R using var. For example, to find $\sum_{i=1}^{72}(x_i - \bar{x})^2$ for the height change variance in the Spruce data set,

```
 > (nrow(Spruce) -1 ) * var(Spruce$Ht.change)
[1] 8668.38
```

9.3.1 Regression Toward the Mean

If x and y are perfectly correlated ($r = 1$), then the slope is s_y/s_x; so every change of one standard deviation in x results in a change of one standard deviation of y. But if $r \neq 1$, then for a change of one standard deviation in x, the vertical change is less than one standard deviation of y; so $\hat{y}$ is less responsive to a change in x. If $\rho = 0$, then the regression line is flat. See Figure 9.6.

This phenomenon is the origin of the name "regression." Sir Francis Galton studied the heights of parents and children. He found that although the children of tall parents were tall, on average, they were less so than their parents. Similarly, children of short parents averaged less short than their parents. The data, from Verzani (2010), are shown in Figure 9.7. He termed this "regression toward mediocrity," with the implication that in the long run, everyone would become of average height. This is of course not true (just look at the coauthors of this book!)[1] The regression line does

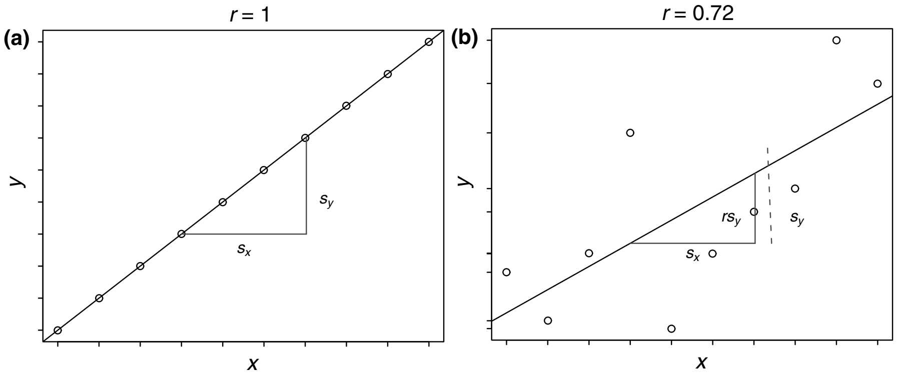

FIGURE 9.6 (a) The relationship between the regression slope, s_y and s_x with perfectly correlated data (b) The relationship without perfect correlation. For every change in one standard deviation of x, $\hat{y}$ changes less than one standard deviation in y.

[1] One of us is 5 ft 2 in., the other 6 ft 3 in.!

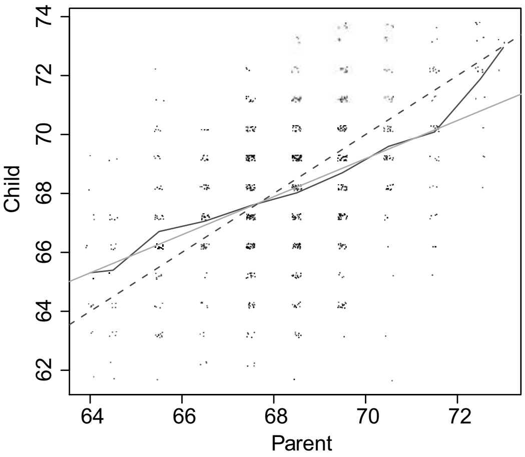

FIGURE 9.7 Heights of parents and children. The data are *jittered*—a small amount of random noise is added so that multiple points with the same x and y are visible. The x-axis contains the "midparent" height—the average of the father's height and 1.08 times the mother's height. The y-axis contains the average adult child's height, with female heights multiplied by 1.08. The dashed line is the 45° line. The solid line is the least-squares regression line and the zigzag line connects the mean child height with each midparent height.

not give the whole picture. There is also substantial variability above and below the regression line. Some offspring of average-height parents end up tall or short, so over time the variability above and below the mean remains roughly constant.

9.3.2 Variation

Let us look at the different components of variation in regression in more detail to see where Galton went wrong. We will use the Spruce data.

We partition the difference between an observed y value and the mean of the observed y values into two parts (Figure 9.8),

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

By algebraic manipulation, we can show that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2. \tag{9.12}$$

This is related to the Pythagorean theorem for the sides of triangles; the vectors $(y_1 - \hat{y}_i, \ldots, y_n - \hat{y}_i)$ and $(\hat{y}_1 - \bar{y}, \ldots, \hat{y}_n - \bar{y})$ are orthogonal. We say that the *total variation* (of the y's) equals the *variation of the residuals* plus the *variation of the predicted values*.

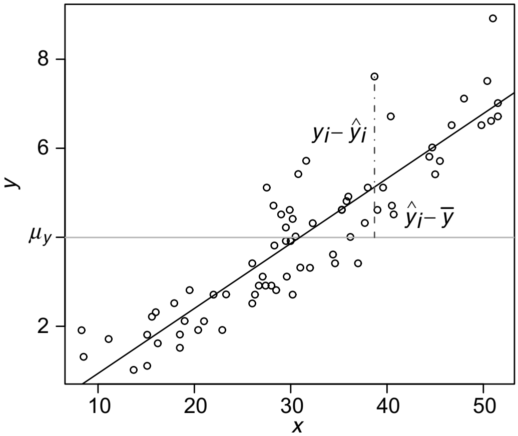

FIGURE 9.8 The partitioning of the variability. The horizontal line is the mean of the observed *y*'s.

Also,

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n}(a + bx_i - (a + b\bar{x}))^2 \\ &= \sum_{i=1}^{n}\big(b(x_i - \bar{x})\big)^2 \\ &= b^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= b^2 \mathrm{ss}_x \\ &= r^2 \mathrm{ss}_y. \end{aligned} \tag{9.13}$$

Thus,

$$\begin{aligned} r^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{1/(n-1)}{1/(n-1)}\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{\text{variance of predicted } y\text{'s}}{\text{variance of observed } y\text{'s}}, \end{aligned}$$

r^2 is the proportion of the variation of the observed $y's$ that is explained by the regression line.

R-SQUARED = PROPORTION OF VARIANCE EXPLAINED

The R-squared coefficient is the square of the correlation r^2. In other words, $r^2 \times 100\%$ of the variance, or variation, of y is explained by the linear regression. We say that r^2 is the proportion of the variance explained by the regression model.

Example 9.4 For the Spruce Case Study, $r^2 = 0.9021^2 = 0.8138$, so about 81% of the variability in the diameter changes is explained by this model. □

9.3.3 Diagnostics

We can fit a linear regression line to any two variables, whether or not there is a linear relationship. But for predictions from the line to be accurate, the relationship should be approximately linear. A linear relationship is also required (together with some additional assumptions) for the standard errors and confidence intervals in Section 9.4 to be correct. So our next step is to check if it is appropriate to model the relationship between these two variables with a straight line.

Definition 9.4 Let (x_i, y_i) be one of the data points. The number $y_i - \hat{y}_i$ is called a *residual.*

A *residuals plot* is a plot of $y_i - \hat{y}_i$ against x_i for $i = 1, 2, \ldots, n$. ||

The residual is the difference between an observed y value and the corresponding fitted value; it provides information on how far off the least-squares line is in predicting the y_i value at a particular data point x_i. If the residual is positive, then the predicted value is an underestimate, whereas if the residual is negative, then the predicted value is an overestimate.

Example 9.5 In the example on page 255, we computed the least-squares line $\hat{y} = -0.519 + 0.146x$. Thus, for the first tree in the data set, the predicted diameter change is $\hat{y} = -0.519 + 0.146 \times 5.416 = 6.049$; so the corresponding residual is $5.416 - 6.049 = -0.633$ (Table 9.1). The least-squares line overestimates the diameter change for this tree. □

The plot of residuals against the predictor variable $(x_i, y_i - \hat{y}_i)$ provides visual information on the appropriateness of a straight line model (Figure 9.9). Ideally, points should be scattered randomly about the reference line $y = 0$.

TABLE 9.1 Partial View of `Spruce` Data

Tree	Height Change	Diameter Change	Predicted $\hat{y}$	Residual
1	45.0	5.416	6.049	−0.633
2	36.2	4.009	4.764	−0.755
3	30.0	3.914	3.859	0.054
4	35.8	4.813	4.706	0.106
5	35.3	4.6125	4.633	−0.021

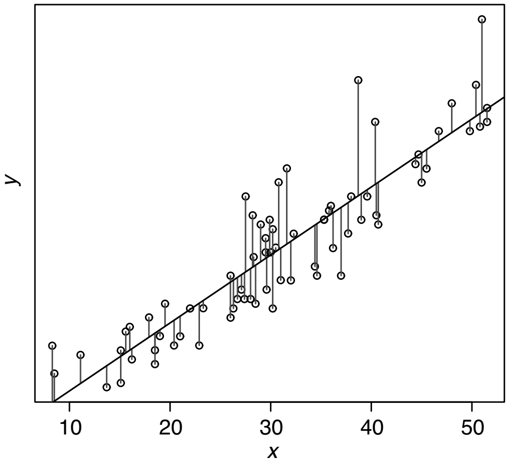

FIGURE 9.9 Residuals are the (signed) lengths of the line segments drawn from each observed y to the corresponding predicted $\hat{y}$.

Residuals plots are useful for the following:

- Revealing curvature—that is, for indicating that the relationship between the two variables is not linear.
- Spotting outliers.

If outliers are noticed, then you should check if they are influential: Does their removal dramatically change the model?

See Figure 9.10 for examples illustrating these points.

For the Spruce data, the residuals plot reveals that the distribution of the residuals is positively skewed—most residuals are negative but small, and there are a smaller number of positive residuals, but they are larger (Figure 9.11). This does not mean that a linear relationship is inappropriate, but it does cause problems for some methods that assume that residuals are normally distributed.

A bigger issue is whether there may be curvature. Here, some caution is needed—the human eye can be good at creating patterns out of nothing (the ancient star constellations are one example). Here, ignoring the most negative residual, the residuals appear to have a curved bottom. This impression is reinforced by a second set of points slightly higher, also curved upward. But these may be purely random artifacts. There do appear to be a large number of negative residuals in the middle—but there are also a number of even bigger positive residuals in the middle. There do appear to be a lack of large negative residuals on both sides—but this may be simply because there are fewer observations on each side.

A more effective way to judge curvature is to add a *scatter plot smooth* to the plot, a statistical procedure that tries to find a possibly curved relationship in the data. There are many such procedures, for example, the "connect-the-dot" procedure shown in

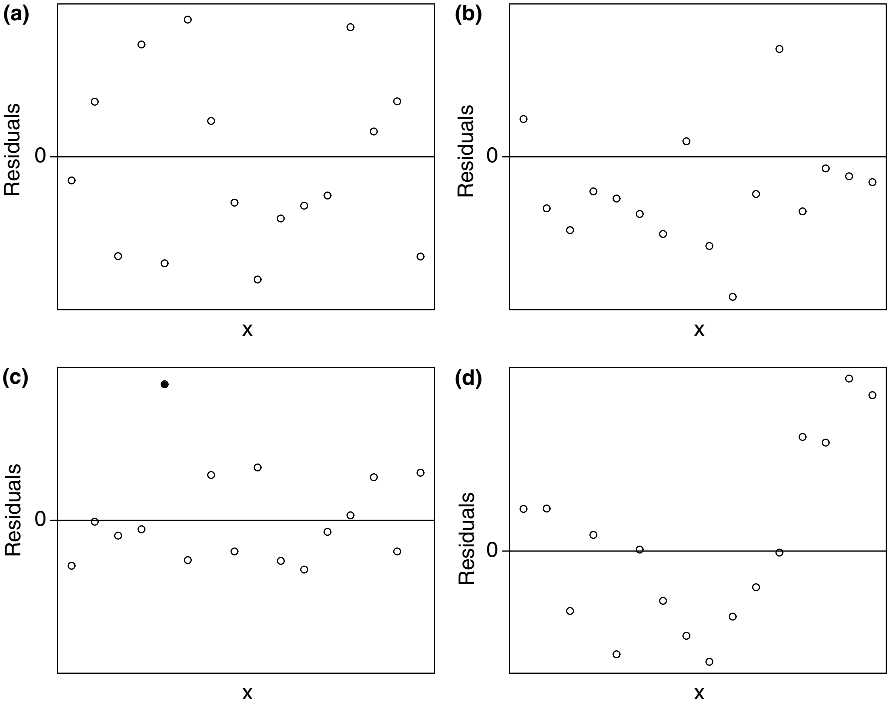

FIGURE 9.10 Examples of residual plots. (a) A good straight line fit. (b) The regression line is consistently overestimating the *y* values. (c) An outlier. (d) Curvature—a straight line is not an appropriate model for the relationship between the two variables.

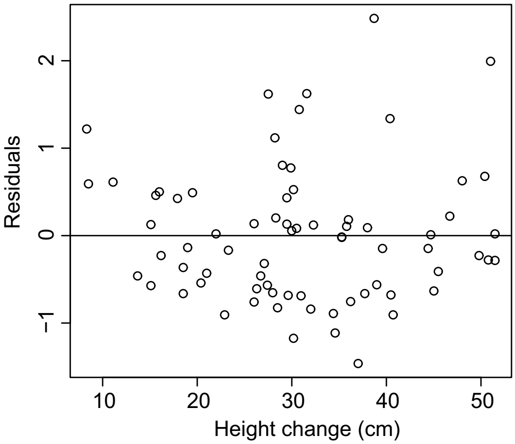

FIGURE 9.11 Residuals plot for the Spruce data.

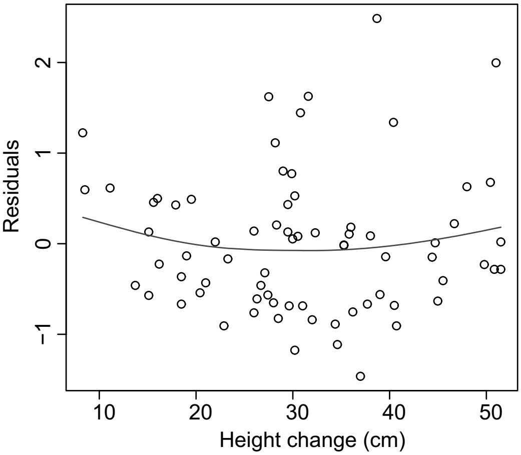

FIGURE 9.12 Residuals plot for the Spruce data, with a scatter plot smooth indicating slight curvature.

Figure 9.7. Figure 9.12 shows another procedure, a *smoothing spline*, a mathematical analog of the old-time draftsman's device—a "spline" was a thin piece of wood that was bent as needed for tracing smooth curves. Calculating these is beyond the scope of this book, but most statistical software offers options to create these smoothers.

The smoother added to the residuals plot (Figure 9.12) indicates slight curvature that should lead the researcher to some more investigation. Indeed, for these data (refer to the description on page 8), the observations do not come from one population since the seedlings were planted under different conditions.

R Note

Spruce regression example, continued from page 257.

The `resid` command gives the residuals for a regression.

```
plot(Spruce$Ht.change, resid(spruce.lm), ylab = "Residuals")
abline(h=0)
lines(smooth.spline(Spruce$Ht.change, resid(spruce.lm), df = 3),
      col="blue")
```

Example 9.6 Here are sugar and fat content (in grams per half cup serving) for a random sample of 20 brands of vanilla ice cream.

Sugar	15	13	20	23	11	21.5	12	23	23.0	19
Fat	8	8	16	14	7	15.5	8	21	15.5	16
Sugar	19.0	19	21.8	17	20	17	20	16	11	12
Fat	4.5	13	13.5	12	16	8	15	8	7	6

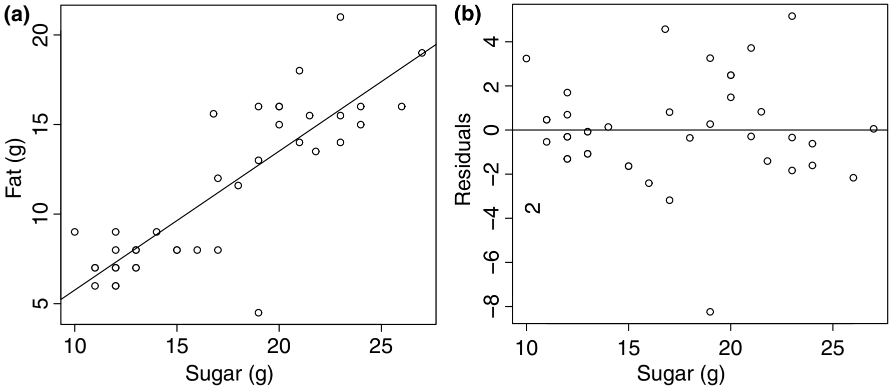

FIGURE 9.13 (a) Scatter plot of fat content against sugar content in ice cream. (b) Residuals plot for the least-squares regression.

The mean and standard deviation of the fat values are 11.6 and 4.5236 g, respectively, while those for sugar are 17.665 and 4.1323 g. The correlation between fat and sugar is 0.792.

For a least-squares line of fat on sugar, the slope of the least-squares line is $b = 0.792 \times (4.5236/4.1323) = 0.867$, while the intercept is $a = 11.6 - 0.867 \times 17.655 = -3.707$. Thus, the equation is $\widehat{\texttt{fat}} = -3.707 + 0.867 \cdot \texttt{sugar}$.

For every gram increase in sugar, there is an associated increase of 0.867 g in fat content.

About 62.7% of the variability in fat content can be explained by this least-squares line.

The residuals plot (Figure 9.13), reveals a large negative outlier at about 19 g of sugar.

Removing this observation results in a least-squares line of $\widehat{\texttt{Fat}} = -3.912 + 0.903 \cdot \texttt{Sugar}$. The slope of the regression line does not change very much, although the proportion of the variability in fat that is explained by the model does change (78.3%). □

9.3.4 Multiple Regression

The ideas of linear regression can be applied in the case where there are multiple predictors; then instead of a prediction equation $\hat{y} = a + bx$, the typical equation is of the form $\hat{y} = a + b_1 x_1 + \cdots , b_p x_p$, where p is the number of predictors. In some cases, such as when Google uses regression to improve web search answers, the number of predictors can be in millions.

One special case is when there are multiple groups in the data. In the Spruce example, there are two additional predictors—whether the tree was fertilized or not and whether the tree faced competition. One relatively simple model in this case is

$$\hat{y} = 0.51 + 0.104 \cdot \texttt{Ht.change} + 1.03 \cdot \texttt{Fertilizer} - 0.49 \cdot \texttt{Competition},$$

where we convert the categorical predictors to *dummy variables*— 1 if Fertilizer = "F" and 0 for "NF"; 1 if Competition = "C" and 0 for "NC." This equation suggests that trees that grew taller tended to grow thicker; that for a given change in height, these trees that were fertilized tended to grow thicker; and that for a given change in Height, trees that were in competition did not grow as thick—it seems they spend more energy growing taller rather than thicker.

This model has a single slope and different intercepts for the four groups defined by Fertilizer and Competition. Other models can be fit, for example, we may allow different slopes in different groups.

The formulas for calculating multiple regression coefficients are beyond the scope of this course, but can be performed using statistical software; for example, the `R` command for the model above is

```
lm(Di.change ~ Ht.change + Fertilizer + Competition, data = Spruce).
```

For more about multiple regression, see Kutner et al. (2005), Weisberg (2005), and Draper and Smith (1998).

9.4 THE SIMPLE LINEAR MODEL

The least-squares regression line is the "best fit" line for a set of n ordered pairs, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ irrespective of whether this set represents a population or is a sample from a larger population. If this set is just a sample from a larger population, then the least-squares line is an estimate of a "true" least-squares line fit to the entire population.

Thus, in the case of a sample, after we calculate sample estimates such as the sample correlation or regression slope, we often want to quantify how accurate these estimates are using, for example, standard errors, confidence intervals, and significance tests. We will do so here using our usual bag of tricks—permutation tests, bootstrapping, and formulas based on certain assumptions.

Example 9.7 In figure skating competitions, skaters perform twice: a 2 min short program and a 4 min free skate program. The scores from the two segments are combined to determine the winner. What is the relationship, if any, between the score on the short program and the score on the free skate portion? We will investigate this by looking at the scores of 24 male skaters who competed in the 2010 Olympics in Vancouver. We will consider these observations as a sample taken from a larger population of all Olympic-level male figure skaters.

Figure 9.14a displays the scores, together with the least-squares regression line. The scores are highly correlated, with a correlation of 0.84. The regression line for predicting the score on the free skate program, based on the short program scores, is $\widehat{\texttt{Free}} = 7.97 + 1.735 \cdot \texttt{Short}$.

But how accurate are these numerical results? Are the correlation and regression slope significantly different from zero? What are standard errors or confidence intervals for the correlation, slope, or the prediction for $\hat{Y}$ at a particular value of x?

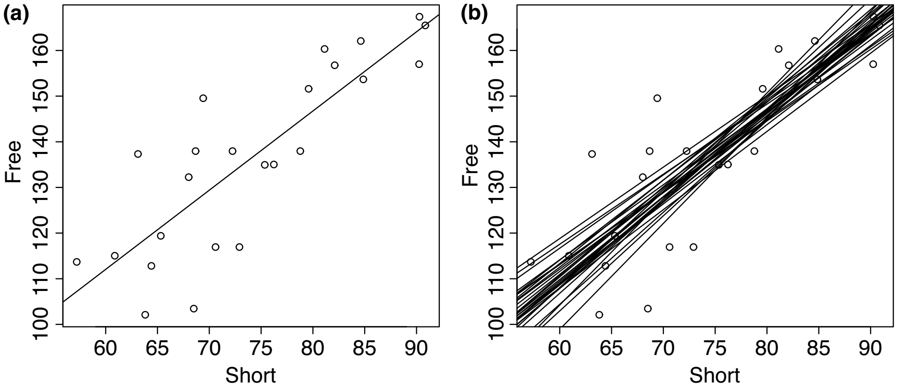

FIGURE 9.14 Scores of the 24 finalists in the 2010 Olympics men's figure skating contest for the short program and free program. (a) The least-squares regression line. (b) The regression lines from 30 bootstrap samples.

Figure 9.14b shows regression lines from 30 bootstrap samples. This gives a useful impression of the variability of the regression predictions. The predictions are most accurate for values of x near the center, and become less accurate as we extrapolate more in either direction. □

In the following sections, we will obtain standard errors and confidence intervals in two different ways, first using formulas and then using resampling. These are complementary—the bootstrap is better at visual impressions, while the formula approach gives mathematical expressions that quantify the visual impressions.

The least-squares regression line is derived without making any assumption about the data. That is, we do not require one or both variables to be an independent random sample drawn from any particular distribution or even for the relationship to be linear. However, in order to draw inferences or calculate confidence intervals, we need to make some assumptions.

ASSUMPTIONS FOR THE SIMPLE LINEAR MODEL

Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ be points with fixed x values and random Y values, independent of other Y's, where the distribution of Y given x is normal, $Y_i \sim N(\mu_i, \sigma^2)$, with

$$\mu_i = \mathrm{E}[Y_i] = \alpha + \beta x_i,$$

for constants α and β.
In other words, we assume the following:

- x values are fixed, not random.
- Relationship between the x values and the means μ_i is linear, $\mathrm{E}[Y_i|x_i] = \alpha + \beta x_i$.
- Residuals $\epsilon_i = Y_i - \mu_i$ are independent.

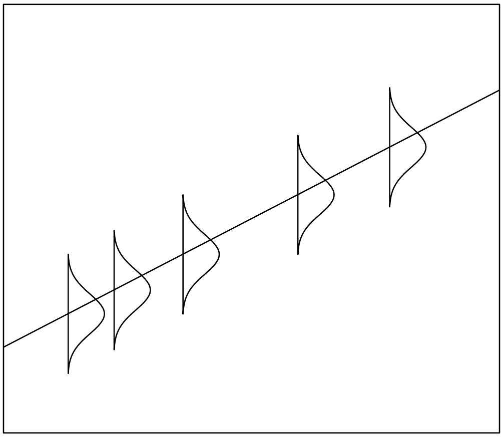

FIGURE 9.15 Linear regression assumptions: each Y_i is normal with mean $\alpha + \beta x_i$ and constant variance. Assumptions not shown are that the x values are fixed and observations are independent.

- Residuals have constant variance.
- Residuals are normally distributed.

See Figure 9.15.

In practice, the linear and independence assumptions are very important, the others less so—the reasons will be explained in Section 9.4.3.

Theorem 9.2 *Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ satisfy the assumptions for a linear model. Then, the maximum likelihood estimates are*

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}, \tag{9.14}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \tag{9.15}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum (Y_i - \hat{Y})^2. \tag{9.16}$$

Proof Since the Y_i's are normally distributed, we can form the likelihood function:

$$\begin{aligned} L(\alpha, \beta, \sigma) &= \prod_1^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_i - \mu_i)^2/(2\sigma^2)} \\ &= \prod_1^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_i - \alpha - \beta x_i)^2/(2\sigma^2)} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-1/(2\sigma^2) \sum_1^n (Y_i - \alpha - \beta x_i)^2}. \end{aligned}$$

Thus, the log-likelihood is

$$\ln(L) = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_1^n (Y_i - \alpha - \beta x_i)^2.$$

We take the partial derivatives with respect to α, β, and σ, respectively:

$$\frac{\partial \ln(L)}{\partial \alpha} = \frac{-1}{\sigma^2}\sum_1^n (Y_i - \alpha - \beta x_i)(-1),$$

$$\frac{\partial \ln(L)}{\partial \beta} = \frac{-1}{\sigma^2}\sum_1^n (Y_i - \alpha - \beta x_i)(-x_i),$$

$$\frac{\partial \ln(L)}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^2}\sum_1^n (Y_i - \alpha - \beta x_i)^2.$$

Equating each partial derivative to 0 and doing some algebra yield the maximum likelihood estimates:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2},$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

□

Note that these maximum likelihood estimates for β and α are exactly the same as the least-squares estimates (page 257).

We state without proof:

Theorem 9.3 *Suppose $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ satisfy the assumptions for a linear model. Then,*

1. *$\hat{\sigma}^2$, $\hat{\beta}$, and $\bar{Y}$ are mutually independent.*
2. *$n\,\hat{\sigma}^2/\sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom.*

We would not actually use $\hat{\sigma}^2$ much; instead, we will use an unbiased version:

Corollary 9.5

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum (Y_i - \hat{Y}_i)^2$$

is an unbiased estimator of σ^2.

We call S the *residual standard deviation* or *residual standard error*. Note that it is computed with a divisor of $n-2$, corresponding to the degrees of freedom (this is because the means are affected by two estimated parameters $\hat{\alpha}$ and $\hat{\beta}$).

Proof From Theorem B.12, the expected value of a chi-square distribution with $n-2$ degrees of freedom is $n-2$. Thus, from Theorem 9.3 (2), $\mathrm{E}\left[n\,\hat{\sigma}^2/\sigma^2\right] = n-2$, or upon rearranging,

$$E[S^2] = \mathrm{E}\left[\frac{n}{n-2}\hat{\sigma}^2\right] = \sigma^2.$$

□

9.4.1 Inference for α and β

We now consider some properties of the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ of the intercept and slope for the linear model $\mathrm{E}[Y] = \alpha + \beta x$.

Theorem 9.4 *Let* $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ *satisfy the assumptions for a simple linear model, and let* $\hat{\alpha}$ *and* $\hat{\beta}$ *denote the estimators of* α *and* β*, respectively. Then,*

1. $\hat{\alpha}$ *and* $\hat{\beta}$ *are normal random variables,*
2. $\mathrm{E}[\hat{\alpha}] = \alpha$ *and* $\mathrm{E}\left[\hat{\beta}\right] = \beta$,
3. $\mathrm{Var}[\hat{\beta}] = \sigma^2/\mathrm{ss}_x$,
4. $\mathrm{Var}[\hat{\alpha}] = \sigma^2\left[1/n + (x-\bar{x})^2/\mathrm{ss}_x\right]$,

where ss_x *is given in Equation 9.6.*

Proof

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\ &= \frac{1}{\mathrm{ss}_x}\sum_{i=1}^{n}((x_i-\bar{x})Y_i - (x_i-\bar{x})\bar{Y})\\ &= \frac{1}{\mathrm{ss}_x}\left(\sum_{i=1}^{n}(x_i-\bar{x})Y_i - \sum_{i=1}^{n}(x_i-\bar{x})\bar{Y}\right)\\ &= \frac{1}{\mathrm{ss}_x}\sum_{i=1}^{n}(x_i-\bar{x})Y_i.\end{aligned}$$

Note that $\hat{\beta}$ is a linear combination of independent normal random variables, so is also a normal random variable (Theorem A.11).

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}\right] &= \mathrm{E}\left[\frac{1}{\mathrm{ss}_x}\sum_{i=1}^{n}(x_i-\bar{x})Y_i\right] \\
&= \frac{1}{\mathrm{ss}_x}\sum_{i=1}^{n}(x_i-\bar{x})\mathrm{E}\left[Y_i\right] \\
&= \frac{1}{\mathrm{ss}_x}\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\beta x_i) \\
&= \alpha\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{\mathrm{ss}_x}+\beta\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\mathrm{ss}_x} \\
&= \beta,
\end{aligned}$$

where the last equality follows from $\sum_{i=1}^{n}(x_i-\bar{x})=0$ and $\mathrm{ss}_x=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})x_i$.

Thus, $\hat{\beta}$ is an unbiased estimator of β.

The variance is

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}] &= \mathrm{Var}\left[\frac{1}{\mathrm{ss}_x}\sum_{i=1}^{n}(x_i-\bar{x})Y_i\right] \\
&= \frac{1}{(\mathrm{ss}_x)^2}\sum_{i=1}^{n}\mathrm{Var}[(x_i-\bar{x})Y_i] \\
&= \frac{1}{(\mathrm{ss}_x)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\mathrm{Var}[Y_i] \\
&= \frac{1}{(\mathrm{ss}_x)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 \\
&= \frac{\sigma^2}{\mathrm{ss}_x}.
\end{aligned}$$

Thus, the sampling distribution of $\hat{\beta}$ is normal with mean β and variance σ^2/ss_x.

The proof for $\hat{\alpha}$ is similar. □

Now, since $\hat{\beta}$ follows a normal distribution, we can form the z statistic,

$$Z=\frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/\mathrm{ss}_x}}=\frac{\hat{\beta}-\beta}{\sigma/\sqrt{\mathrm{ss}_x}},$$

which follows a standard normal distribution.

In practice, σ is unknown, so we plug in the estimate S to obtain

$$\frac{\hat{\beta}-\beta}{S/\sqrt{\mathrm{ss}_x}}.$$

As in earlier chapters, replacing the population standard deviation with an estimate results in a t rather than standard normal distribution.

Let $\widehat{\mathrm{SE}}[\hat{\beta}] = S/\sqrt{\mathrm{ss}_x}$, the estimate of the standard error of $\hat{\beta}$; then, we have the following theorem:

Theorem 9.5 *Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ satisfy the assumptions for the simple linear model. Then,*

$$T = \frac{\hat{\beta} - \beta}{\widehat{\mathrm{SE}}[\hat{\beta}]}$$

follows a t distribution with $n - 2$ degrees of freedom.

Proof From Theorem 9.4

$$Z = \frac{\hat{\beta} - \beta}{\sigma/\sqrt{\mathrm{ss}_x}}$$

follows a standard normal distribution. Also, from Theorem 9.3,

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2}$$

has a χ^2 distribution with $n - 2$ degrees of freedom, and Z and $(n-2)S^2/\sigma^2$ are independent. Thus, from Theorem B.17, the ratio

$$\frac{Z}{\sqrt{((n-2)S^2/\sigma^2)/(n-2)}} = \frac{\hat{\beta} - \beta}{S/\sqrt{\mathrm{ss}_x}}$$

has a t distribution with $n - 2$ degrees of freedom. □

In practice, we are often interested in testing if the slope β is zero or we will want a confidence interval for β.

INFERENCE FOR β

To test the hypothesis H_0: $\beta = 0$ versus H_a: $\beta \neq 0$, form the test statistic

$$T = \frac{\hat{\beta}}{\widehat{\mathrm{SE}}[\hat{\beta}]}.$$

Under the null hypothesis, T has a t distribution with $n - 2$ degrees of freedom.

A $(1 - \alpha) \times 100\%$ confidence interval for β is given by

$$\hat{\beta} \pm q\,\widehat{\mathrm{SE}}[\hat{\beta}],$$

where q is the $1 - \alpha/2$ quantile of the t distribution with $n - 2$ degrees of freedom and $\widehat{\mathrm{SE}}[\hat{\beta}] = S/\sqrt{\mathrm{ss}_x}$.

Example 9.8 The data set `Skating2010` contains the scores from the short program and free skate for men's figure skating in the 2010 Olympics.

R Note:

```
> skate.lm <- lm(Free ~ Short, data=Skating2010)
> summary(skate.lm)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.9691    18.1175   0.440    0.664
Short         1.7347     0.2424   7.157 3.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.36 on 22 degrees of freedom
Multiple R-squared: 0.6995,     Adjusted R-squared: 0.6859
F-statistic: 51.22 on 1 and 22 DF,  p-value: 3.562e-07
```

From the R output, we obtain $S = 11.36$, the estimate for σ.

To test H_0: $\beta = 0$ versus H_A: $\beta \neq 0$, we use $t = \hat{\beta}/\widehat{\text{SE}}[\hat{\beta}] = 1.7347/0.2424 = 7.157$. We compare this to a t distribution with 22 degrees of freedom to obtain a P-value of $2 \times 1.780036 \times 10^{-7} = 3.56 \times 10^{-7}$. Thus, we conclude that $\beta \neq 0$.

To compute a 95% confidence interval for the true β, we first find the 0.975 quantile for the t distribution with 22 degrees of freedom, $q = 2.0738$. Then,

$$1.7347 \pm 2.0738 \times 0.2424 = (1.232, 2.2374).$$

Thus, we are 95% confident that the true slope β is between 1.23 and 2.34. □

Similarly, we could give a standard error for $\hat{\alpha}$ and calculate a t statistic for testing H_0: $\hat{\alpha} = 0$. Statistical software routinely provides these. We caution, however, that it is rarely appropriate to do that test. It may be tempting to test whether one can simplify a regression model by omitting the intercept. But unless you have a physical model that omits the intercept, you should include the intercept in describing a linear relationship. And even when there is such a physical model, in practice including the intercept provides a useful fudge factor for adjusting the discrepancy between theory and reality.

9.4.2 Inference for the Response

In many applications, we will be interested in estimating the mean response for a specific value of x, say $x = x_s$. If $\hat{Y}_s = \hat{\alpha} + \hat{\beta}x_s$ denotes the point estimate of $\text{E}[Y_s]$, we need the sampling distribution of $\hat{Y}_s$.

We state results for both $\bar{Y}$ and $\hat{Y}$.

Theorem 9.6 *Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ satisfy the assumptions for a simple linear model. Then,*

1. *$\bar{Y}$ is a normal random variable.*
2. $\mathrm{E}\left[\bar{Y}\right] = \alpha + \beta\bar{x}$.
3. $\mathrm{Var}[\bar{Y}] = \sigma^2/n$.
4. *$\hat{Y}_\mathrm{s}$ is a normal random variable.*
5. $\mathrm{E}\left[\hat{Y}_\mathrm{s}\right] = \mathrm{E}\left[Y_\mathrm{s}\right] = \alpha + \beta x_\mathrm{s}$.
6. $\mathrm{Var}[\hat{Y}_\mathrm{s}] = \sigma^2\left[1/n + (x_\mathrm{s} - \bar{x})^2/\mathrm{ss}_x\right]$.

Proof We leave the proof for the normality, mean, and variance of $\bar{Y}$ as an exercise. From Theorem 9.4,

$$\mathrm{E}\left[\hat{Y}_\mathrm{s}\right] = \mathrm{E}\left[\hat{\alpha} + \hat{\beta}x_\mathrm{s}\right] = \mathrm{E}\left[\hat{\alpha}\right] + \mathrm{E}\left[\hat{\beta}\right]x_\mathrm{s} = \alpha + \beta x_\mathrm{s}.$$

Using Equation 9.15, we have $\hat{Y}_\mathrm{s} = \bar{Y} + (x_\mathrm{s} - \bar{x})\hat{\beta}$, which is a linear combination of two independent normal variables. Also, by Theorems 9.3 and 9.4,

$$\begin{aligned}\mathrm{Var}[\hat{Y}_\mathrm{s}] &= \mathrm{Var}[\bar{Y}_\mathrm{s} + (x_\mathrm{s} - \bar{x})\hat{\beta}]\\ &= \mathrm{Var}[\bar{Y}_\mathrm{s}] + (x_\mathrm{s} - \bar{x})^2\mathrm{Var}[\hat{\beta}]\\ &= \frac{\sigma^2}{n} + (x_\mathrm{s} - \bar{x})^2\frac{\sigma^2}{\mathrm{ss}_x}.\end{aligned}$$

□

Again, using the residual standard error S as an estimate of σ, we have that

$$T = \frac{\hat{Y}_\mathrm{s} - \mathrm{E}\left[\hat{Y}_\mathrm{s}\right]}{S\sqrt{1/n + (x_\mathrm{s} - \bar{x})^2/\mathrm{ss}_x}}$$

follows a t distribution.

Let $\hat{\mathrm{SE}}[\hat{Y}_\mathrm{s}] = S\sqrt{1/n + (x_\mathrm{s} - \bar{x})^2/\mathrm{ss}_x}$, the estimate of the standard error of $\hat{Y}_\mathrm{s}$. We summarize without formal proof:

Theorem 9.7 *Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ satisfy the assumptions for a simple linear model. Let $x = x_\mathrm{s}$ be a specific value of the independent variable and $\hat{Y}_\mathrm{s} = \hat{\alpha} + \hat{\beta}x_\mathrm{s}$. Then,*

$$T = \frac{\hat{Y}_\mathrm{s} - \mathrm{E}\left[\hat{Y}_\mathrm{s}\right]}{\hat{\mathrm{SE}}[\hat{Y}_\mathrm{s}]}$$

follows a t distribution with $n - 2$ degrees of freedom.

CONFIDENCE INTERVAL FOR E[Y_s]

A $(1-\alpha)\times 100\%$ confidence interval for $\mathrm{E}\,[Y_{\mathrm{s}}]$ at $x = x_{\mathrm{s}}$ is given by

$$\hat{Y}_{\mathrm{s}} \pm q\,\hat{\mathrm{SE}}[\hat{Y}_{\mathrm{s}}] = \hat{Y}_s \pm q\,S\sqrt{\frac{1}{n} + \frac{(x_{\mathrm{s}}-\bar{x})^2}{\mathrm{ss}_x}},$$

where q is the $1-\alpha/2$ quantile of the t distribution with $n-2$ degrees of freedom and S is the residual standard error.

We see that the variance of $\hat{Y}_{\mathrm{s}}$ is smallest at $x_{\mathrm{s}} = \bar{x}$ and increases as $(x_{\mathrm{s}}-\bar{x})^2$ increases. In other words, the farther the x_{s} from $\bar{x}$, the less accurate the predictions.

Example 9.9 In the Olympic skating in Example 9.8, suppose we consider a short program score of 60. Then the estimate of the mean free skate score is $\hat{\mathrm{E}}[Y_{\mathrm{s}}] = 7.969 + 1.735 \times 60 = 112.07$. From the data set, we find the mean and standard deviation of the short score to be $\bar{x} = 74.132$ and $s_x = 9.771$, respectively. Thus, with $n = 24$, $S = 11.36$, and $\mathrm{ss}_x = (n-1)s_x^2 = 2195.691$, the standard error is

$$11.36\sqrt{\frac{1}{24} + \frac{(60-74.132)^2}{2195.61}} = 4.137;$$

the 0.975 quantile for the t distribution with 22 degrees of freedom $q = 2.074$. Thus, the 95% confidence interval for the mean free skate score when the short score is 60 is

$$112.07 \pm 2.074 \times 4.137 = (103.5, 120.7).$$

We conclude that with 95% confidence, the expected free skate score is between 103.5 and 120.7 when the short program score is 60 points. □

What if instead of the mean free skate score corresponding to a short score of 60, we want an estimate of an individual free skate score? In this case, we need to take into account the uncertainty in the expected value as well as the random variability of a single observation. Thus, the variance of the prediction error is

$$\mathrm{Var}[Y-\hat{Y}] = \mathrm{Var}[Y] + \mathrm{Var}[\hat{Y}] = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\mathrm{ss}_x}\right].$$

Thus, the estimate of the prediction standard error is

$$\hat{\mathrm{SE}}[\text{prediction}] = S\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\mathrm{ss}_x}}. \qquad (9.17)$$

PREDICTION INTERVAL FOR $\mathbf{Y_s}$

A $(1-\alpha)\times 100\%$ prediction interval for Y_s at $x = x_s$ is given by

$$\hat{Y}_s \pm q\,\hat{\mathrm{SE}}[\text{prediction}] = \hat{Y}_s \pm q\,S\sqrt{1 + \frac{1}{n} + \frac{(x_s-\bar{x})^2}{\mathrm{ss}_x}}, \tag{9.18}$$

where q is the $1-\alpha/2$ quantile of the t distribution with $n-2$ degrees of freedom and S is the residual standard error.

This interval is very sensitive to normality—if the residual distribution is not normal, this should not be used even if n is huge.

Example 9.10 Suppose a male skater scores 60 on his short program. Find a 95% prediction interval for his score on the free skate.

Solution Referring to Example 9.9, we have $\hat{Y}_s = 112.07$. The standard error of prediction is

$$11.36\sqrt{1 + \frac{1}{24} + \frac{199 \times 7393}{2195.61}} \approx 12.09.$$

Thus,

$$112.07 \pm 2.074 \times 12.09 = (86.995, 137.146).$$

Note that the prediction interval is much wider than the confidence interval; also see Figure 9.16. □

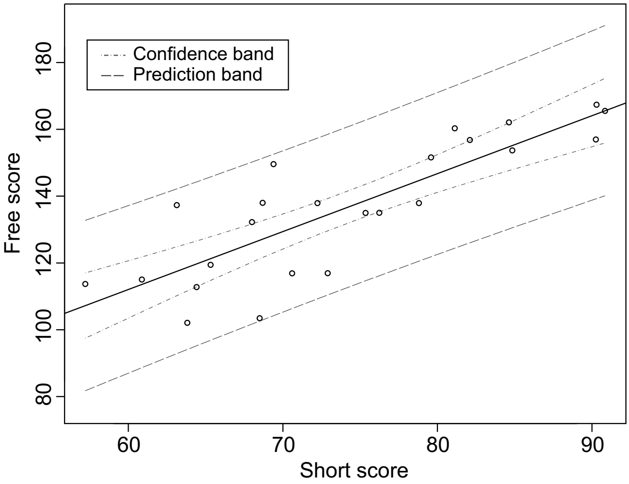

FIGURE 9.16 Plot of pointwise confidence and prediction intervals for the Olympic skating data.

Example 9.11 A wildlife biologist is interested in the relationship between the girth and length of grizzly bears. Suppose from a sample of 17 bears, she finds the following linear relationship: $\texttt{Girth} = 32.67 + 0.55\texttt{Length}$, where the measurements are in centimeters. Suppose the residual standard error and r-squared are 4.69 and 0.792, respectively. In addition, suppose the mean and standard deviation of the length measurements are 141.59 and 16.19 cm, respectively.

(a) Find the standard deviation of the girth measurements.

(b) Find a 95% confidence interval for the average girth of a bear whose length is 120 cm.

(c) Find a 95% prediction interval for the girth of an individual bear whose length is 120 cm.

Solution We are given $s = 4.69$, $r = \sqrt{0.792} = 0.890$, $\hat{\beta} = 0.55$, and $s_x = 16.19$. Thus,

(a) The standard deviation of the girth measurements is $s_y = \hat{\beta}s_x/r = 0.55 \times 16.19/0.890 = 10.01$.

(b) $Y_{\mathrm{s}} = 32.67 + 0.55 \times 120 = 98.67$. Since $\mathrm{ss}_x = (n-1)s_x^2$, we have

$$\hat{\mathrm{SE}}[\bar{Y}_{\mathrm{s}}] = 4.69\sqrt{\frac{1}{17} + \frac{(120 - 141.59)^2}{(16 \times 16.19^2)}} = 1.934.$$

The 0.975 quantile for a t distribution on 15 degrees of freedom is $q = 2.131$, so a 95% confidence interval for the mean girth of grizzlies that are 120 cm in length is $98.67 \pm 2.131 \times 1.934 = (94.5, 102.8)$ cm.

(c) The standard error of prediction is

$$\hat{\mathrm{SE}}[\text{prediction}] = 4.69\sqrt{1 + \frac{1}{17} + \frac{(120 - 141.59)^2}{(16 \times 16.19^2)}} = 5.07.$$

Thus, a 95% prediction interval for the girth of an individual grizzly that is 120 cm in length is $98 \pm 2.131 \times 5.07 = (87.9, 108.5)$ cm.

□

9.4.3 Comments About Assumptions for the Linear Model

We began Section 9.4 with certain assumptions. We now discuss how important these assumptions are.

The x Values are Fixed In practice, this assumption holds in some designed experiments, say when estimating the relationship between crop yield and fertilizer, where the amount of fertilizer applied to each plot or plant is specified in advance. It does not hold in the more common case that both the X and Y values are random, and the pairs (X_i, Y_i) are drawn at random from a joint distribution.

This assumption played a key role in derivations of the properties of $\hat{\beta}$ and other estimates. But in practice this assumption is relatively unimportant. We routinely use the output from regression printouts, even when the X values are random.

In fact, it is typically better to do the analysis as if the X values are fixed, even if they are random. In doing so, we are *conditioning on the observed information*. Information is a concept that relates to how accurately we can make estimates. For example, a larger sample size corresponds to more information and more precise estimates. In simple linear regression, the information also depends on how spread out the x values are. Recall that $\text{Var}[\hat{\beta}] = \sigma^2/\text{ss}_x$—the more spread out the x values, the more accurate the estimate of slope.

Now, what does it mean to condition on the observed information? Suppose you are planning a survey and your roommate agrees to help. Each of you will poll 100 people to end up with a total sample size of 200. However, she gets sick and cannot help. When analyzing the results of your survey and computing standard errors, should you take into account that the eventual sample size was random, with a high probability of being 200? No, you should not. You should just analyze the survey based on the amount of information you have, not what might have been. Similarly, when computing standard errors for the regression slope, it is generally best to compute them based on how spread out the X values actually are, rather than adjusting for the fact that they could have been more or less spread out.

The Relationship Between the Variables Is Linear This is critical. Suppose, for example, that the real relationship is quadratic. Then the residuals standard error is inflated, because $\sum(Y_i - \hat{Y}_i)^2$ includes not only the random deviations but also the systematic error, the differences between the line and the curve.

In practice, this linearity assumption is often violated. If the violation is small, we may proceed anyway, but if it is larger, then standard errors, confidence intervals, and P-values are all incorrect.

The Residuals are Independent This assumption is also critical. This assumption is often violated when the observations are collected over time. Often data are collected over time, and successive residuals are positively correlated, in which case the actual variances are larger than indicated by the usual formulas.

Remark This issue of variances being larger when observations are correlated arises in other contexts too. I (Hesterberg) consulted for Verizon in a case for the Public Utilities Commission of New York. I was a young guy and on the other side was an eminent statistician, who used ordinary two-sample t tests but neglected to take the correlation of the observations into account. I showed that this completely invalidated the results, using simulation to help the PUC understand. An eye-witness reported that the other side was "furious, but of course they couldn't refute any of it," and Verizon won handily. ||

The Residuals Have Constant Variance This is less important when doing inferences for $\hat{\beta}$. The assumption is often violated: in particular, we often see the residual variance increase (or decrease) with x, with an average value in the middle. Then, the differences between reality and the assumptions tend to cancel out in computing $\mathrm{Var}[\hat{\beta}]$.

However, when computing $\mathrm{Var}[\hat{Y}]$ when $x \neq \bar{x}$ or $\mathrm{Var}[\hat{\alpha}]$, this assumption does matter. We will see an example below.

The Residuals Are Normally Distributed Here we benefit from a version of the Central Limit Theorem—if the sample size is large and the information contained in ss_x is not concentrated in a small number of observations, then $\hat{\alpha}$, $\hat{\beta}$, and $\hat{Y}$ are approximately normally distributed even when the residuals are not normal and confidence intervals are approximately correct.

Prediction intervals are another story. They are a prediction for a single value, not an average, and a large sample size does not make these approximately correct if the residual distribution is nonnormal.

SUMMARY OF ASSUMPTIONS FOR LINEAR MODEL

The critical assumptions are that the relationship between the two variables is linear and that the observations are independent. The constant variance assumption can be important, but the most common violations have little effect on inferences for $\hat{\beta}$. Normality and fixed X values are relatively unimportant for confidence intervals, but normality matters when computing a prediction interval.

9.5 RESAMPLING CORRELATION AND REGRESSION

Another approach to obtaining inferences is to resample. We begin with the bootstrap for standard errors and confidence intervals and then use permutation tests for hypothesis testing.

To bootstrap, we treat these skaters as a random sample of the population of all Olympic-quality male skaters. Then, to create a bootstrap sample, we resample the skaters. For each bootstrap sample, we calculate the statistic(s) of interest.

Here is the general bootstrap procedure for two variables:

BOOTSTRAP FOR TWO VARIABLES

Given a sample of size n from a population with two variables,

1. Draw a resample of size n with replacement from the sample; in particular, draw n bivariate observations (x_i, y_i). If the observations are rows and variables are columns, we resample whole rows.
2. Compute a statistic of interest, such as the correlation, slope, or for prediction, $\hat{\mathrm{E}}[\hat{Y}] = \hat{\alpha} + \hat{\beta}x$ *at a specific value of* x.
3. Repeat this resampling process many times, say 10,000.
4. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

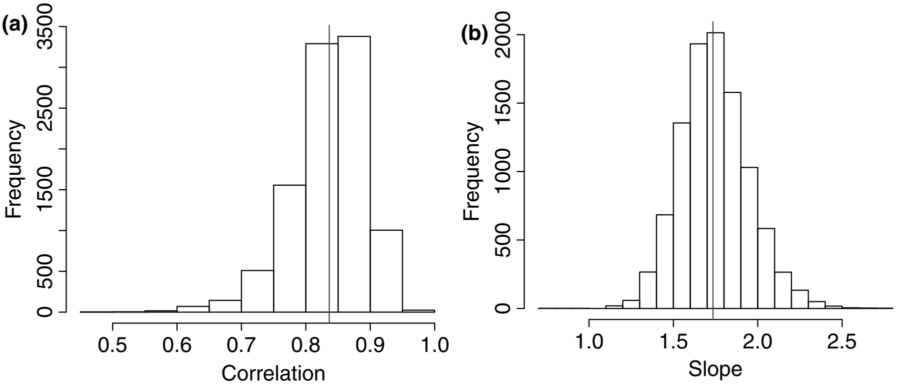

FIGURE 9.17 Bootstrap distributions of correlations (a) and slopes (b) for scores from the 2010 Olympics men's figure skating competition. The vertical lines are the corresponding statistics for the original data.

Figure 9.17a shows the bootstrap distribution for the correlation coefficient. This distribution is skewed, and indicates bias—the mean of the bootstrap distribution is smaller than the correlation of the original data. This is typical for correlations—they are biased toward zero. This is reasonable—if the original correlation is near 1 (or -1), the correlation for a sample cannot get much larger (or smaller), but can get much smaller (larger).

The bootstrap standard errors are 0.57 for the correlation and 0.20 for the slope. As before, these are the standard deviations of the bootstrap distributions.

We can use the range of the middle 95% of the bootstrap values as a rough confidence interval. For instance, in one simulation for the skating data, we found a 95% percentile confidence interval for the correlation to be $(0.70, 0.93)$ and for the slope β of the regression line to be $(1.38, 2.18)$. In Example 9.8, using the t distribution, we found a 95% confidence interval for the slope to be $(1.23, 2.34)$.

In this case, the classical interval is probably more accurate. In most of the regression problems you will encounter, the bootstrap does not offer much improvement in accuracy over classical intervals, except when the assumptions behind classical intervals are violated (one such example is the Bushmeat Case Study on page 283). Still, the bootstrap offers a way to check your work and provides graphics that may help you understand confidence intervals and standard errors in regression problems.

R Note:

The script for bootstrapping the correlation, slope, and mean response is given below. Since we want to resample the observations (x_i, y_i), we will resample the corresponding row numbers: that is, we will draw samples of size 24 (the number of skaters) with replacement from $1, 2, \ldots, 24$ and store these in the vector `index`. The command `Skating2010[index, ]` creates a new data frame from the original with rows corresponding to the rows in `index` and keeping all the columns.

```
N <- 10^4
cor.boot <- numeric(N)
beta.boot <- numeric(N)
alpha.boot <- numeric(N)
yPred.boot <- numeric(N)
n <- 24                              # number of skaters
for (i in 1:N)
{
  index <- sample(n, replace = TRUE) # sample from 1,2,...,n
  Skate.boot <- Skating2010[index, ] # resampled data

  cor.boot[i] <- cor(Skate.boot$Short, Skate.boot$Free)

  #recalculate linear model estimates
  skateBoot.lm <- lm(Free ~ Short, data = Skate.boot)
  alpha.boot[i] <- coef(skateBoot.lm)[1] # new intercept
  beta.boot[i] <- coef(skateBoot.lm)[2]  # new slope
  yPred.boot[i] <- alpha.boot[i] + 60 * beta.boot[i] # recompute Y^
}

mean(cor.boot)
sd(cor.boot)
quantile(cor.boot, c(.025,.975))

hist(cor.boot, main="Bootstrap distribution of correlation",
  xlab = "Correlation")
observed <- cor(Skating2010$Short, Skating2010$Free)
abline(v = observed, col = "blue")    # add line at original cor.
```

The commands for the summaries and plot of the slope and response results are similar.

For a specific x, bootstrapping gives a percentile confidence interval for the expected value $\mathrm{E}[Y]$.

Figure 9.18 shows the bootstrap distribution for the mean free skate program score corresponding to a short program score of 60. This distribution is centered at the original prediction and is roughly normally distributed, perhaps with a long left tail. The range of the middle 95% of the bootstrap lines (for a given x) gives the percentile confidence interval for that x. For instance, for the skating data, a 95% bootstrap percentile confidence interval for $\mathrm{E}[Y]$ at $x = 60$ is (103.3, 120.2): we are 95% confident that at $x = 60$, the corresponding mean Y value is between 103.3 and 120.2.

On the other hand, a prediction interval gives a range for an individual—for a male who scores 60 on the short program, a 95% prediction interval should have a 95%

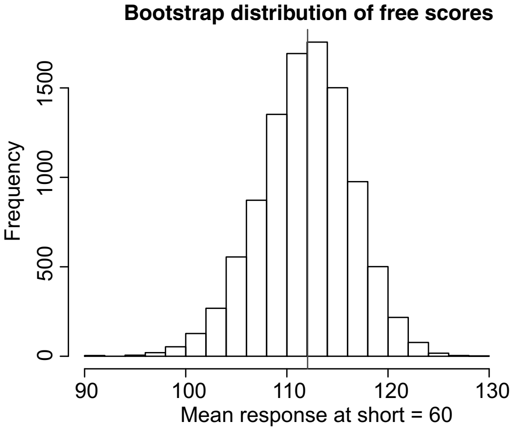

FIGURE 9.18 Distribution of bootstrapped free program scores when the short program score is 60.

chance of containing the free program score for that individual. The algorithm for a prediction interval for a response at a given x is more involved since we need to take into account the variability of an individual, and the Central Limit Theorem does not apply. See Davison and Hinkley (1997) for a way to compute prediction intervals.

9.5.1 Permutation Tests

To test whether there is a relationship between x and y or whether they are independent, we turn to permutation tests. The procedure here is to create a permutation sample by randomly permuting just one (not both) of the two variables and then computing a statistic such as correlation or slope.

PERMUTATION TEST FOR INDEPENDENCE OF TWO VARIABLES

Given a sample of size n from a population with two variables,

1. Draw a permutation resample of size n without replacement from one of the variables; keep the other variable in its original order.
2. Compute a statistic that measures the relationship, such as the correlation or slope.
3. Repeat this resampling process many times, say 9999.
4. Calculate the P-value.

For the Skating scores, the P-values are essentially zero; the probability of random chance alone producing a correlation as strong as 0.84 is minuscule, so we conclude that the two scores are not independent.

R Note:

Script for testing to see whether the short program score and the free skate score are independent. We permute just one of the variables (`Short`) while leaving `Free` fixed.

```
N <- 9999
n <- nrow(Skating2010)  # number of observations
result <- numeric(N)
observed <- cor(Skating2010$Short, Skating2010$Free)
for (i in 1:N)
{
  index <- sample(n, replace=FALSE)
  Short.permuted <- Skating2010$Short[index]
  result[i] <- cor(Short.permuted, Skating2010$Free)
}

(sum(observed <= result) + 1) / (N + 1)  # P-value
```

9.5.2 Bootstrap Case Study: Bushmeat

Many species of wildlife are going extinct due to habitat loss, climate change, and hunting. Brashares et al. (2004) found evidence of a direct link between fish supply (in kg) and subsequent demand for bushmeat[2] in Ghana. Table 9.2 and Figure 9.19 contain data of 30 years of local fish supply and biomass of 41 species in nature preserves.

TABLE 9.2 Bushmeat: Local Supply of Fish Per Capita and Biomass of 41 Species in Nature Preserves

Year	Fish	Biomass	Year	Fish	Biomass	Year	Fish	Biomass
1970	28.6	942.54	1980	21.8	862.85	1990	25.9	529.41
1971	34.7	969.77	1981	20.8	815.67	1991	23.0	497.37
1972	39.3	999.45	1982	19.7	756.58	1992	27.1	476.86
1973	32.4	987.13	1983	20.8	725.27	1993	23.4	453.80
1974	31.8	976.31	1984	21.1	662.65	1994	18.9	402.70
1975	32.8	944.07	1985	21.3	625.97	1995	19.6	365.25
1976	38.4	979.37	1986	24.3	621.69	1996	25.3	326.02
1977	33.2	997.86	1987	27.4	589.83	1997	22.0	320.12
1978	29.7	994.85	1988	24.5	548.05	1998	21.0	296.49
1979	25.0	936.36	1989	25.2	524.88	1999	23.0	228.72

[2] From Wikipedia: Bushmeat is the term commonly used for meat of terrestrial wild animals, killed for subsistence or commercial purposes throughout the humid tropics of the Americas, Asia, and Africa.

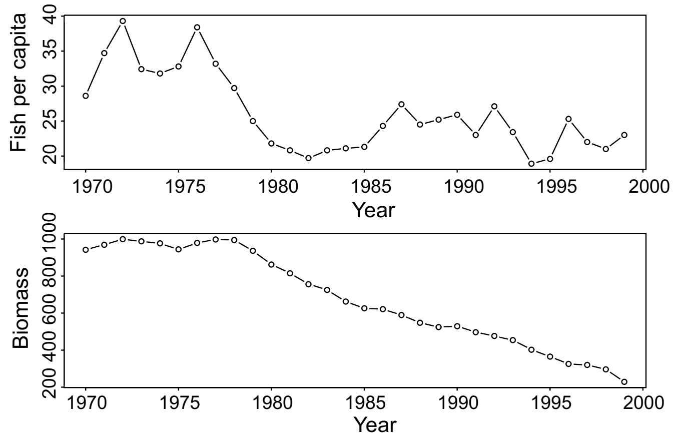

FIGURE 9.19 Bushmeat data; fish per capita and biomass of 41 species of wildlife in nature preserves for 30 years.

There is a general decline in biomass over the study period, but a closer look suggests that the decline is steeper in years with a small supply of fish. Rather than looking at the biomass for each year, we look at the percentage change. In Figure 9.20, we observe a positive relationship between fish supply and percentage change in biomass. The correlation is 0.67 and a regression of percent change in biomass against fish supply gives a slope of 0.64, suggesting that each increase of 1 kg fish per capita results in 0.64% loss of biomass and that with sufficiently large fish supplies, estimated at 33.3 (the x intercept of the least-squares line), there would be no loss in biomass.

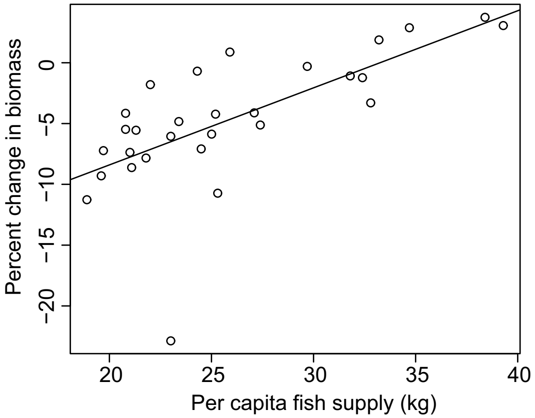

FIGURE 9.20 Scatter plot of percent change in biomass against fish supply with least-squares line imposed.

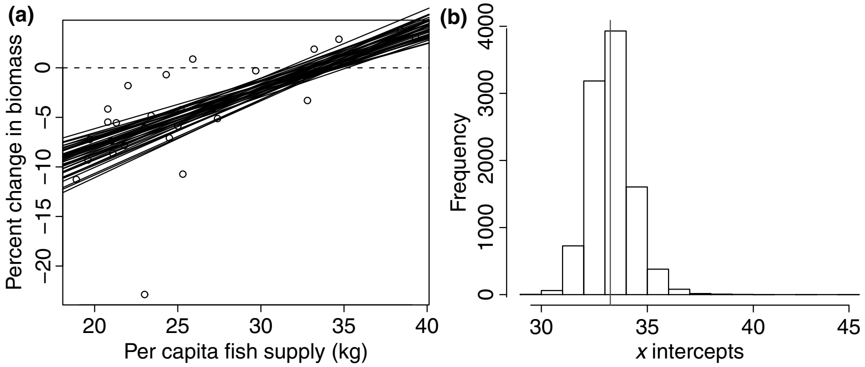

FIGURE 9.21 (a) Regression lines from 40 bootstrap samples of the bushmeat data. (b) Bootstrap distribution of the x intercept.

However, these are estimates based on a limited amount of data, so we turn to the bootstrap for more accuracy.

Figure 9.21 shows two views of the bootstrap output. Figure 9.21a is a graphical bootstrap—for 40 bootstrap samples, we calculate the slopes and intercepts and draw the corresponding lines over the original data and original line. We note that there appears to be a moderate amount of variability in the regression slopes, but not to the extent that any slopes are negative. There is also variability in the height of the regression lines, especially as we move to the right or to the left. If we extrapolate to the left, all the way to zero fish, it would show the variability in the intercept $\hat{\alpha}$.

Figure 9.21a shows that the regression lines have the smallest variability in the middle; the farther one goes to either side, the less accurate the answers are. However, the smallest variance occurs not at $\bar{x} = 26.1$, as implied by Theorem 9.6, but farther right. This is because the assumption of constant variance is violated. Looking back at the original scatter plot in Figure 9.20, we see that the residuals are smaller on the right. This is probably not just random variation, for two reasons, related to the numerator and denominator of the definition of relative change. When fish are more plentiful, there is less demand for bushmeat and correspondingly less variability in that demand; at the extreme, if there is zero demand, the variance is zero. Second, the observations on the left typically occur in the later years when the denominator is smaller, hence, even constant variability in the numerator results in greater variability in the ratio.

The results suggest that increasing the fish supply would reduce bushmeat harvest. An important question is what level of fish would stop the loss of wildlife? Based on the original regression line, the value would be 33.25 (the x intercept of the line). We can use the bootstrap to get an idea how accurate that number is.

We will use more bootstrap samples to do this. We use only 40 samples for plotting the regression lines because otherwise the figure becomes a mass of black ink. But now, for better accuracy in estimating the intercept, we will use 10^4 samples.

Figure 9.21b displays the bootstrap distribution of the x intercept, that is, the estimated supply of fish needed to stop the loss of wildlife. The original value of 33.25 falls in the middle of this distribution. The middle 95% range is 31.78 and 35.04, giving a rough idea of the reliability of the estimate. We are 95% confident that the supply of fish needed to forestall loss of biomass lies in that interval. Curiously, the interval $(31.5, 35.43) = (33.25 - 1.75, 33.25 + 2.18)$ stretches farther to the right, which gives a pessimistic story—it takes a lot of fish to gain confidence on the positive side.

We must admit that the bootstrap we just did, sampling with replacement from the data, assumes that the original data are i.i.d. from a bivariate population. This assumption is violated, because the data occur over time; they are neither independent nor identically distributed. The 22% drop in biomass in one year, for example, occurs in the final year when the denominator is small, making a large change in either direction relatively easy. There are procedures intended for time series data, both bootstrap and formula based; these are more complicated and are beyond the scope of this book.

9.6 LOGISTIC REGRESSION

According to the Centers for Disease Control and Prevention, the leading cause of death for people under the age of 34 is motor vehicle-related injuries.[3] Since many of these accidents are due to impaired driving—driving while under the influence of alcohol or drugs—there is a lot of emphasis on educating young drivers on the dangers of combining drinking and driving. But is there evidence that in fatal accidents, drinking and age are linked?

The Fatal Analysis Reporting System (FARS) (`http://www.nhtsa.gov/FARS`) database contains data on all fatal traffic accidents in the United States, the District of Columbia and Puerto Rico, since 1975. FARS is maintained by the National Center for Statistics and Analysis, part of the National Highway Traffic Safety Administration. We investigate the relationship between the involvement of alcohol and age of the driver in a random sample of 100 driver fatalities in 2009 in Pennsylvania; the drivers were driving a car, SUV, or light pickup truck (vehicles such as motor homes, convertibles, or commercial vehicles are excluded) (Table 9.3). One variable is a binary variable coded 1 if alcohol was involved and 0 otherwise; another variable is age of the driver, in years.

Let Y_i denote the alcohol involvement variable and x_i the independent variable age, $i = 1, 2, \ldots, 100$. For these individuals, we assume that the Y_i's are independent Bernoulli random variables with $P(Y_i = 1) = p_i$.

We want to understand the relationship between $p_i = \mathrm{E}[Y_i]$ and x_i. In Section 9.4, we considered linear regression of the form $\mathrm{E}[Y_i] = p_i = \alpha + \beta x_i$. But if β is nonzero, for sufficiently large and small x this would give probabilities less than zero or greater

[3] `http://www.cdc.gov/Motorvehiclesafety/`

TABLE 9.3 Part of the Data on Driver Fatalities in Pennsylvania

ID	Alcohol	Age
1	0	86
2	0	38
3	0	40
4	0	20
5	1	27

than one. Furthermore, linear regression assumes the same variances for every observation, but for Bernoulli data $\mathrm{Var}[Y_i] = p_i(1 - p_i)$.

So linear regression is not appropriate for these data. In this chapter, we discuss a type of regression suitable for zero–one data, *logistic regression*. We begin by introducing odds.

Definition 9.5 Let p denote the probability of some event. The *odds* of the event is defined by $p/(1 - p)$. ||

For instance, if $p = 0.8$ is the probability of a soccer team winning its next game, then $0.8/(1 - 0.8) = 4$ is its odds of winning the next game: the odds of the team winning (to not winning) the next game is 4 to 1. If $p = 0.25$ is the probability of dying from a certain disease, then $0.25/0.75 = 0.33$: the odds of dying (to not dying) is 1 to 3.

Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ be a set of ordered pairs where $x_1, x_2, \ldots, x_n$ are fixed and $Y_1, Y_2, \ldots, Y_n$ are Bernoulli random variables with $P(Y_i = 1) = p_i$. In logistic regression, we model the logarithm of the odds as a linear function of x:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i, \quad i = 1, 2, \ldots, n. \tag{9.19}$$

Equivalently,

$$p_i = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}. \tag{9.20}$$

This gives an S-shaped relationship between x and $\mathrm{E}[Y]$ (see Figure 9.22).

The slope coefficient β describes how quickly the estimated probability increases; the maximum slope is $\beta/4$ (where $p = 0.5$). We can also interpret β by evaluating the odds at two different values, say x and $x + \Delta x$,

$$\frac{p_1}{1 - p_1} = e^{\alpha+\beta x},$$
$$\frac{p_2}{1 - p_2} = e^{\alpha+\beta(x+\Delta x)}.$$

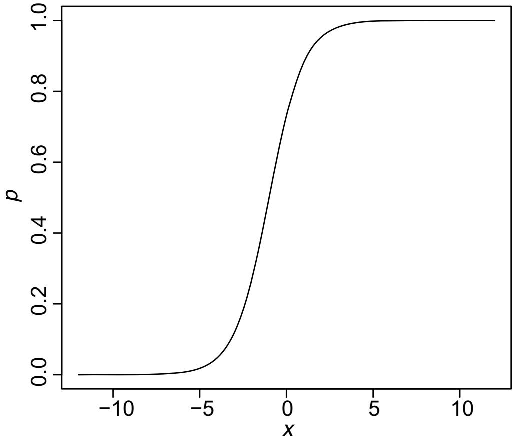

FIGURE 9.22 Plot of a typical logistic curve, Equation 9.20.

The *odds ratio* is

$$\frac{p_2/(1-p_2)}{p_1/(1-p_1)} = \frac{e^{\alpha+\beta(x+\Delta x)}}{e^{\alpha+\beta x}} = e^{\beta \Delta x}$$

and the *log odds ratio* is $\beta\Delta x$. So β measures how quickly the log odds ratio changes.

The parameters α and β are estimated using maximum likelihood. The likelihood function is given by

$$L(\alpha, \beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i},$$

and then taking the logarithm, we find

$$\begin{aligned}\ln(L(\alpha, \beta)) &= \sum_{i=1}^{n} Y_i \ln(p_i) + (1-Y_i)\ln(1-p_i) \\ &= \sum_{i=1}^{n} Y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i).\end{aligned}$$

Setting the partial derivatives with respect to α and β (using the chain rule, because p_i is a function of α and β) equal to 0 and simplifying yields equations:

$$\frac{\partial \ln(L)}{\partial \alpha} = \sum_{i=1}^{n} y_i - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \sum_{i=1}^{n} y_i - p_i = 0,$$

$$\frac{\partial \ln(L)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \frac{e^{\alpha+\beta x_i} x_i}{1+e^{\alpha+\beta x_i}} = \sum_{i=1}^{n} (y_i - p_i) x_i = 0.$$

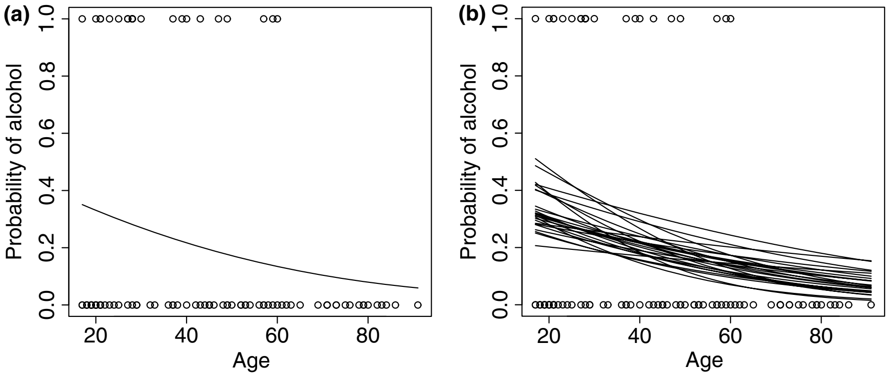

FIGURE 9.23 (a) Plot of estimated probability of alcohol being involved against age of driver. (b) Estimates from 25 bootstrap samples (see Section 9.6.1).

There is no closed form solution to these two equations, so a numerical algorithm must be used to find estimates $\hat{\alpha}$ and $\hat{\beta}$. For instance, R uses a procedure called iteratively reweighted least-squares, a multivariate version of Newton's method for finding the zero of a function; for those who have taken linear algebra, it uses the gradient and Hessian of $\ln L(\alpha, \beta)$.

Example 9.12 For the FARS fatalities data from Pennsylvania, the estimated logistic equation is

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.123 - 0.029x$$

(see Figure 9.23). For a 25-year-old driver, the estimated probability that alcohol was involved is

$$\hat{p} = \frac{e^{-0.123-0.029\times 25}}{1+e^{-0.123-0.029\times 25}} = 0.3$$

and the odds of alcohol being involved to not being involved is

$$\frac{\hat{p}}{1-\hat{p}} = e^{-0.123-0.029\times 25} = 0.428.$$

Similarly, we find the odds of alcohol being involved in the case of a 35-year-old driver, $\hat{p}_2/(1-\hat{p}_2) = 0.32$. The odds ratio is $0.428/0.32 = 1.34$. The odds of alcohol being involved in the fatal accident of a 25-year-old driver is 1.34 times greater than the odds of alcohol being involved in the fatal accident of a 35-year-old driver. Equivalently, we can say that the odds of alcohol being involved in the fatal accident of a 25-year-old driver is 34% higher than the odds of alcohol being involved in the fatal accident of a 35-year-old driver.

R Note:

The FARS data are in a file called `Fatalities.csv`. In `R`, logistic regression is performed using the `glm` command. The syntax is similar to that of the `lm` command, except that we must specify that the Y variable follows a binomial (Bernoulli) distribution:

```
> glm(Alcohol ~ Age, data = Fatalities, family = binomial)
...
Coefficients:
(Intercept)          Age
   -0.12262     -0.02898

x <- seq(17, 91, length=500)      # vector spanning the age range
y <- exp(-.123-.029*x) / (1+exp(-.123-.029*x))

plot(x, y, type = "l", ylim = c(0,1), xlab = "age",
 ylab = "Probability of alcohol")
points(Fatalities$Age, Fatalities$Alcohol)  # observations
```

□

Example 9.13 Suppose a hospital conducts a study to see if there is a link between patients getting an infection ($y = 1$ if yes) and their length of stay in the hospital (x, in days). A logistic regression performed on their data gives

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.942 + 0.023x.$$

How do the odds of getting an infection change for somebody who stays an additional week in this hospital?

Solution To compare the odds of infection for somebody who stays $\Delta x = 7$ days longer than another patient, compute $e^{0.023 \cdot 7} = 1.175$. Thus, staying an additional week increases the odds of infection by about 17.5%. Alternatively, we can express this result: the odds of infection are 1.175 times greater for every extra 7 days in the hospital. □

Logistic regression is a special case of a *generalized linear model*; another common version is *Poisson regression* in which Y has a Poisson distribution with mean given by $\exp(\alpha + \beta x)$; see Collett (2003), Dobson (2002), Kutner et al. (2005), McCullagh and Nelder (1989), for details. One of us (Hesterberg) consults at Google on a project to predict web traffic for billions of search phrases based on time of day and day of week, using Poisson regression. Searches that receive more traffic than predicted may be flagged for *Google Trends*. For example, a current hot search on November 21, 2010 was "ben roethlisberger punched."

9.6.1 Inference for Logistic Regression

Standard errors in logistic regression are based on assumptions similar to those in Section 9.4; we assume the following:

- x values are fixed, not random.
- Relationship between the x values and $\log(p/(1-p))$ is linear.
- Y values have Bernoulli distributions, with parameters p_i.
- Y values are independent.

We will rely on software to do the calculations for standard errors for coefficients and predictions ($\hat{p}_i$).

We can use the standard errors to produce confidence intervals, t statistics, P-values, and hypothesis tests, but be warned that sample sizes may need to be quite large for these to be accurate. Some software calculates t statistics, but then admirably declines to print P-values based on these t statistics because the P-values cannot be trusted.

Alternatively, we may bootstrap for standard errors and confidence intervals. We resample individuals, that is resample paired values (age, alcohol). For each bootstrap data set, we estimate the logistic regression parameters and calculate the desired predictions. Figure 9.23b shows the predictions from 25 bootstrap samples. This gives a rough idea of variability and suggests that some distributions are skewed—for example, predictions for the probability of alcohol involvement at age 80 are mostly near zero, but with a few larger values.

Figure 9.24 shows a histogram and normal quantile plot for b for 1000 bootstrap samples, and Figure 9.25 shows a histogram and normal quantile plot for the probabilities of alcohol involvement at age 20.

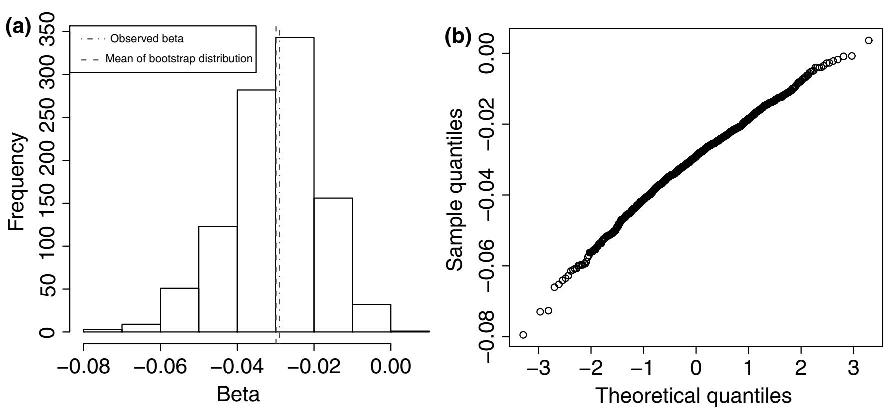

FIGURE 9.24 Histogram and normal quantile plot for b, the logistic regression slope coefficient for alcohol involvement versus age of driver.

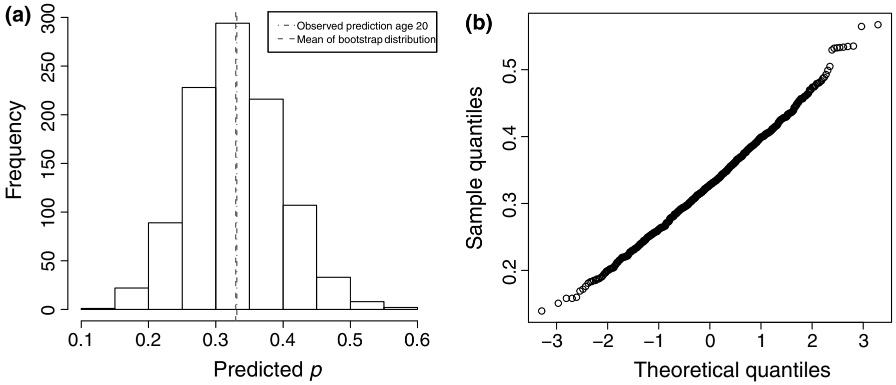

FIGURE 9.25 Histogram and normal quantile plot for probability of involvement at age 20.

Bootstrap percentile confidence intervals are $(-0.056, -0.008)$ for β and $(0.20, 0.47)$ for the prediction of alcohol involvement at age 20. These are quite wide, a range of 20–47% for the probability, and a ratio of 7:1 for the slope. For comparison, we note that intervals based on the formula standard errors are similar: $(-0.056, -0.002)$ for β and $(0.187, 0.475)$ for the probability.

R Note:

The command to extract coefficients from a `glm` object is `coef`. The command `plogis` is the cdf for a logistic random variable and is a handy way to compute $\exp(x)/(1 + \exp(x))$.

```
fit <- glm(Alcohol ~ Age, data = Fatalities, family = binomial)
data.class(fit)  # is a "glm" object, so for help use:
help(glm)

fit          # prints the coefficients and other basic info
coef(fit)    # the coefficients as a vector
summary(fit) # gives standard errors for coefficients, etc.

x <- seq(17, 91, length=500) # vector spanning the age range
# compute predicted probabilities
y1 <- exp(-.123-.029*x) / (1+exp(-.123-.029*x))
y2 <- plogis(coef(fit)[1] + coef(fit)[2] * x)

plot(Fatalities$Age, Fatalities$Alcohol,
     ylab = "Probability of alcohol")
lines(x, y2)

# Full bootstrap - slope coefficient, and prediction at age 20
N <- 10^3
```

```
n <- nrow(Fatalities)                   # number of observations
alpha.boot <- numeric(N)
beta.boot <- numeric(N)
pPred.boot <- numeric(N)

for (i in 1:N)
{
  index <- sample(n, replace = TRUE)
  Fatal.boot <- Fatalities[index, ]     # resampled data

  fit.boot <- glm(Alcohol ~ Age, data = Fatal.boot,
                  family = binomial)
  alpha.boot[i] <- coef(fit.boot)[1]    # new intercept
  beta.boot[i] <- coef(fit.boot)[2]     # new slope
  pPred.boot[i] <- plogis(alpha.boot[i] + 20 * beta.boot[i])
}

quantile(beta.boot, c(.025, .975))      # 95% percentile intervals
quantile(pPred.boot, c(.025, .975))

par(mfrow=c(2,2))                       # set layout
hist(beta.boot, xlab = "beta", main = "")
qqnorm(beta.boot, main = "")

hist(pPred.boot, xlab = "p^", main = "")
qqnorm(pPred.boot, main = "")
# Figures in the text also use abline to add lines, and legend.
par(mfrow=c(1,1))                       # reset layout
```

The `predict` command can also be used to give predicted values. The required arguments of `predict` are the model object (the output from the `glm` command) and a data frame containing the values of the explanatory variable at which you wish to predict. By default, `predict` returns the predicted $a + bx$. To obtain the predicted probabilities, provide the argument `type = "response"`. We illustrate the use of `predict` in reproducing Figure 9.23:

```
help(predict.glmm)                # for more help on predict

n <- nrow(Fatalities)             # number of observations
x <- seq(17, 91, length=500)      # vector spanning the age range
df.Age <- data.frame(Age = x)     # data frame to hold
    # explanatory variables, will use this for making predictions

plot(Fatalities$Age, Fatalities$Alcohol,
     ylab = "Probability of alcohol")
for (i in 1:25)
 {
  index <- sample(n, replace = TRUE)
  Fatal.boot <- Fatalities[index, ]     # resampled data
  fit.boot <- glm(Alcohol ~ Age, data = Fatal.boot,
                  family = binomial)
  pPred <- predict(fit.boot, newdata = df.Age, type = "response")
  lines(x, pPred)
}
```

9.7 EXERCISES

1. Let X and Y be random variables with joint probability density function given by

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2), & 0 \le x \le 1,\ 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

   Find the covariance $\text{Cov}[X, Y]$.

2. Let X be a random variable with mean μ and variance σ^2. Let $Y = C$, a constant. Find the covariance between X and Y.

3. Let X and Y be random variables with $\text{Var}[X] = 4$, $\text{Var}[Y] = 9$, and $\text{Cov}[X, Y] = 3$. Find the variance of $2X + 3Y$.

4. Let X and Y be random variables with $\text{Var}[X] = 5$, $\text{Var}[Y] = 7$, and $\text{Cov}[X, Y] = 2$. Find the variance of $2X - 5Y$.

5. Let X, Y, and Z be random variables with $\text{Var}[X] = 3$, $\text{Var}[Y] = 2$, and $\text{Var}[Z] = 3$ and $\text{Cov}[X, Y] = -2$, $\text{Cov}[X, Z, =] - 4$, and $\text{Cov}[Y, Z] = 7$. Find $\text{Var}[5X - Y + 2Z]$.

6. Import the data from data set `corrExerciseA`.
   (a) Create a scatter plot of X and Y and then find the correlation between X and Y.
   (b) Note that there is another variable Z that puts each observation into group A or B. Create a scatter plot of A points and then of B points. Describe the relationship between X and Y for each group.
   (c) Find the correlation between X and Y for each group.
   (d) What is the lesson here?

7. Import the data from data set `corrExerciseB`.
   (a) Find the correlation between X and Y.
   (b) Note that observations are put into one of four groups: A, B, C, and D. Find the mean of X and the mean of Y for each group.
   (c) Create a scatter plot of the means $(\bar{X}_{\text{A}}, \bar{Y}_{\text{A}})$, $(\bar{X}_{\text{B}}, \bar{Y}_{\text{B}})$, ... and then find the correlation between the $\bar{X}'$s, and the $\bar{Y}'$s. Compare to (a).

   Ecological Correlation Correlations based on rates or groups are often higher than correlations based on individuals. This is a common problem in the social/behavior sciences where many data sets are based on summaries (e.g., census data for the 50 states: mean income levels, mean literacy rate, etc.).

8. Compare the roundoff error of two ways of computing sample variances. Write functions that compute `(mean(x^2)- mean(x)^2) * n/(n-1)` and `mean((x-mean(x))^2) * n/(n-1)` and calculate the variance of $x_1 = c, x_2 = c + 1, x_3 = c + 2$ for $c = 0$, $c = 10^5$, $c = 10^6$, ... using both. What do you find? Compare to the `R` `var` function.

9. Verify that $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$ — that is, the average of the residuals sum to 0. *Hint*: Use the fact that a and b are solutions to $\partial g/\partial a = 0$, where $g(a, b) = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$.

10. Let $s_{\hat{y}}$ denote the standard deviation of the predicted y's; that is, the standard deviation of $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$.
    (a) Verify that $s_{\hat{y}} = rs_y$ (Equation 9.13).
    (b) Let s_e denote the standard deviation of the residuals, $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$. Show that

$$s_e = \sqrt{1 - r^2}s_y.$$

11. Suppose the height and weight of 30 girls in Sodor are measured. The mean and standard deviation of the heights are 46 and 7 in., respectively, and the mean and standard deviation of the weights are 94 and 15 pounds, respectively. Suppose the correlation between height and weight is 0.75.
    (a) Find the equation of the least-squares regression line of weight against height.
    (b) Find the predicted weight of a girl who is 5 ft. tall.
    (c) Find R-squared and give a sentence interpreting this statistic.

12. Refer to the Beer and Hot Wings Case Study in Section 1.8.
    (a) Create a scatter plot of beer consumed against hot wings eaten and find the correlation between these two variables.
    (b) Find the equation of the least-squares regression line (take `hotwings` as the independent variable) and give a sentence interpreting the slope.
    (c) Compute R-squared and state the interpretation of this statistic.

13. Figure 9.26 contains residuals plot for several least-squares regression models. Describe the unusual features.

14. Is there a relationship between female literacy and birth rate? The data set `Illiteracy` contains data on a sample of countries where female illiteracy is more that 5%. The variable `Illit` is the percentage of women over 15 years of age who are illiterate (2003) and the variable `Births` is the number of births per woman in that country (2005).[4]

[4] http://www.unesco.org, www.data.worldbank.org.

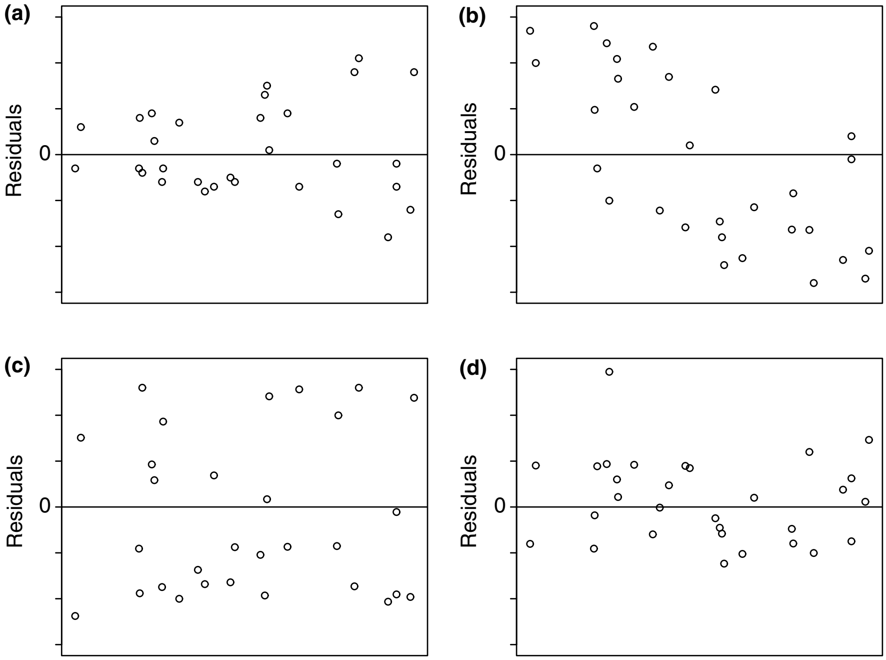

FIGURE 9.26 Residuals plot for regression models.

(a) Create a scatter plot of birth rate against illiteracy and comment on the relationship.

(b) Find the equation of the least-squares line and interpret the slope and r^2.

(c) Create a residuals plot and comment on the appropriateness of a straight line model.

(d) Can we say that improving literacy (reducing illiteracy) will cause the birth rate to go down? Explain.

15. The data set `Volleyball2009` contains data on 30 Division I women volleyball teams from the 2009 season (from `http://web1.ncaa.org/stats/StatsSrv/rankings`).[5]

(a) Create a scatter plot of the number of kills per set (`Kills`) against assists per set (`Assts`) and describe the relationship.

(b) Find the least-squares equation for the line and interpret the slope and r^2.

(c) Create the residuals plot and comment on the appropriateness of a straight line model.

16. Refer to Exercise 15.

(a) Find a 95% confidence interval for the slope.

(b) Suppose a team not listed records 12.4 assists per set. Find the predicted number of kills per set and a 95% prediction interval.

[5] © 2009 American Volleyball Coaches Association; © 2009 National Collegiate Athletic Association.

17. Refer to the North Carolina Births Case Study in Section 1.2.
    (a) Create a scatter plot of weight against gestation period and compute the correlation between the two variables.
    (b) Find the least-squares regression line.
    (c) Give a sentence with the interpretation of the slope and the R-squared.
    (d) Create the residuals plot and comment on any unusual features. Is a linear model appropriate for the relationship between gestation period and weight?
18. North Carolina Births (continued)
    (a) What is the estimate of σ for the linear model of weight against gestation period?
    (b) Find a 95% confidence interval for the true slope β.
19. In the ice cream example (Example 9.6), the sum of squares for the sugar variable (ss_x) is 324.446 and the residual standard error is 2.069. Find a 95% confidence interval for the true slope (use the model including all the data).
20. Suppose a company offers tutoring to students interested in improving their math SAT scores. From a random sample of 100 previous customers, they find that $\widehat{\texttt{Score}} = 502.7 + 1.45 \cdot \texttt{Hours}$, where `Score` is the SAT math score and `Hours` is the number of hours the student was tutored. The R-squared and residual standard error for this model is 0.855 and 16.54, respectively.
    (a) For every additional 10 h of tutoring, what is the corresponding change in the test score?
    (b) If the standard deviation of the hours variable is 27.5, what is the standard deviation of the score variable?
    (c) Find a 95% confidence interval for the true slope.
    (d) Find a 95% confidence interval for the mean score for students who are tutored 50 h. Assume the mean number of hours tutored is 55 h.
21. The Mauna Loa Observatory, located on the Hawaiian islands of Mauna Loa, specializes in research in the atmospheric sciences. The facility has been collecting data on carbon dioxide levels since 1950s. The data file `Maunaloa` contains data on average CO_2 levels (ppm) for the month of May from 1990 to 2010 (from `www.esrl.noaa.gov/gmd/ccgg/trends/`).
    (a) Create a scatter plot of CO_2 levels against year and describe the relationship.
    (b) Find the equation for the least-squares equation.
    (c) Plot the residuals against year. Is a straight line model appropriate? Discuss.
22. The data set `Walleye` from the Minnesota Pollution Control Agency contains data on length (inches) and weight (pound) measurements for a sample of 60 walleye caught in Minnesota lakes during 1990s (Monson (2010)).
    (a) Create a scatter plot of weight against length. Does the relationship appear linear?

In fact, biologists have determined that the relationship between length and weight of fish is given by $W = aL^b$, where a and b depend on the species (Ricker (1973, 1975)). We will consider $\log(W) = \log(a) + b\ \log(L)$ (base 10).

(b) Transform the weight and length variables by log base 10 and then create a scatter plot of log(`Weight`) against log(`Length`). Describe the relationship.

(c) Use least-squares to find estimates of a and b based on this sample.

(d) What is the 95% confidence interval for b?

23. Import from the data set `Alelager` the data on alcohol content (per volume) and calories for a sample of beers (12 ounces). Find the correlation between alcohol and calories and then compute a 95% bootstrap percentile confidence interval for the true correlation.

24. Using the data set `Illiteracy`,

    (a) find the correlation between illiteracy rates and births via the bootstrap and also find a 95% bootstrap percentile interval; and

    (b) using a permutation test, find if illiteracy rates and births are independent.

25. Prove Proposition 9.2

26. Prove the second part of Proposition 9.4.

27. Prove Proposition 9.5.

28. Prove the results about $\bar{Y}$ in Theorem 9.6.

29. In Theorem 9.3, we stated without proof that $\hat{\beta}$ and $\bar{Y}$ are independent. Instead, prove that they are uncorrelated.

30. A campaign manager conducts a survey to gauge voter support for his candidate Lopez. He gathers data on the age of a registered voter (x) and whether this person supports Lopez ($Y = 1$) or somebody else ($Y = 0$). An analysis yields the following logistic equation:

$$\ln\left(\hat{p}/\left(1 - \hat{p}\right)\right) = -0.324 + 0.012x,$$

where p is the probability of a vote for Lopez.

    (a) Find the estimated probability that a 21-year-old voter will vote for Lopez.

    (b) Compare the odds of support for Lopez between two people who are 10 years apart in age and give your answer in a complete sentence.

    (c) At what age is the expected response equal to 0.5?

31. On January 28, 1986, the space shuttle Challenger exploded during lift-off, killing all seven astronauts aboard.[6] In the follow-up investigation, attention was focused on the rubber O-rings that sealed the booster rockets. Engineers

[6] `http://history.nasa.gov/sts51l.html`.

had concerns earlier that the ambient temperature at the time of lift-off could affect the integrity of the O-ring. The data set `Challenger` contains data on 23 Challenger flights before the January 21 flight. The binary variable `Incident` records 1 if one of the O-rings on one of the booster rockets was damaged on this flight. The variable `Temperature` records the temperature (Fahrenheit) at the time of lift-off.

(a) Find the logistic regression equation modeling the log-odds of an O-ring incident against temperature. Plot the graph of the predicted probabilities against temperature and add the observed incidents also.

(b) How does a 10°F degrees decrease in temperature affect the odds of an incident? State your answer in a complete sentence.

(c) On the day of the Challenger accident, the temperature was 33°F. What is the predicted probability of an O-ring incident?

(d) Some would argue that it is not appropriate to use this model to predict an O-ring incident at 33°F. Why not?

32. The biologist in the Black Spruce Seedings Case Study in Section 1.9 was also interested in the relationship between seedling growth and water table depth. The data set `Watertable` contains data on a sample of the seedlings. The variable binary `Alive` indicates 1 if the seedling was alive at the end of the second year of the study and 0 otherwise. The variable `Depth` gives the depth of the water table (centimeter).

(a) Find the logistic equation modeling the log-odds of a seedling being alive against water table depth. Plot the graph of the predicted probabilities against depth and add the observed data points.

(b) Interpret the slope of the regression equation (in terms of odds).

(c) For a seedling growing in soil with a water table depth of 15 cm, find the predicted probability of being alive at the end of the second year.

33. Import from the data set `Phillies2009` the data on the Philadelphia Phillies baseball team.

(a) Find the logistic equation modeling the log-odds of the team winning (`Outcome`) against the number of hits in a game (`Hits`). (`R` will automatically convert `Lose` to 0 and `Win` to 1).

(b) Interpret the slope of the equation (in terms of odds).

(c) Find a 95% bootstrap percentile interval for the slope.

(d) Predict the probability of winning if the team has 17 hits and then find a 95% bootstrap percentile interval.

(e) For inference, we need to assume that the observations are independent. Is that condition met here?

34. Import from the data set `Titanic` the data on male passengers on the Titanic. The variable `Survived` is 1 if the passenger survived the sinking and 0 if the passenger died.

(a) Find the logistic equation modeling the log-odds of a male passenger surviving against age.

(b) Compare the odds of survival for a 30-year-old male with a 40 year old male.

(c) Find a 95% bootstrap percentile interval for the slope.

(d) Estimate the probability of a 69-year-old male surviving, and find a 95% bootstrap percentile interval for the probability.

35. Verify that Equation 9.20 is equivalent to $p_i = 1/\left(1 + e^{-(\alpha+\beta x)}\right)$.