3

Covariance, Regression, and Correlation

In the previous chapter, the variance was introduced as a measure of the dispersion of a univariate distribution. Additional statistics are required to describe the joint distribution of two or more variables. The **covariance** provides a natural measure of the association between two variables, and it appears in the analysis of many problems in quantitative genetics including the resemblance between relatives, the correlation between characters, and measures of selection.

As a prelude to the formal theory of covariance and regression, we first provide a brief review of the theory for the distribution of pairs of random variables. We then give a formal definition of the covariance and its properties. Next, we show how the covariance enters naturally into statistical methods for estimating the linear relationship between two variables (least-squares linear regression) and for estimating the goodness-of-fit of such linear trends (correlation). Finally, we apply the concept of covariance to several problems in quantitative-genetic theory. More advanced topics associated with multivariate distributions involving three or more variables are taken up in Chapter 8.

JOINTLY DISTRIBUTED RANDOM VARIABLES

The probability of joint occurrence of a pair of random variables (x, y) is specified by the **joint probability density function**, $p(x, y)$, where

$$\mathrm{P}(\, y_1 \le y \le y_2,\, x_1 \le x \le x_2 \,) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} p(x, y)\, dx\, dy \tag{3.1}$$

We often ask questions of the form: What is the distribution of y given that x equals some specified value? For example, we might want to know the probability that parents whose height is 68 inches have offspring with height exceeding 70 inches. To answer such questions, we use $p(y|x)$, the **conditional density** of y given x, where

$$\mathrm{P}(\, y_1 \le y \le y_2 \,|\, x \,) = \int_{y_1}^{y_2} p(\, y \,|\, x \,)\, dy \tag{3.2}$$

Joint probability density functions, $p(x, y)$, and conditional density functions,

$p(y|x)$, are connected by

$$p(x,y) = p(\,y\,|\,x\,)\,p(x) \tag{3.3a}$$

where $p(x) = \int_{-\infty}^{+\infty} p(\,y\,|\,x\,)\,dy$ is the marginal (univariate) density of x.

Two random variables, x and y, are said to be **independent** if $p(x,y)$ can be factored into the product of a function of x only and a function of y only, i.e.,

$$p(x,y) = p(x)\,p(y) \tag{3.3b}$$

If x and y are independent, knowledge of x gives no information about the value of y. From Equations 3.3a and 3.3b, if $p(x,y) = p(x)\,p(y)$, then $p(\,y\,|\,x\,) = p(y)$.

Expectations of Jointly Distributed Variables

The expectation of a bivariate function, $f(x,y)$, is determined by the joint probability density

$$E[\,f(x,y)\,] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,p(x,y)\,dx\,dy \tag{3.4}$$

Most of this chapter is focused on **conditional expectation,** i.e., the expectation of one variable, given information on another. For example, one may know the value of x (perhaps parental height), and wish to compute the expected value of y (offspring height) given x. In general, conditional expectations are computed by using the conditional density

$$E(\,y\,|\,x\,) = \int_{-\infty}^{+\infty} y\,p(\,y\,|\,x\,)\,dy \tag{3.5}$$

If x and y are independent, then $E(y|x) = E(y)$, the unconditional expectation. Otherwise, $E(\,y\,|\,x\,)$ is a function of the specified x value. For height in humans (Figure 1.1), Galton (1889) observed a linear relationship,

$$E(\,y\,|\,x\,) = \alpha + \beta x \tag{3.6}$$

where α and β are constants. Thus, the conditional expectation of height in offspring (y) is linearly related to the average height of the parents (x).

COVARIANCE

Consider a set of paired variables, (x,y). For each pair, subtract the population mean μ_x from the measure of x, and similarly subtract μ_y from y. Finally, for each pair of observations, multiply both of these new measures together to obtain $(x-\mu_x)(y-\mu_y)$. The **covariance** of x and y is defined to be the average of this quantity over all pairs of measures in the population,

$$\sigma(x,y) = E[\,(x-\mu_x)\,(y-\mu_y)\,] \tag{3.7}$$

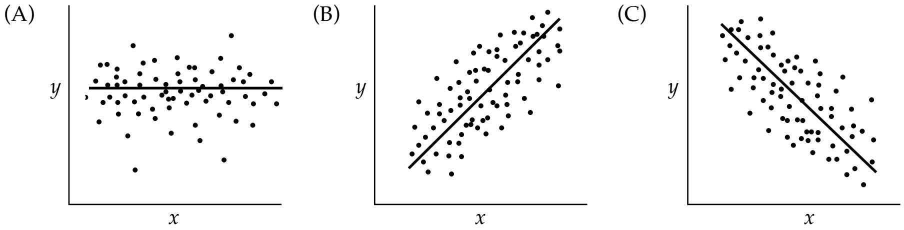

Figure 3.1 Scatterplots for the variables x and y. Each point in the x-y plane corresponds to a single pair of observations (x, y). The line drawn through the scatterplot gives the expected value of y given a specified value of x. (A) There is no linear tendency for large x values to be associated with large (or small) y values, so $\sigma(x, y) = 0$. (B) As x increases, the conditional expectation of y given x, $E(y|x)$, also increases, and $\sigma(x, y) > 0$. (C) As x increases, the conditional expectation of y given x decreases, and $\sigma(x, y) < 0$.

We often denote covariance by $\sigma_{x,y}$. Because $E(x) = \mu_x$ and $E(y) = \mu_y$, expansion of the product leads to further simplification,

$$\begin{aligned}\sigma(x, y) &= E[\,(x - \mu_x)\,(y - \mu_y)\,]\\ &= E\,(xy - \mu_y\, x - \mu_x\, y + \mu_x\, \mu_y\,)\\ &= E(x\,y) - \mu_y\, E(x) - \mu_x\, E(y) + \mu_x\, \mu_y\\ &= E(x\,y) - \mu_x\, \mu_y \end{aligned} \tag{3.8}$$

In words, the covariance is the mean of the pairwise cross-product $x\,y$ minus the cross-product of the means. The sampling estimator of $\sigma(x, y)$ is similar in form to that for a variance,

$$\mathrm{Cov}(x, y) = \frac{n\,(\,\overline{xy} - \overline{x} \cdot \overline{y}\,)}{n - 1} \tag{3.9}$$

where n is the number of pairs of observations, and

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i\, y_i$$

The covariance is a measure of association between x and y (Figure 3.1). It is positive if y increases with increasing x, negative if y decreases as x increases, and zero if there is no *linear* tendency for y to change with x. If x and y are independent, then $\sigma(x, y) = 0$, but the converse is not true — a covariance of zero does not necessarily imply independence. (We will return to this shortly; see Figure 3.3.)

Useful Identities for Variances and Covariances

Since $\sigma(x, y) = \sigma(y, x)$, covariances are symmetrical. Furthermore, from the definition of the variance and covariance,

$$\sigma(x, x) = \sigma^2(x) \tag{3.10a}$$

i.e., *the covariance of a variable with itself is the variance of that variable*. It also follows from Equation 3.8 that, for any constant a,

$$\sigma(a, x) = 0 \tag{3.10b}$$

$$\sigma(a\,x, y) = a\,\sigma(x, y) \tag{3.10c}$$

and if b is also a constant

$$\sigma(a\,x, b\,y) = a\,b\,\sigma(x, y) \tag{3.10d}$$

From Equations 3.10a and 3.10d,

$$\sigma^2(a\,x) = a^2\sigma^2(x) \tag{3.10e}$$

i.e., *the variance of the transformed variable ax is a^2 times the variance of x*. Likewise, for any constant a,

$$\sigma[\,(a + x), y\,] = \sigma(x, y) \tag{3.10f}$$

so that *simply adding a constant to a variable does not change its covariance with another variable*.

Finally, the covariance of two sums can be written as a sum of covariances,

$$\sigma[\,(x + y), (w + z)\,] = \sigma(x, w) + \sigma(y, w) + \sigma(x, z) + \sigma(y, z) \tag{3.10g}$$

Similarly, the variance of a sum can be expressed as the sum of all possible variances and covariances. From Equations 3.10a and 3.10g,

$$\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y) + 2\sigma(x, y) \tag{3.11a}$$

More generally,

$$\sigma^2\left(\sum_i^n x_i\right) = \sum_i^n\sum_j^n \sigma(x_i, x_j) = \sum_i^n \sigma^2(x_i) + 2\sum_{i<j}^n \sigma(x_i, x_j) \tag{3.11b}$$

Thus, *the variance of a sum of uncorrelated variables is just the sum of the variances of each variable.*

We will make considerable use of the preceding relationships in the remainder of this chapter and in chapters to come. Methods for approximating variances and covariances of more complex functions are outlined in Appendix 1.

REGRESSION

Depending on the causal connections between two variables, x and y, their true relationship may be linear or nonlinear. However, regardless of the true pattern of association, a linear model can always serve as a first approximation. In this case, the analysis is particularly simple,

$$y = \alpha + \beta x + e \tag{3.12a}$$

where α is the y-intercept, β is the slope of the line (also known as the **regression coefficient**), and e is the **residual error**. Letting

$$\widehat{y} = \alpha + \beta x \tag{3.12b}$$

be the value of y predicted by the model, then the residual error is the deviation between the observed and predicted y value, i.e., $e = y - \widehat{y}$. When information on x is used to predict y, x is referred to as the **predictor** or **independent variable** and y as the **response** or **dependent variable**.

The objective of linear regression analysis is to estimate the model parameters, α and β, that give the "best fit" for the joint distribution of x and y. The true parameters α and β are only obtainable if the entire population is sampled. With an incomplete sample, α and β are approximated by sample estimators, denoted as a and b. Good approximations of α and β are sometimes obtainable by visual inspection of the data, particularly in the physical sciences, where deviations from a simple relationship are due to errors of measurement rather than biological variability. However, in biology many factors are often beyond the investigator's control. The data in Figure 3.2 provide a good example. While there appears to be a weak positive relationship between maternal weight and offspring number in rats, it is difficult to say anything more precise. An objective definition of "best fit" is required.

Derivation of the Least-Squares Linear Regression

The mathematical method of **least-squares linear regression** provides one such best-fit solution. Without making any assumptions about the true joint distribution of x and y, least-squares regression minimizes the average value of the squared (vertical) deviations of the observed y from the values predicted by the regression line. That is, the **least-squares** solution yields the values of a and b that minimize the mean squared residual, $\overline{e^2}$. Other criteria could be used to define "best fit." For example, one might minimize the mean absolute deviations (or cubed deviations) of observed values from predicted values. However, as we will now see, least-squares regression has the unique and very useful property of maximizing the amount of variance in y that can be explained by a linear model.

Consider a sample of n individuals, each of which has been measured for x and y. Recalling the definition of a residual

$$e = y - \widehat{y} = y - a - bx \tag{3.13a}$$

and then adding and subtracting the quantity $(\,\overline{y} + b\,\overline{x}\,)$ on the right side, we obtain

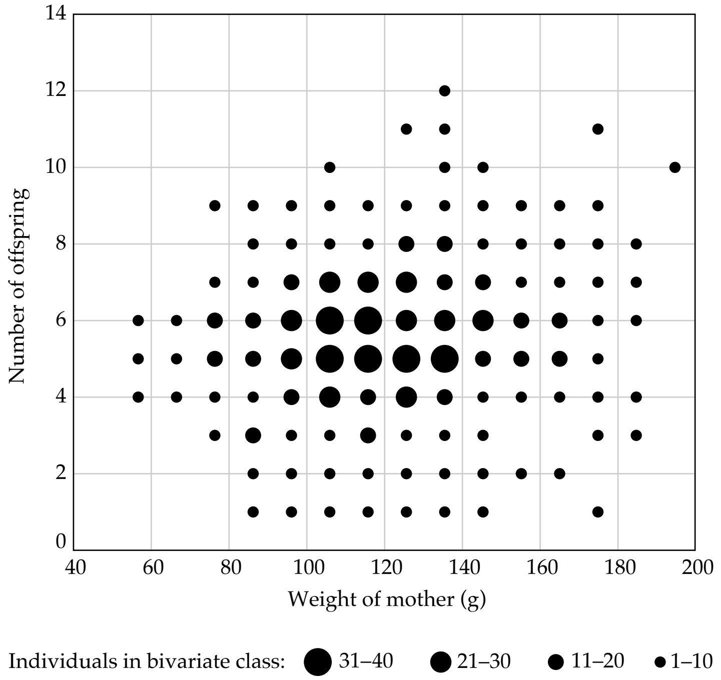

Figure 3.2 A bivariate plot of the relationship between maternal weight and number of offspring for the sample of rats summarized in Table 2.2. Different-sized circles refer to different numbers of individuals in the bivariate classes.

$$e = (\,y - \overline{y}\,) - b\,(\,x - \overline{x}\,) - (\,a + b\,\overline{x} - \overline{y}\,) \tag{3.13b}$$

Squaring both sides leads to

$$\begin{aligned} e^2 = {} & (\,y - \overline{y}\,)^2 - 2\,b\,(\,y - \overline{y}\,)\,(\,x - \overline{x}\,) + b^2(\,x - \overline{x}\,)^2 + (\,a + b\,\overline{x} - \overline{y}\,)^2 \\ & -\; 2\,(\,y - \overline{y}\,)\,(a + b\,\overline{x} - \overline{y}\,) + 2\,b\,(\,x - \overline{x}\,)(\,a + b\,\overline{x} - \overline{y}\,) \end{aligned} \tag{3.13c}$$

Finally, we consider the average value of e^2 in the sample. The final two terms in Equation 3.13b drop out here because, by definition, the mean values of $(x - \overline{x})$ and $(y - \overline{y})$ are zero. However, by definition, the mean values of the first three terms are directly related to the sample variances and covariance. Thus,

$$\overline{e^2} = \left(\frac{n-1}{n}\right)\big[\,\mathrm{Var}(y) - 2\,b\,\mathrm{Cov}(x,y) + b^2\,\mathrm{Var}(x)\,\big] + (\,a + b\,\overline{x} - \overline{y}\,)^2 \tag{3.13d}$$

The values of a and b that minimize $\overline{e^2}$ are obtained by taking partial derivatives of this function and setting them equal to zero:

$$\frac{\partial\,(\overline{e^2})}{\partial\,a} = 2\,(a + b\,\overline{x} - \overline{y}\,) = 0$$

$$\frac{\partial\,(\overline{e^2})}{\partial\,b} = 2\left[\left(\frac{n-1}{n}\right)[-\text{Cov}(x,y) + b\,\text{Var}(x)\,] + \overline{x}\,(\,a + b\,\overline{x} - \overline{y}\,)\right] = 0$$

The solutions to these two equations are

$$a = \overline{y} - b\,\overline{x} \tag{3.14a}$$

$$b = \frac{\text{Cov}(x,y)}{\text{Var}(x)} \tag{3.14b}$$

Thus, the least-squares estimators for the intercept and slope of a linear regression are simple functions of the observed means, variances, and covariances. From the standpoint of quantitative genetics, this property is exceedingly useful, since such statistics are readily obtainable from phenotypic data.

Properties of Least-squares Regressions

Here we summarize some fundamental features and useful properties of the least-squares approach to linear regression analysis:

1. *The regression line passes through the means of both x and y.* This relationship should be immediately apparent from Equation 3.14a, which implies $\overline{y} = a + b\,\overline{x}$.

2. *The average value of the residual is zero.* From Equation 3.13a, the mean residual is $\overline{e} = \overline{y} - a - b\,\overline{x}$, which is constrained to be zero by Equation 3.14a. Thus, the least-squares procedure results in a fit to the data such that the sum of (vertical) deviations above and below the regression line are exactly equal.

3. *For any set of paired data, the least-squares regression parameters, a and b, define the straight line that maximizes the amount of variation in y that can be explained by a linear regression on x.* Since $\overline{e} = 0$, it follows that the variance of residual errors about the regression is simply $\overline{e^2}$. As noted above, this variance is the quantity minimized by the least-squares procedure.

4. *The residual errors around the least-squares regression are uncorrelated with the predictor variable x.* This statement follows since

$$\begin{aligned}\text{Cov}(x,e) &= \text{Cov}[\,x,(y - a - b\,x)\,] = \text{Cov}(x,y) - \text{Cov}(x,a) - b\,\text{Cov}(x,x)\\ &= \text{Cov}(x,y) - 0 - b\,\text{Var}(x)\\ &= \text{Cov}(x,y) - \frac{\text{Cov}(x,y)}{\text{Var}(x)}\,\text{Var}(x) = 0\end{aligned}$$

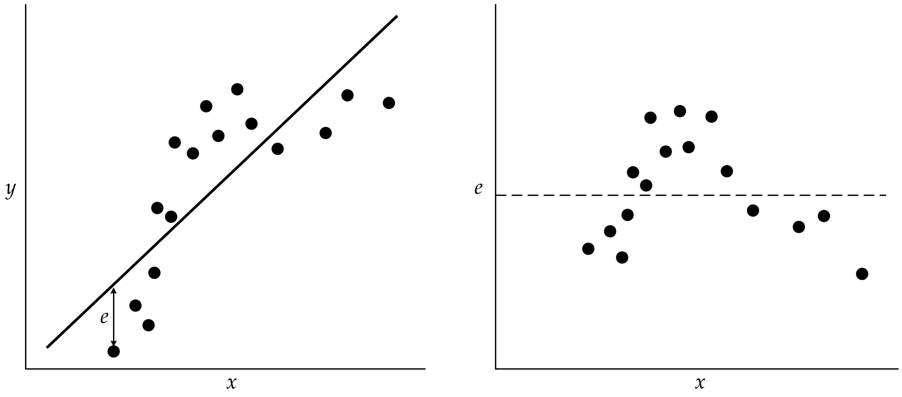

Figure 3.3 A linear least-squares fit to an inherently nonlinear data set. Although there is a systematic relationship between the residual error (e) and the predictor variable (x), the two are uncorrelated (show no net *linear* trend) when viewed over the entire range of x. The mean residual error ($\overline{e} = 0$) is denoted by the dashed line on the right graph.

Note, however, that $\text{Cov}(x, e) = 0$ does not guarantee that e and x are independent. In Figure 3.3, for example, because of a nonlinear relationship between y and x, the residual errors associated with extreme values of x tend to be negative while those for intermediate values are positive. Thus, if the true regression is nonlinear, then $E(\,e\,|\,x\,) \neq 0$ for some x values, and the predictive power of the linear model is compromised. Even if the true regression is linear, the variance of the residual errors may vary with x, in which case the regression is said to display **heteroscedasticity** (Figure 3.4). If the conditional variance of the residual errors given any specified x value, $\sigma^2(\,e\,|\,x\,)$, is a constant (i.e., independent of the value of x), then the regression is said to be **homoscedastic**.

5. There is an important situation in which *the true regression, the value of $E(\,y\,|\,x\,)$, is both linear and homoscedastic — when x are y are bivariate normally distributed*. The requirements for such a distribution are that the univariate distributions of both x and y are normal and that the conditional distributions of y given x, and x given y, are also normal (Chapter 8). Since statistical testing is simplified enormously, it is generally desirable to work with normally distributed data. For situations in which the raw data are not so distributed, a variety of transformations exist that can render the data close to normality (Chapter 11).

6. It is clear from Equations 3.14a,b that *the regression of y on x is different from the regression of x on y unless the means and variances of the two variables are equal*. This distinction is made by denoting the regression coefficient by $b(y, x)$ or $b_{y,x}$ when x is the predictor and y the response variable.

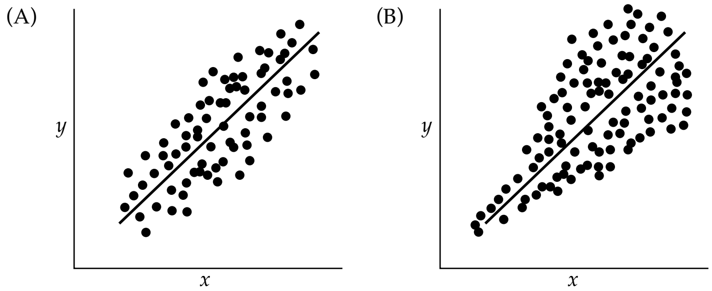

Figure 3.4 The dispersion of residual errors around a regression. (A) The regression is homoscedastic — the variance of residuals given x is a constant. (B) The regression is heteroscedastic — the variance of residuals increases with x. In this case, higher x values predict y with less certainty.

For practical reasons, we have expressed properties 1 – 6 in terms of the estimators a, b, Cov(x, y), and Var(x). They also hold when the estimators are replaced by the true parameters α, β, $\sigma(x, y)$, and $\sigma^2(x)$.

Example 1. Suppose Cov(x, y) = 10, Var$(x) = 10$, Var(y) = 15, and $\overline{x} = \overline{y} = 0$. Compute the least-squares regressions of y on x, and of x on y.

From Equation 3.14a, $a = 0$ for both regressions. However,

$$b(y, x) = \text{Cov}(x, y)/\text{Var}(x) = 10/10 = 1$$

while $b(x, y)$ = Cov(x, y)/Var$(y) = 2/3$. Hence, $\widehat{y} = x$ is the least-squares regression of y on x, while $\widehat{x} = (2/3)y$ is the regression of x on y.

CORRELATION

For purposes of hypothesis testing, it is often desirable to use a dimensionless measure of association. The most frequently used measure in bivariate analysis is the **correlation coefficient**,

$$r(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\,\text{Var}(y)}} \tag{3.15a}$$

Note that $r(x, y)$ is symmetrical, i.e., $r(x, y) = r(y, x)$. Thus, where there is no ambiguity as to the variables being considered, we abbreviate $r(x, y)$ as r. The parametric correlation coefficient is denoted by $\rho(x, y)$ (or ρ) and equals $\sigma(x, y)/\sigma(x)\sigma(y)$. The least-squares regression coefficient is related to the correlation coefficient by

$$b(y, x) = r\sqrt{\frac{\text{Var}(y)}{\text{Var}(x)}} \tag{3.15b}$$

An advantage of correlations over covariances is that the former are *scale independent*. This can be seen by noting that if w and c are constants,

$$r(w\,x, c\,y) = \frac{\text{Cov}(w\,x, c\,y)}{\sqrt{\text{Var}(w\,x)\,\text{Var}(c\,y)}} = \frac{w\,c\,\text{Cov}(x, y)}{\sqrt{w^2\,\text{Var}(x)\,c^2\,\text{Var}(y)}} = r(x, y) \tag{3.16a}$$

Thus scaling x and/or y by constants does not change the correlation coefficient, although the variances and covariances are affected. Since r is dimensionless with limits of ± 1, it gives a direct measure of the degree of association: if $|r|$ is close to one, x and y are very strongly associated in a linear fashion, while if $|r|$ is close to zero, they are not.

The correlation coefficient has other useful properties. First, *r is a standardized regression coefficient (the regression coefficient resulting from rescaling x and y such that each has unit variance)*. Letting $x' = x/\sqrt{\text{Var}(x)}$ and $y' = y/\sqrt{\text{Var}(y)}$ gives $\text{Var}(x') = \text{Var}(y') = 1$, implying

$$b(y', x') = b(x', y') = \text{Cov}(x', y') = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\,\text{Var}(y)}} = r \tag{3.16b}$$

Thus, when variables are standardized, the regression coefficient is equal to the correlation coefficient regardless of whether x' or y' is chosen as the predictor variable.

Second, *the squared correlation coefficient measures the proportion of the variance in y that is explained by assuming that $E(y|x)$ is linear.* The variance of the response variable y has two components: $r^2\,\text{Var}(y)$, the amount of variance accounted for by the linear model (the **regression variance**), and $(1 - r^2)\,\text{Var}(y)$, the remaining variance not accountable by the regression (the **residual variance**). To obtain this result, we derive the variance of the residual deviation defined in Equation 3.13a,

$$\begin{aligned}
\text{Var}(e) &= \text{Var}(\,y - a - bx\,) = \text{Var}(\,y - bx\,) \\
&= \text{Var}(y) - 2\,b\,\text{Cov}(x, y) + b^2\,\text{Var}(x) \\
&= \text{Var}(y) - \frac{2\,[\,\text{Cov}(x, y)\,]^2}{\text{Var}(x)} + \frac{[\,\text{Cov}(x, y)\,]^2\,\text{Var}(x)}{[\,\text{Var}(x)\,]^2} \\
&= \left(1 - \frac{[\,\text{Cov}(x, y)\,]^2}{\text{Var}(x)\,\text{Var}(y)}\right)\text{Var}(y) = (1 - r^2)\,\text{Var}(y)
\end{aligned} \tag{3.17}$$

Example 2. Returning to Table 2.1, the preceding formulae can be used to characterize the relationship between maternal weight and offspring number in rats. Here we take offspring number as the response variable y and maternal weight as the predictor variable x. The mean and variance for maternal weight were found to be $\overline{x} = 118.90$ and Var$(x) = 623.06$ (Table 2.1). For offspring number, $\overline{y} = 5.49$ and Var$(y) = 2.94$. In order to obtain an estimate of the covariance, we first require an estimate of $E(x\,y)$. Taking the xy cross-product of all classes in Table 2.1 (using the midpoint of the interal for the value of x) and weighting them by their frequencies,

$$\overline{xy} = \frac{(1 \cdot 4 \cdot 55) + (3 \cdot 5 \cdot 55) + (1 \cdot 6 \cdot 55) + \cdots + (1 \cdot 10 \cdot 195)}{1003} = 660.14$$

The covariance estimate is then obtained using Equation 3.9,

$$\text{Cov}(x, y) = \frac{1003}{1002} \left[\, 660.14 - (118.90 \times 5.49) \,\right] = 7.39$$

From Equation 3.14b, the slope of the regression is found to be

$$b(y, x) = \frac{7.39}{623.06} = 0.01$$

Thus, the expected increase in number of offspring per gram increase in maternal weight is about 0.01. How predictable is this change? From Equation 3.15a, the correlation coefficient is estimated to be

$$r = \frac{7.39}{\sqrt{623.06 \times 2.94}} = 0.17$$

Squaring this value, $r^2 = 0.03$. Therefore, only about 3 percent of the variance in offspring number can be accounted for with a model that assumes a linear relationship with maternal weight.