17 Regression

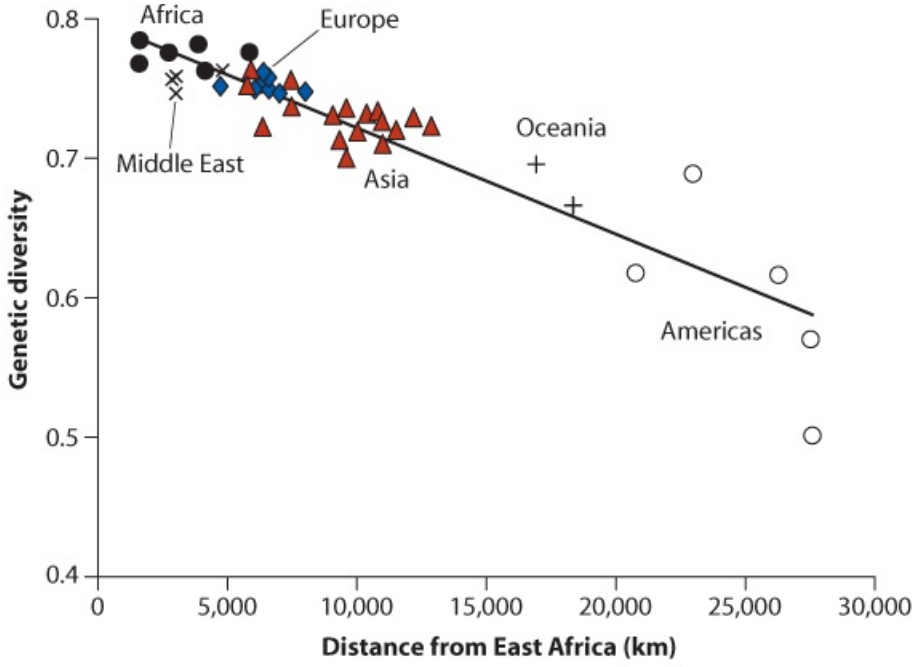
Red campion, *Silene dioica*
haraldmuc/Shutterstock

Regression is the method used to predict values of one numerical variable from values of another. For example, the scatter plot on the following page shows how genetic diversity in a local contemporary human population is predicted by its dispersal distance from East Africa by fitting a straight line to the data points.[1] Modern humans emerged from Africa around 60,000 years ago, and our ancestors lost some genetic variation at each step as they spread to new lands.

> ***Regression*** is a method that predicts values of one numerical variable from values of another numerical variable.

The line fitted to the data is the regression line. The line can be used to *predict* the genetic

diversity of a local human population (the response variable), even for a locale not included in this study, based on its dispersal distance from East Africa (the explanatory variable). The slope or steepness of the regression line indicates the *rate of change* of genetic diversity with distance. It shows that humans lose 0.076 units of genetic diversity, about 10% of the maximum, with every 10,000-km distance from East Africa. Both features of the relationship are captured in the equation for the line.



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

In this chapter, we show how to estimate the regression line, how to put bounds on its predictions, and how to test hypotheses about the slope. Our focus is linear regression, but we also introduce some general principles of nonlinear regression.

Like correlation (Chapter 16), linear regression measures aspects of the linear relationship between two numerical variables. However, there are important differences. Regression fits a line through the data to predict one variable from another and to measure how steeply one variable changes with changes in the other. Correlation does none of these things. It measures strength of association between two variables, reflecting the amount of scatter in the data.

Regression is used on data from either of two study designs. In the first, individuals are randomly sampled from a population. Two variables are measured on the sample, and one of them (deemed the explanatory variable) is used to predict the other (response) variable. In the second design, the researcher fixes or chooses values of the explanatory variable, which represent treatments or doses. The response variable is then measured on one or more individuals assigned to each treatment. The calculations are the same in both cases. Examples 17.1 and 17.3 in this chapter illustrate these two designs.

Linear regression

The most common type of regression is **linear regression,** which draws a straight line through the data to predict the response variable (Y, shown on the vertical axis) from the explanatory variable (X, shown on the horizontal axis). One important assumption of the linear regression method is that the relationship between the two variables really is linear. Example 17.1 shows how to use linear regression to predict the value of a response variable.

EXAMPLE 17.1 The lion's nose

Managing the trophy hunting of African lions is an important part of maintaining viable lion populations. Knowing the ages of the male lions helps, because removing males older than six years has little impact on lion social structure, whereas taking younger males is more disruptive. Whitman et al. (2004) showed that the amount of black pigmentation on the nose of male lions increases as they get older and so might be used to estimate the age of unknown lions. The relationship between age and the proportion of black pigmentation on the noses of 32 male lions of known age in Tanzania is shown in the scatter plot in Figure 17.1-1. The raw data are listed in Table 17.1-1. We can use these data to predict a lion's age from the proportion of black on his nose.

Deborah Kolb/Shutterstock

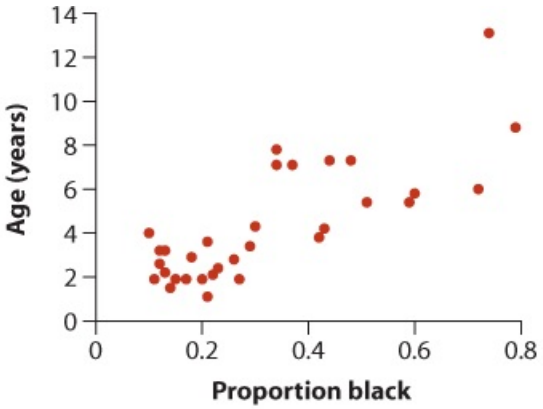


Figure 17.1-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.1-1 Scatter plot of the known ages of 32 male lions (Y, vertical axis) and the proportion of black on their noses (X, horizontal axis).

TABLE 17.1-1 The proportion of black on the noses of 32 male lions of

known age.

Proportion black	Age (years)
0.21	1.1
0.14	1.5
0.11	1.9
0.13	2.2
0.12	2.6
0.13	3.2
0.12	3.2
0.18	2.9
0.23	2.4
0.22	2.1
0.20	1.9
0.17	1.9
0.15	1.9
0.27	1.9
0.26	2.8
0.21	3.6
0.30	4.3
0.42	3.8
0.43	4.2
0.59	5.4
0.60	5.8
0.72	6.0
0.29	3.4
0.10	4.0
0.48	7.3
0.44	7.3
0.34	7.8
0.37	7.1
0.34	7.1
0.74	13.1
0.79	8.8
0.51	5.4

The scatter plot in Figure 17.1-1 puts age as the response variable (the vertical axis) and the proportion of black on the nose as the explanatory variable (the horizontal axis), rather than the reverse, because we want to predict age from the proportion of black, not the other way around.

The method of least squares

Many straight lines can be drawn through a scatter of points, so how do we find the “best” one?

Ideally, we would find a line that leads to the most accurate predictions of *Y* from *X*. This is the line that has the smallest possible deviations in *Y* (the vertical axis) between the data points and the regression line (Figure 17.1-2).

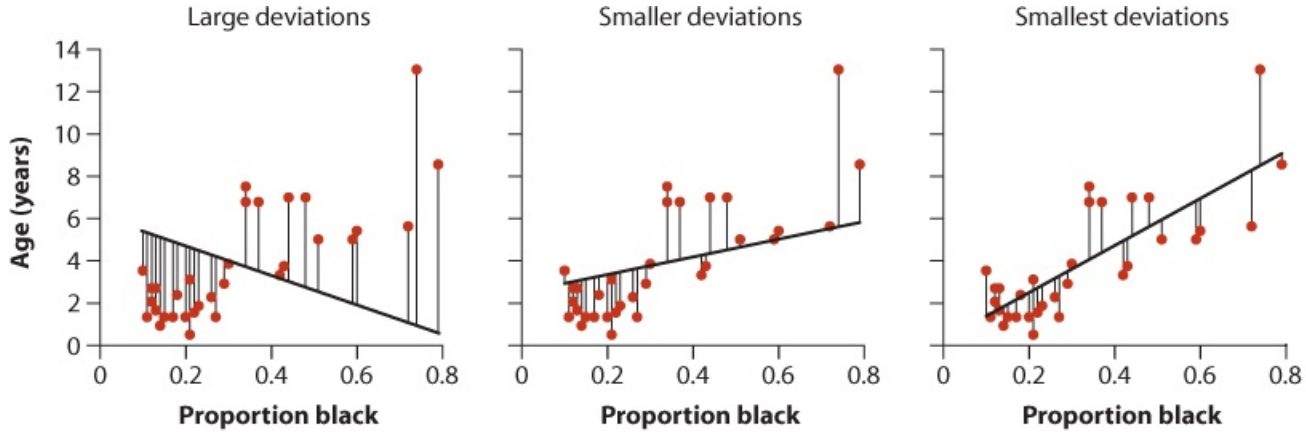


Figure 17.1-2
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.1-2 Illustration of the deviations between the data and several possible regression lines (the heavy black lines) drawn through the data points originally plotted in Figure 17.1-1. Vertical lines are the deviations in *Y* between each point and the regression line. The line in the right panel is the least-squares regression line.

The **least-squares regression** line is the line for which the sum of all the *squared* deviations in *Y* is smallest. We square the deviations from the regression line for the same reason that we square deviations from the mean when calculating an ordinary variance—to overcome the fact that some deviations are positive (the points above the regression line) and others are negative (the points below the regression line), which would cancel each other out in a simple average.

Formula for the line

The regression line through a scatter of points is described mathematically by the following equation:

Y=a+bX.$Y = a + bX.$

The symbol *Y* is the response variable (displayed on the vertical axis in a scatter plot), and *X* is the explanatory variable (the horizontal axis). The formula has two coefficients, *a* and *b*. The coefficient *a* is the *Y*-intercept, or just the **intercept.** Mathematically, *a* is the value of *Y* when *X* is zero (hence, it is the *Y*-value where the regression line "intercepts" the y-axis). Its units are the same as the units of *Y*. The right panel in Figure 17.1-3 shows two regression lines that have different intercepts.

The coefficient *b* is the **slope** of the regression line. It measures how much *Y* changes per unit change in *X*. Its units are the ratio of the units of *Y* and *X*. If *b* is positive, then larger values of *X* predict larger values of *Y*. If *b* is negative, then larger values of *X* predict smaller values of *Y*. The first three panels of Figure 17.1-3 show the slope of the line when *b* is positive, negative, and equal to zero.

The ***slope*** of a linear regression is the rate of change in *Y* per unit of *X*.

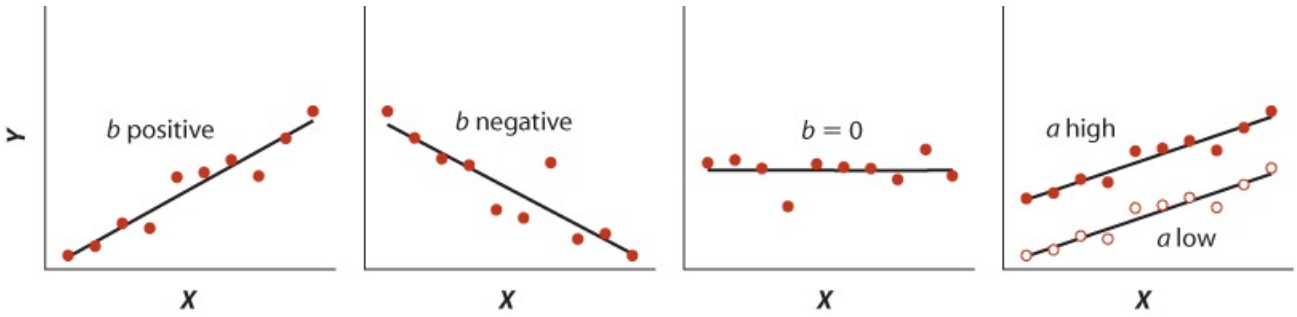


Figure 17.1-3
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.1-3 Comparing the slope of a line when *b* is positive (*farleft)*, negative (*left)*, and zero (*right);* comparing a line with a high intercept *a* and one with a low intercept *a* (*far right).*

Calculating the slope and intercept

Typically, you would use a computer to calculate the regression line, but we provide the formulas here for use with a calculator. The slope of the least-squares regression line is computed as[2]

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

b=∑i(Xi-X¯)(Yi-Y¯)∑i(Xi-X¯)2,
where X¯ $\overline{X}$ and Y¯ $\overline{Y}$ are the sample means of the two variables, and X_i and Y_i refer to the X and Y measurements of individual i. The top of this formula is the sum of products, something first encountered in Section 16.1. The bottom is the sum of squares for X. Shortcut formulas for these sums are given in the Quick Formula Summary (Section 17.10).

Once we have the slope *b*, getting the intercept is relatively straightforward, because the least-squares regression line always goes through the point (X¯,Y¯ $\overline{X}, \overline{Y}$). As a result, we know that
Y¯=a+bX¯. $\overline{Y} = a + b\overline{X}$.
So, we find *a* by simple algebra:
a=Y¯-bX¯. $a = \overline{Y} - b\overline{X}$.
We can now use these formulas to calculate the coefficients of the least-squares regression line for the lion data in Example 17.1. First, though, we need the following quantities calculated from the data in Table 17.1-1:
X¯=0.3222Y¯=4.3094∑i(Xi−X¯)2=1.2221∑i(Yi−Y¯)2=222.0872∑i(Xi−X¯)

$$\overline{X} = 0.3222 \qquad \overline{Y} = 4.3094$$

$$\sum_i (X_i - \overline{X})^2 = 1.2221 \qquad \sum_i (Y_i - \overline{Y})^2 = 222.0872$$

$$\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = 13.0123.$$

(Yi−Y¯)=13.0123.
The slope is then

$$b = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} = \frac{13.0123}{1.2221} = 10.647.$$

b=∑i(Xi-X¯)(Yi-Y¯)∑i(Xi-X¯)2=13.01231.2221=10.647.
The slope *b* measures the change in age of male lions per unit increase in the proportion of black on the nose. Its units are years per unit proportion black.

The intercept, in years, is
$a=\bar{Y}-b\bar{X}=4.3094-10.647(0.3222)=0.879.$
The formula for the line that predicts age from the proportion of black pigmentation on the nose in these lions can be written by putting all of this together, with appropriate rounding:
$Y=0.88+10.65X.$
This equation could also be written as
$\text{Age}=0.88+10.65(\text{proportion black}).$

Figure 17.1-4 shows what this line looks like when it is plotted on the scatter plot[3] shown originally in Figure 17.1-1. The slope of the line indicates that on average, lion age increases by 10.65 years per unit change in the proportion of the nose that is black. We can say, equivalently, that age goes up by 1.065 years for each 0.1 increase of black on the nose.

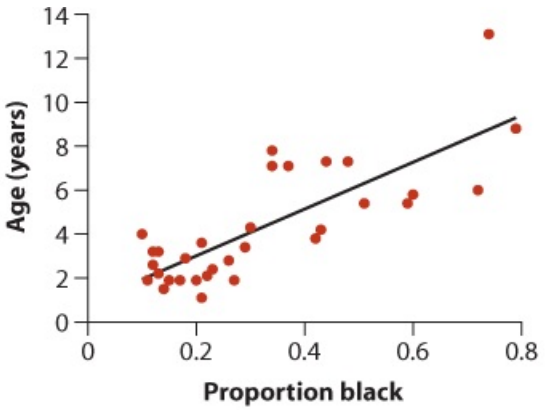


Figure 17.1-4
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 17.1-4 The regression line for the lion data from Example 17.1.

Populations and samples

The regression line is not just calculated for the sake of the data. It is typically used to estimate the true regression of Y on X in the population from which the data are a sample. The regression equation for the population is
$Y=\alpha+\beta X,$
where β is the slope in the population, and α is the intercept. The quantities α and β are population parameters, whereas a and b are their sample estimates.

Under one sampling scenario for linear regression, we have a random sample of (X,Y) pairs of measurements from a population. The lion data correspond to this scenario. Or, under a second scenario, the researcher fixes or chooses values of X to include in the study, and Y is then measured on a sample of one or more individuals for each X-value included in the study. In either case, regression assumes that there is a population of *possible* Y-values for every possible value of X. The mean Y-value for each value of X lies on the true regression line.

For example, one of the lions in the data for Example 17.1 has a value of 0.6 for the proportion of black on its nose. We assume that there is a population of lions having the value $X = 0.6$ for the proportion of black on their noses, even though the data includes just one lion with the value. The mean age of all lions in the population having $X = 0.6$ is assumed to lie on the true regression line.

Predicted values

Now that we have the regression line, we can use it to determine points on the line that correspond to specified values of *X*. These points on the regression line are called **predictions.** We will symbolize predictions as Y^$\hat{Y}$ ("Y-hat") to distinguish them from values of *Y* (i.e., actual data points), which lie above or below the line but not usually on it. The predicted value of *Y* for a given value of *X* estimates the mean of *Y* for the whole population of individuals having that value of *X*. For example, to predict the age of a male lion corresponding to a proportion of 0.50 black on the nose, plug the value *X* = 0.50 into the regression formula:
Y^=a+b(0.50)=0.88+10.65(0.50)=6.2.$\hat{Y} = a + b\,(0.50) = 0.88 + 10.65\,(0.50) = 6.2.$
In other words, the regression line predicts that lions with a proportion of black *X* = 0.50 will be 6.2 years old on average. If we observed a lion with 0.5 proportion black on its nose, we could predict its age, even though we had never seen a lion exactly like that before.

According to Table 17.1-1, the value *X* = 0.50 was not represented in the sample, although it falls within the range of observed *X*-values (i.e., 0.10 to 0.79). For reasons that are explained in Section 17.2, we can reliably make predictions only by using values of *X* that lie within the range of values in the sample.

> The predicted value of ***Y*** from a regression line estimates the mean value of *Y* for all individuals having a given value of *X*.

Residuals

Residuals measure the scatter of points above and below the least-squares regression line. They are crucial for evaluating the fit of the line to the data. Each observation in the data has a corresponding residual, measuring the vertical deviation from the least-squares regression line (see the right panel in Figure 17.1-2). The point on the regression line used to calculate the residual for individual *i* is Y^i,$\hat{Y}_i$, the value predicted when its corresponding value for X_i is plugged into the regression formula:
Y^i=a+bXi.$\hat{Y}_i = a + bX_i.$
For example, the 31st lion in the sample (*i* = 31) has a proportion X_{31} = 0.79 of black on its nose (see Table 17.1-1). The corresponding age Y^31$\hat{Y}_{31}$ predicted for a lion with this much black on the nose is
Y^31=0.88.+10.65(0.79)=9.29.$\hat{Y}_{31} = 0.88. + 10.65\,(0.79) = 9.29.$
The actual age of the lion was 8.8 years, which is below the predicted value. The residual is the observed value minus the predicted value:
residual31=(Y31−Y^31)=(8.8−9.3)=−0.49 years.$\text{residual}_{31} = \left(Y_{31} - \hat{Y}_{31}\right) = (8.8 - 9.3) = -0.49\,\text{years}.$
The variance of the residuals, symbolized as $MS_{residual}$, quantifies the spread of the scatter of points above and below the line. In regression jargon, this variance is called the "residual mean square":
MSresidual=∑i(Yi-Y^i)2n-2.$MS_{\text{residual}} = \dfrac{\sum_i \left(Y_i - \hat{Y}_i\right)^2}{n-2}.$
The $MS_{residual}$ is like an ordinary variance, but it has *n* − 2 degrees of freedom[4] rather than *n* - 1. It is analogous to the error mean square in the analysis of variance (Section 15.1). The following alternate formula is easier to use, though, because you don't need to calculate each $\hat{Y}_i$:

MSresidual=∑i(Yi-Y¯)2-b∑i(Xi-X¯)(Yi-Y¯)n-2. $$\mathrm{MS}_{\mathrm{residual}} = \frac{\sum_i (Y_i - \overline{Y})^2 - b\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{n-2}.$$
Shortcuts for calculating the sum of squares and products are provided in the Quick Formula Summary (Section 17.10).

All of the quantities needed to determine $\mathrm{MS}_{\mathrm{residual}}$ for the lion data have been calculated previously on page 544. Inserting these values into the equation for $\mathrm{MS}_{\mathrm{residual}}$ yields

MSresidual=222.0872-10.647(13.0123)32-2=2.785. $$\mathrm{MS}_{\mathrm{residual}} = \frac{222.0872 - 10.647(13.0123)}{32-2} = 2.785.$$

Standard error of slope

Like any other estimate, there is uncertainty associated with the sample estimate *b* of the population slope *β*. Uncertainty is measured by the standard error, the standard deviation of the sampling distribution of *b*. The smaller the standard error, the higher the precision and the lower the uncertainty of the estimate of the slope. If the assumptions of linear regression are met (Section 17.5), then the sampling distribution of *b* is a normal distribution having a mean equal to *β* and a standard error estimated from data as

SEb=MSresidual∑i(Xi-X¯)2. $$\mathrm{SE}_b = \sqrt{\frac{\mathrm{MS}_{\mathrm{residual}}}{\sum_i (X_i - \overline{X})^2}}.$$
The quantity on top of the fraction under the square-root sign is the residual mean square, and the quantity on the bottom is the sum of squares for *X*.

The standard error of *b* for the lion data is

SEb=MSresidual∑i(Xi-X¯)2=2.7851.2221=1.510. $$\mathrm{SE}_b = \sqrt{\frac{\mathrm{MS}_{\mathrm{residual}}}{\sum_i (X_i - \overline{X})^2}} = \sqrt{\frac{2.785}{1.2221}} = 1.510.$$
The standard error of the slope has the same units as the slope itself (i.e., years per unit of proportion black for the lion data in Example 17.1).

Confidence interval for the slope

A confidence interval for the parameter *β* is given by

b−tα(2),dfSEb<β<b+tα(2),dfSEb, $$b - t_{\alpha(2),df}\mathrm{SE}_b < \beta < b + t_{\alpha(2),df}\mathrm{SE}_b,$$
where $t_{\alpha(2),df}$ is the two-tailed critical value of the *t*-distribution having $df = n - 2$ degrees of freedom. For a 95% confidence interval, $\alpha = 0.05$, and for a 99% confidence interval, $\alpha = 0.01$. For the lion data, $t_{0.05(2),30} = 2.042$ (Statistical Table C), so the 95% confidence interval for the slope is

10.647−2.042(1.510)<β<10.647+2.042(1.510)7.56<β<13.73. $$10.647 - 2.042\,(1.510) < \beta < 10.647 + 2.042\,(1.510)\;7.56 < \beta < 13.73.$$
This is a modest range of most-plausible values for the slope. The mean age of lions increases by as little as 7.6 years per unit proportion of black on the nose, or by as much as 13.7 years.

Confidence in predictions

The regression line calculated from data predicts the mean value of Y for any specified value of X lying between the smallest and largest X in the data. This line is calculated with error, however, which affects how precise the predictions are. Here in Section 17.2 we quantify the precision of predictions. We also discuss the hazards of *extrapolating*—making predictions when the values of X lie beyond the range of X-values in the data.

Confidence intervals for predictions

Two subtly different types of predictions can be made using the regression line. The first predicts the *mean* Y for a given X. What, for example, is the mean age of all male lions in the population whose noses are 60% black (i.e., $X = 0.60$)? The second type predicts a *single* Y for a given X. (For example, how old is that lion over there, given that 60% of its nose is black?) Usually we just want to predict the mean Y for each X (i.e., the first prediction) because we are interested in the overall trend. In special situations, though, we also want to predict an individual Y-value (i.e., the second prediction). This is especially true in the lion study (Example 17.1). A hunter who encounters a male lion would want to know the age of that specific lion if he or she wishes to avoid shooting a young male.

Both types of predictions generate the same value for Y^.$\hat{Y}$. They differ in the precision of the predictions. In the case of lions with 60% black on their noses, Y^=a+bX=0.88+10.65(0.60)=7.27 years.$\hat{Y} = a + bX = 0.88 + 10.65\,(0.60) = 7.27$ years.

Regardless of the prediction goals, this is the best prediction of age. The precision of the prediction is lower, however, if the goal is to predict the age of an individual lion rather than the mean age of lions having the specified proportion of black on their noses. This is because the prediction for a single Y-value includes uncertainty stemming from variation in Y among the individuals in the population having the same value of X (i.e., not all male lions having 60% black noses are the same age). The two graphs in Figure 17.2-1 illustrate these differences in precision of predictions using confidence intervals.

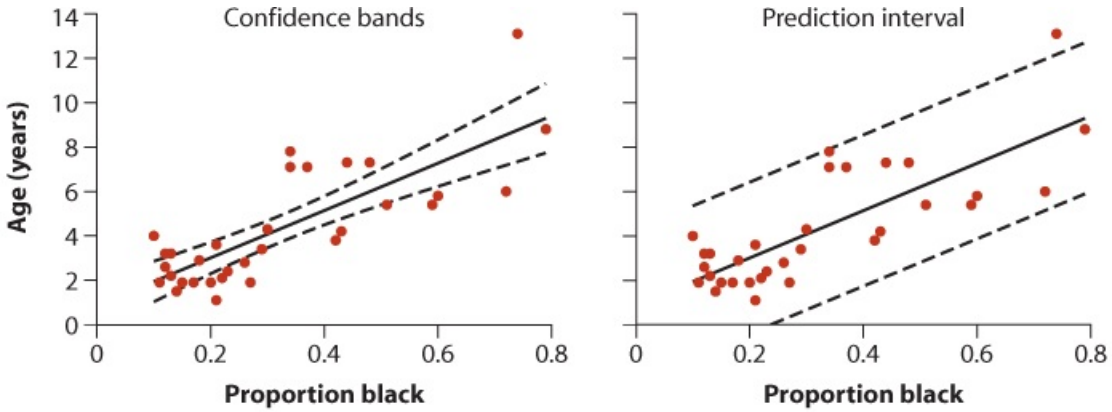


Figure 17.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.2-1 *Left:* 95% confidence bands for the predicted mean age of male lions at every value of proportion of black on their noses. *Right:* 95% prediction intervals for the predicted age of single lions. $n = 32$.

The left panel of Figure 17.2-1 shows the 95% confidence intervals for the predicted mean lion age at every *X*. The upper curve connects the upper bounds of all of the 95% confidence intervals for the predicted mean *Y*-values, one for every *X* between the smallest and largest *X* in the data. The lower curve connects the lower bounds of these same confidence intervals. Together the upper and lower curves showing the confidence intervals for the mean *Y* are called the 95% **confidence bands.** These bands are narrowest in the vicinity of $\bar{X}$, the mean value for proportion of black on the nose, and they flare outward toward the extremes of the range of data. The uncertainty of predictions always increases the farther the *X*-value is from the mean *X* in the data. In 95% of samples, the confidence bands will bracket the true regression line in the population.

The right panel of Figure 17.2-1 shows the 95% **prediction intervals.** The upper and lower curves connect the upper and lower limits of the 95% prediction intervals for a *single Y* over the range of *X*-values in the data. These are much wider than the confidence bands because predicting an individual lion's age from the color of its nose is more uncertain than predicting the mean age of all lions having the same proportion of black on their noses. Prediction intervals bracket most of the individual data points in the sample, because they incorporate the variability in *Y* from individual to individual at a given *X*.

> ***Confidence bands*** measure the precision of the predicted mean *Y* for each value of *X*.
> *Prediction intervals* measure the precision of the predicted single *Y*-values for each *X*.

Most statistical packages on the computer will calculate and display confidence bands and prediction intervals. We haven't given calculation details, but we provide the formulas in the Quick Formula Summary (Section 17.10).

Extrapolation

We've stressed that regression can be used to predict *Y* for any value of *X* lying between the smallest and largest values of *X* in the data set. Regression cannot be used to predict the value of the response variable when an *X*-value lies well outside the range of the data. This is because there is no way to ensure that the relationship between *X* and *Y* continues to be linear beyond the range of the data. Predicting *Y* for *X*-values beyond the range of the data is called **extrapolation.** The graph in Figure 17.2-2 illustrates the problem.

> ***Extrapolation*** is the prediction of the value of a response variable outside the range of *X*-values in the data.

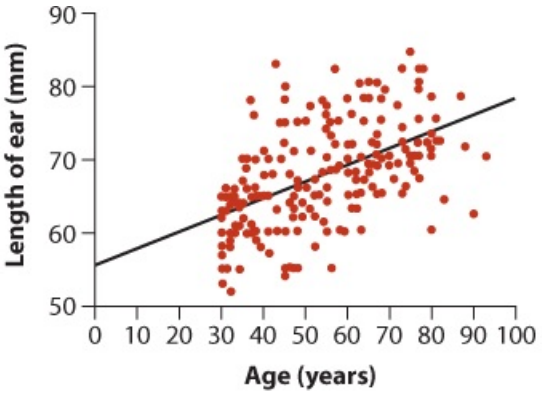


Figure 17.2-2
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 17.2-2 Ear lengths of 206 adults 30 years old or more as a function of their ages. Modified from Heathcote (1995).

The data are measurements of ear length taken on a sample of adults at least 30 years old (Heathcote 1995). The linear regression equation calculated from these data (in millimeters) is ear length=55.9+0.22(age). $\text{ear length} = 55.9 + 0.22\,(\text{age})$.
The results suggest that our ears grow longer by about 0.22 mm per year on average as we age. The intercept of this equation, which predicts the ear length at birth (i.e., when age is zero), is 56 mm. This makes no sense, though. To quote the authors of the study (Altman and Bland 1998), “A baby with ears 5.6 cm long would look like Dumbo.” The relationship between ear length and age is not linear from birth, but we wouldn’t know this unless we took measurements over the complete range of ages.

Michael Whitlock

17.3 Testing hypotheses about a slope

Hypothesis testing in regression is used to evaluate whether the population slope equals a null hypothesized value, β_0, which is typically (but not always) zero. The test statistic t is

t=b-β0SEb, $t = \frac{b - \beta_0}{\text{SE}_b}$,

where b is the estimate of the slope in the sample and SE_b is the standard error of b. Under the null hypothesis, this test statistic has a t-distribution with $n - 2$ degrees of freedom. Example 17.3 shows how to use this test.

EXAMPLE 17.3 Prairie Home Campion

© Cedar Creek Ecosystem Science Reserve

Human activity is reducing species numbers in many of the world's ecosystems. Does this decrease affect basic ecosystem properties? Or are different plant species largely substitutable, with lost species compensated by those species remaining? To find out, Tilman et al. (2006) seeded 161 plots, each measuring 9 × 9 meters, at the Cedar Creek Reserve in Minnesota. They used a varying number of prairie plant species and measured plant biomass production over 10 subsequent years. Treatments of either 1, 2, 4, 8, or 16 plant species (randomly chosen from a set of 18 perennials) were randomly assigned to plots. After 10 years of measurement, the researchers measured the "stability" of plant biomass production in every plot as mean biomass divided by the standard deviation in biomass over the 10 years (the reciprocal of the coefficient of variation; Section 3.1). Results are plotted in Figure 17.3-1. Stability has been log-transformed to reduce skew. The data are available at whitlockschluter.zoology.ubc.ca.

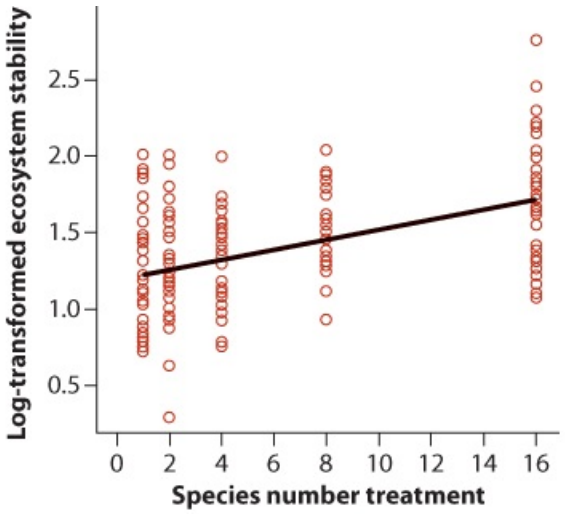


Figure 17.3-1
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 17.3-1 Stability of plant biomass production over 10 years in 161 plots and the initial number of plant species assigned to plots. Stability was log-transformed (natural log) to better meet the assumptions of regression. The line is the least-squares regression line. Data are from Tilman et al. (2006).

Unlike the previous example, involving lions, the data in Example 17.3 are from an experiment in which the values of the explanatory variable were fixed treatments. In contrast to correlation, regression does not require the explanatory variable to follow a normal distribution.

The *t*-test of regression slope

The null hypothesis is that the measure of ecosystem stability cannot be predicted from the species number treatment—that is, the slope of the linear regression of ecosystem stability on number of species is zero. The alternative hypothesis is that stability either increases or decreases with increasing number of species. This is a two-tailed test.

H_0: The slope of the regression of log ecosystem stability on species number is zero ($\beta = 0$).

H_A: The slope of the regression of log ecosystem stability on species number is not zero ($\beta \neq 0$).

The following quantities calculated from the data are needed for the *t*-test of zero slope:
X¯=6.3168Y¯=1.4063∑i(Xi-X¯)2=5088.8447∑i(Yi-Y¯)2=24.8149∑i(Xi-X¯)(Yi-

$$\overline{X} = 6.3168 \quad \overline{Y} = 1.4063$$

$$\sum_i (X_i - \overline{X})^2 = 5088.8447 \qquad \sum_i (Y_i - \overline{Y})^2 = 24.8149$$

$$\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = 167.5548 \quad n = 161.$$

Y¯)=167.5548n=161.
The best estimate of the slope is
b=∑i(Xi-X¯)(Yi-Y¯)∑i(Xi-X¯)2=167.55485088.8447=0.03293.

$$b = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} = \frac{167.5548}{5088.8447} = 0.03293.$$

Calculating the intercept as before, and rounding, we get the least-squares regression line as
Y=1.20+0.033X,$Y = 1.20 + 0.033X,$

which can also be written as
Log stability=1.20+0.033(number of species).$\text{Log stability} = 1.20 + 0.033\,(\text{number of species})\,.$
This line has a positive slope, as shown in Figure 17.3-1. The estimate of slope (b = 0.033) indicates that log stability of biomass production rises by the amount 0.033 for every species added to plots.

To calculate the standard error of the slope, we need the mean square residual:
MSresidual=∑i(Yi-Y¯)2-b∑i(Xi-X¯)(Yi-Y¯)n-2=24.8149-0.03293(167.5548)161-2=0.12137.

$$\begin{aligned}\text{MS}_{\text{residual}} &= \frac{\sum_i (Y_i-\overline{Y})^2 - b\sum_i (X_i-\overline{X})(Y_i-\overline{Y})}{n-2} \\ &= \frac{24.8149-0.03293(167.5548)}{161-2} \\ &= 0.12137.\end{aligned}$$

Thus, the standard error of b is
SEb=MSresidual∑i(Xi-X¯)2=0.121375088.8447=0.004884.

$$\text{SE}_b = \sqrt{\frac{\text{MS}_{\text{residual}}}{\sum_i (X_i-\overline{X})^2}} = \sqrt{\frac{0.12137}{5088.8447}} = 0.004884.$$

We now have all of the elements needed to calculate the t-statistic:
t=b-β0SEb=0.03293-00.004884=6.74.$t = \frac{b-\beta_0}{\text{SE}_b} = \frac{0.03293-0}{0.004884} = 6.74.$
We must compare this t-statistic with the t-distribution having $df = n - 2 = 161 - 2 = 159$ degrees of freedom. Using a computer, we find that $t = 6.74$ corresponds to $P = 2.7 \times 10^{-10}$, so we reject the null hypothesis. We reach the same conclusion if we use the critical value for the t-distribution with $df = 159$ (Statistical Table C):
t0.05(2),159=1.97.$t_{0.05(2),159} = 1.97.$
Since $t = 6.74$ is greater than 1.97, $P < 0.05$, and we reject H_0. In other words, increasing the number of plant species in plots increases the stability of plant biomass production of the ecosystem. A 95% confidence interval for the population slope, calculated from the formula in Section 17.1, is
0.0233<β<0.0426.$0.0233 < \beta < 0.0426.$
indicating that the estimate of slope has fairly tight bounds.

The ANOVA approach

In the literature, and in the output of regression analyses conducted on the computer, you will encounter tests of regression slopes that use an F-test rather than a t-test. Just as ANOVA can be used to compare two population means in place of the two-sample t-test, ANOVA can be used to test for a significant slope in place of the t-test of slope. The resulting P-values are identical. Table 17.3-2 shows the ANOVA table for the ecosystem stability data. Formulas for the quantities are given in the Quick Formula Summary (Section 17.10).

TABLE 17.3-2 ANOVA table testing the effect of plant species number on the stability of bio-mass production.

Source of variation	Sum of Squares	*df*	Mean Squares	*F*-ratio	*P*
Regression	5.5169	1	5.5169	45.45	2.73×10^{-10}

Residual	19.2980	159	0.1214
Total	24.8149	160	

The basic idea behind the ANOVA approach in regression is similar to that when testing differences among means of multiple groups (Chapter 15). If the null hypothesis is true, then the population regression line is flat with a slope of 0. In this case, the amount of variation in *Y* among individual data points having the same value for *X* (represented by the residual mean square) is expected to equal the amount of variation among data points having different *X* values (represented by the regression mean square), except by chance. If the null hypothesis is false, we expect the regression mean square to exceed the residual mean square. The comparison of mean squares is done with an *F*-ratio.

The first step to estimating these two sources of variation in the data is to take the deviation between each *Y*-measurement Y_i and the grand mean Y¯$\bar{Y}$ and break it into two parts. The residual part is the deviation between Y_i and its predicted value on the regression line (i.e., Yi−Y^i,$Y_i - \hat{Y}_i$, analogous to the "error" component in ANOVA). The regression, on the other hand, is the difference between the predicted value for each point and Y¯$\bar{Y}$ (i.e., Y^i−Y¯, $\hat{Y}_i - \bar{Y}$, analogous to the "groups" component in ANOVA). The sum of squared deviations corresponding to each of these two sources of variation and their total are computed and used to calulate the mean squares. The test statistic is an *F*-ratio of the two mean squares (the mean square regression divided by the mean square residuals). If the null hypothesis is true, and the slope of the population regression β is zero, then the *F*-ratio is expected to be 1 (except by chance). If the slope of the regression is not zero, however, then *F* is expected to be greater than 1.

The ANOVA approach can be used when the test is two-sided and the null hypothesized slope is zero.

Using R^2 to measure the fit of the line to data

We can measure the fraction of variation in *Y* that is "explained" by *X* in the estimated linear regression with the quantity R^2:

R2=SSregressionSStotal.$R^2 = \frac{\text{SS}_{\text{regression}}}{\text{SS}_{\text{total}}}.$

R^2 is calculated from the sums of squares in the ANOVA table, and it is analogous to the R^2 in analysis of variance (Section 15.1), which measures the fraction of variation in the sample of Y-values accounted for by differences between groups.[5] If R^2 is close to one (i.e., its maximum possible value), then *X* predicts most of the variation in the values of *Y*. In this case, the *Y*-observations will be clustered tightly around the regression line with little scatter. If R^2 is close to zero (i.e., its minimum value), then *X* does not predict much of the variation in *Y*, and the data points will be widely scattered above and below the regression line.

For the ecosystem stability study, R^2 can be calculated from the quantities in the ANOVA table, Table 17.3-2:

R2=5.516924.8149=0.222.$R^2 = \frac{5.5169}{24.8149} = 0.222.$

Thus, number of plant species explained 22% of the variation in log-transformed ecosystem stability, a moderate percentage.

17.4 Regression toward the mean

Suppose a study measured the cholesterol levels of 100 men randomly chosen from a population. After their initial measurement, the men were put on a new drug therapy, designed to reduce cholesterol levels. After one year, the cholesterol level of each man was measured again and compared with the first measurement. Figure 17.4-1 shows a scatter plot of the results. The researchers were delighted to find that cholesterol levels had dropped on average in the men who had previously had the highest levels. Their excitement dimmed, though, when they realized that cholesterol levels had *ncreased* on average in the men who had previously had low levels of cholesterol. What happened? Had they discovered a drug with complex effects?

In this hypothetical example, there was no effect at all due to the drug—the trend resulted entirely from a general phenomenon known as **regression toward the mean.** If two variables measured on a sample of individuals, such as consecutive measures of cholesterol, have a correlation less than one, then individuals that are far from the mean on the first measure will on average lie closer to the mean for the second measure. Even without an effect of the drug treatment, average cholesterol levels of the men with the highest levels on the first measure were expected to drop by the second measurement, and average levels of the men who originally had low levels were expected to rise.[6]

Regression toward the mean results when two variables measured on a sample of individuals have a correlation less than one. Individuals that are far from the mean for one of the measurements will, on average, lie closer to the mean for the other measurement.

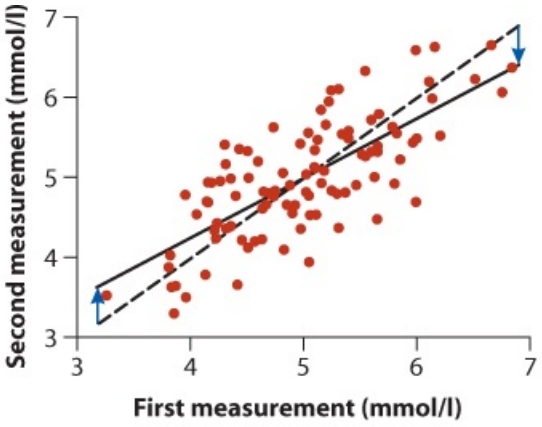


Figure 17.4-1
Whitlock et al., *The Analysis of Biological Data*, 2e,

FIGURE 17.4-1 Regression toward the mean. These hypothetical data are two cholesterol measurements taken on the same 100 men. The dashed line is the one-to-one line with a slope of one. The solid line is the regression line predicting the second measure from the first. It has a slope less than one, as indicated by the blue arrows.

Regression toward the mean is a tricky concept, which is perhaps why it is often overlooked. Think of it this way: each of the men in the study has a "true," underlying cholesterol value, but his "measured" cholesterol value varies randomly with time and

circumstance around the true value. The subset of men who scored highest on the first measurement therefore likely included a disproportionate number of men whose cholesterol measurement was higher than its true value the first time. The second measurement made on each of these men is expected to be closer to its true value on average, bringing down the average for the subset of men as a whole. Similarly, the subset of men who initially scored lowest likely included a disproportionate number of men whose measured values were lower than their true values, so on the second measurement they would seem to improve.[7]

Regression toward the mean is potentially a large problem in any study that tends to focus on individuals in one tail of the distribution. In many medical studies, for example, only sick people are included in the research, as indicated by their initial assessment before the study. Because of regression toward the mean, many of these people will appear to improve even if the treatment has no effect. Interpreting this improvement as if it were a response to the treatment, instead of a mathematical fact of regression, is called the **regression fallacy.** It is one of the reasons that experiments should always include a control group for comparison.

Regression toward the mean is an issue only in observational studies, not in randomized experiments, where the value of the explanatory variable (X) is set by the experimenter. Kelly and Price (2005) discuss ways to disentangle biologically meaningful trends from the effect of regression toward the mean.

Assumptions of regression

When using linear regression, the following assumptions must be met for confidence intervals and hypothesis tests to be accurate:

- At each value of X, there is a population of possible Y-values whose mean lies on the true regression line (this is the assumption that the relationship must be linear).
- At each value of X, the distribution of possible Y-values is normal.
- The variance of Y-values is the same at all values of X.
- At each value of X, the Y-measurements represent a random sample from the population of possible Y-values.

Figure 17.5-1 illustrates the first three of these assumptions. In the next few sections, we explore some of the ways to examine deviations from these assumptions. We also discuss methods to try when the assumptions are not supported by the data.

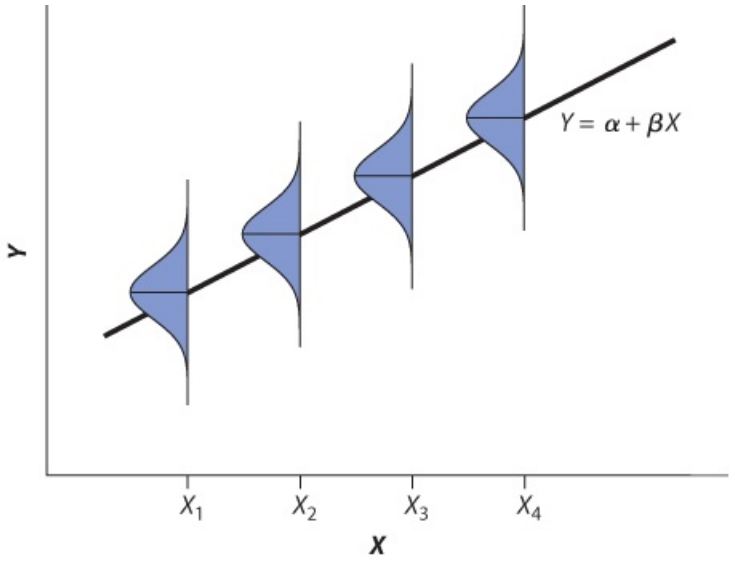


Figure 17.5-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015
W. H. Freeman and Company

FIGURE 17.5-1 Illustration of the assumptions of linear regression. At each value of X, there is a normally distributed population of Y-values with the mean on the true regression line. The variance of the Y-values is assumed to be the same for every value of X.

Unlike correlation analysis, no assumptions are made about the distribution of X when using regression. In regression, for example, it is not necessary that the distribution of X-values is normal. It is not even necessary that X-values are randomly sampled—they might be fixed by the experimenter, instead, as they were in the ecosystem stability study (Example 17.3).

Outliers

Besides creating a non-normal distribution of Y-values at the corresponding value of X, and

violating the assumption of equal variance in Y, outliers disproportionately affect estimates of the regression slope and intercept. If an outlier is present, biologists usually examine and report its influence on the results by comparing the regression line produced with and without the outlier. For example, Figure 17.5-2 shows how the average amount of white in the tails of dark-eyed juncos varies with latitude. One outlier is present, however, indicated in red (a population that formed in 1983 on the campus of the University of California, San Diego). Without the outlier, the estimate of slope is $b = -0.37$ (black line in Figure 17.5-3), and the null hypothesis of zero slope is rejected ($t = -2.66$, $P = 0.024$). Including the outlier changes the slope substantially (red line; $b = -0.18$), and the null hypothesis of zero slope is not rejected ($t = -0.81$, $P = 0.43$).

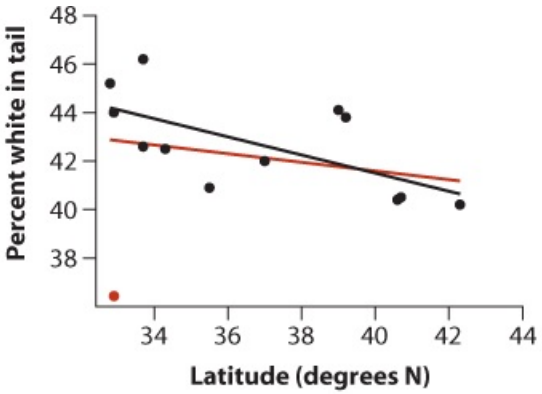


Figure 17.5-2
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 17.5-2 Graph showing the effect of an outlier on an estimate of the regression line. The data are the percentage of white in the tail feathers of the dark-eyed junco at sites at different latitudes in California (Yeh 2004). The black regression line was calculated after excluding the red point on the lower left, whereas the red regression line included it.

Robert L Kothenbeutel/Shutterstock.com

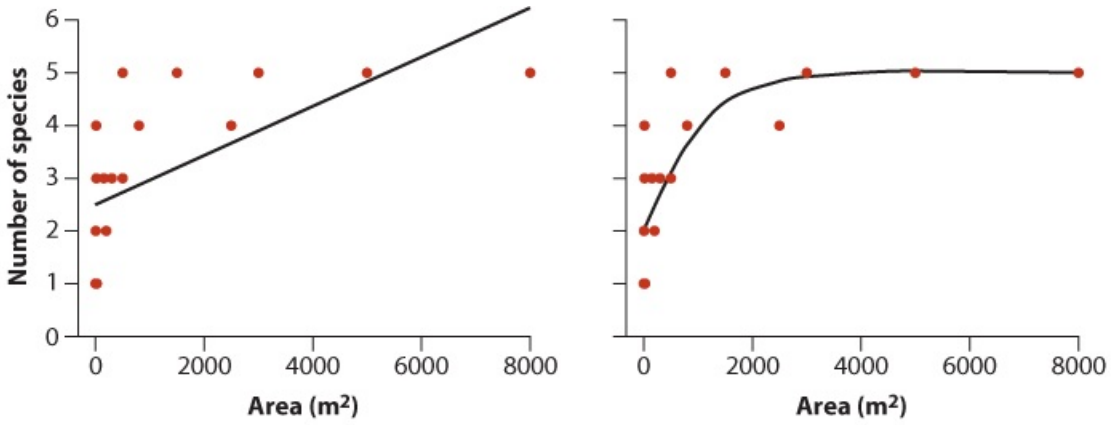


Figure 17.5-3
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.5-3 A scatter plot showing the relationship between the number of fish species and the surface area of 20 desert pools (Kodric-Brown and Brown 1993). The left panel fits a linear regression to the data to highlight how poorly a straight line matches the data. The right panel adds a "smoothed" fit to the same data (see Section 17.8).

Outliers are especially likely to be influential if they occur at or beyond the range of *X*-values in the rest of the data. If the outlier has a large effect, then alternative approaches might also be sought. One approach is to transform *X* or *Y*, such as by taking logarithms, to see if this brings the outlier closer to the rest of the distribution. Further solutions include robust regression methods (Rousseeuw and Leroy 2003) and permutation testing (Chapter 13).

Detecting nonlinearity

Visually inspecting the scatter plot is a useful method for detecting departures from the assumption of a linear relationship between *Y* and *X*. Often, this approach is enough to conclude that the relationship between *X* and *Y* is not linear. Forcing a linear regression through the scatter plot can sometimes make the nonlinearity even more obvious (see the left panel of Figure 17.5-3).

Scatter-plot "smoothing," a method discussed in greater detail in Section 17.8, can also aid the eye in detecting a nonlinear relationship (see the right panel of Figure 17.5-3). Most statistics packages on the computer are able to carry out scatter-plot smoothing.

Detecting non-normality and unequal variance

It is often difficult to decide whether the assumptions of normally distributed residuals and equal variance of residuals are met. Visual inspection of a **residual plot** can help. In a residual plot, the residual for every data point (Yi−Y^i) $\left(Y_i - \hat{Y}_i\right)$ is plotted against X_i, the corresponding value of the explanatory variable. This plot is best made with the aid of a computer. Two examples of residual plots are shown in Figure 17.5-4.

A ***residual plot*** is a scatter plot of the residuals (Yi−Y^i) $\left(Y_i - \hat{Y}_i\right)$ against the X_i, the values of the explanatory variable.

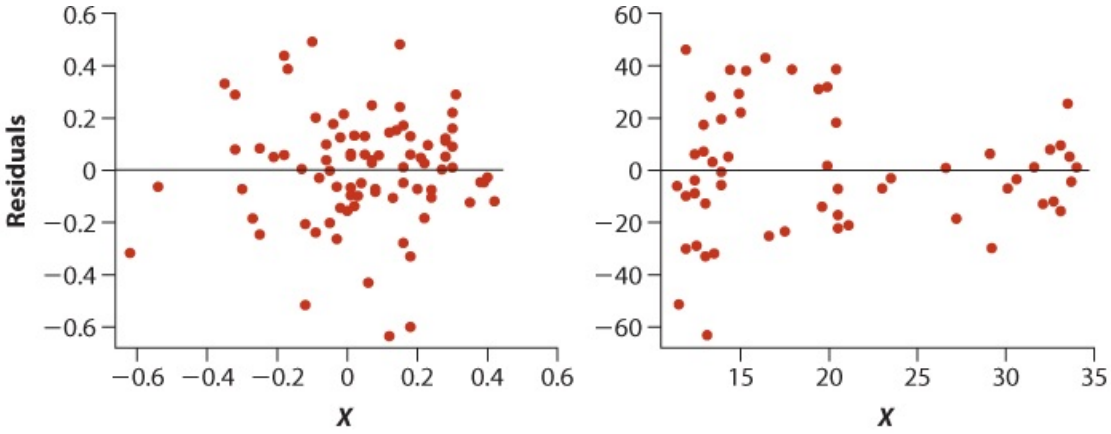


Figure 17.5-4
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.5-4 Two examples of residual plots. Data on the left are from a linear regression of the cap color of offspring on that of their parents in the blue tit, a British bird (Hadfield et al. 2006). Those on the right are from a linear regression of firing rates of cockroach neurons on temperature (Murphy and Heath 1983).

If the assumptions of normality and equal variance of residuals are met, then the residual plot should have all of the following features:

- A roughly symmetric cloud of points above and below the horizontal line at zero, with a higher density of points close to the line than away from the line
- Little noticeable curvature as we move from left to right along the *x*-axis
- Approximately equal variance of points above and below the line at all values of *X*

The blue tit data in the left panel of Figure 17.5-4 fit these requirements reasonably well. The density of observations peaks near the horizontal line and spreads outward above and below in a fairly symmetrical fashion. The spread of points above and below the line is similar across the range of *X*-values. (The spread may seem low at the extreme left end, but we can't tell because there are only two data points). The cockroach data in the right panel of Figure 17.5-4 do not fit these requirements as well. The spread of points above and below the horizontal line is considerably higher at low values of *X* than at high values of *X*.

Normal quantile plots (Section 13.1) and histograms of the residuals are yet other ways to evaluate the assumption that the residuals are normally distributed.

Transformations

Some (but not all) nonlinear relationships can be made linear with a suitable transformation. The most versatile transformation in biology is the log transformation (Section 13.3). The power and exponential relationships, which are commonly encountered in biology, are two nonlinear relationships that can be made linear with a log transformation, as shown in Figure 17.6-1.

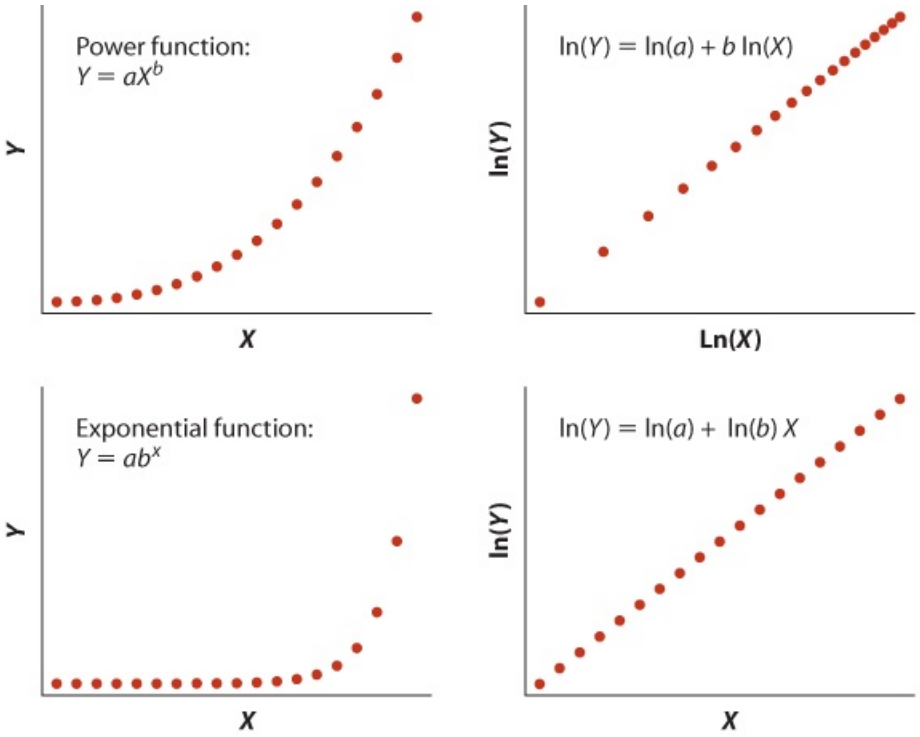


Figure 17.6-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.6-1 The power function (*upper left*) and the exponential function (*lower left*) are two types of nonlinear relationships that can be made linear using the log transformation. Plot log of Y against log of X if the two variables are described by a power function (*upper right*). Plot the log of Y against X if the relationship between these two variables is described by an exponential function (*lower right).*

For example, the relationship between the number of fish species in 20 desert pools and the surface area of pools (Figure 17.5-3) looks like it might fit a power curve or an exponential curve. We tried a log-transformation of both the number of species and the surface area, and we obtained the graph shown in Figure 17.6-2. The straight line fits the transformed data much better than the untransformed data. We can now proceed to estimate parameters and test hypotheses about this relationship using the methods of linear regression, setting Y to be the log of the number of fish species and X to be the log of the surface area of the pools.

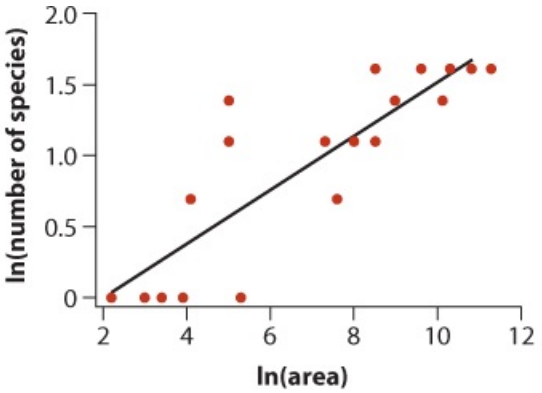


Figure 17.6-2
Whitlock et al., *The Analysis of Biological Data*, 2e,

FIGURE 17.6-2 A scatter plot of the log-transformed number of fish species and surface area of desert pools. This relationship is more linear than the one in Figure 17.5-3, which is based on the untransformed data.

Transformation can also be used to help meet other assumptions of linear regression. For example, if a residual plot reveals that the variance of Y increases with increasing X, then transforming Y can often improve matters (it may be necessary, though, to transform X as well to keep the relationship linear). For example, the number of pollen grains received by flowers of the iris *Lapeirousia anceps*, which is pollinated by long-proboscid flies, increases with increasing flower tube length (Pauw et al. 2009). In a residual plot from a regression using the untransformed data (see the left panel of Figure 17.6-3), the variance of residuals increases from the smallest X-values to larger X-values. This problem goes away when Y is square-root transformed (see the right panel of Figure 17.6-3).

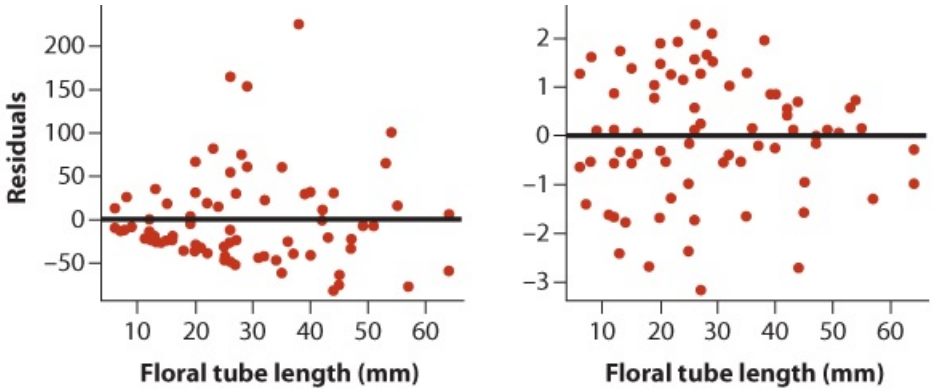


Figure 17.6-3
Whitlock et al., *The Analysis of Biological Data*, 2e,

FIGURE 17.6-3 The effect of a square-root transformation on the residuals from a linear regression of number of pollen grains received on floral tube length of an iris species (Pauw et al. 2009). Residuals from a linear regression calculated on the original data (*left panel*) do not fit the equal-variance assumptions of linear regression, but residuals from a regression using the square root of the number of pollen grains (*right panel*) have more equal variances.

The square-root transformation, originally described in Section 13.3, often resolves unequal variance problems when the data are counts, as in Figure 17.6-3. The log transformation can also be effective when the variance in Y increases with increasing X. Arcsine transformation is often effective when Y is a proportion. When analyzing data that violate the assumptions of

regression, try simple transformations of X and/or Y to see if they help to meet the assumptions of linear regression.

The effects of measurement error on regression

Recall from Section 16.6 that measurement error occurs when a variable is not measured with complete accuracy. Many biological traits, such as behavior or aspects of physiology, can be difficult to measure accurately, so measurement error can be an important component of variation (Section 15.6).

The effects of measurement error on regression differ from the effects on correlation (Section 16.6). The effect depends on the variable. Measurement error in *Y* increases the variance of the residuals, as shown in Figure 17.7-1 when you compare the scatter in the middle panel (measurement error in *Y*) with that in the left panel (no measurement error). This increases the sampling error of the estimate of the slope and of the predictions but has no effect on expected slope.

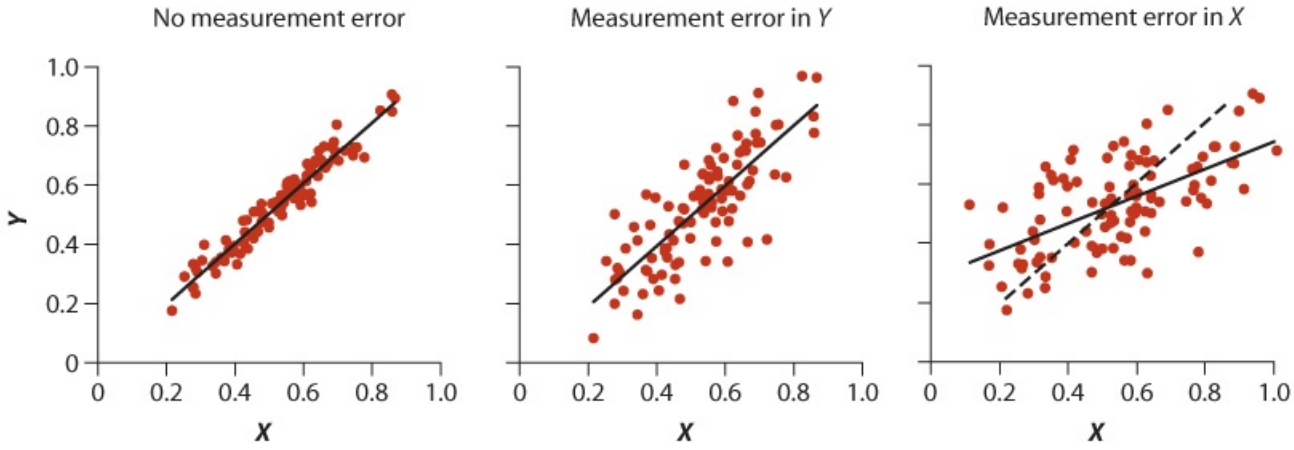


Figure 17.7-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.7-1 The effects of measurement error on the estimate of regression slope. *X* and *Y* are measured without error in the left panel. *Y* is measured with error in the middle panel, which has little effect on expected slope but increases the variability in the residuals. *X* is measured with error in the right panel, which causes the expected estimate of the slope to decline (*solid line*) compared with the slope in the absence of measurement error (*dashed line*). The variability of the residuals also increases.

Measurement error in *X* (the right panel in Figure 17.7-1) also increases the variance of the residuals, and in addition it causes bias in the expected estimate of the slope. With measurement error in *X*, *b* will tend to lie closer to zero on average than the population quantity β. On average, the largest values of *X* in the data will include disproportionately many measurements that were erroneously overestimated. Since the true *X*-values of these points are smaller than their measured values, they predict *Y*-values that on average lie closer to the mean than would the same *X*-values in the absence of measurement error. Conversely, the smallest values of *X* will tend to include disproportionately many measurements that were erroneously underestimated. These underestimated *X*-values will be associated with *Y*-values that lie closer to the mean on average than in the absence of measurement error.

Nonlinear regression

Transformations won't always successfully convert a nonlinear relationship into one that can be analyzed using linear regression. Nonlinear regression methods, however, are readily available in most statistics packages on the computer. Here in Section 17.8, we outline some basic principles for nonlinear regression.

The assumptions of nonlinear regression are almost the same as those of linear regression (Section 17.5), except here we usually assume that the true relationship between X and Y has a specific nonlinear form.

The immediate problem when turning to nonlinear regression is the nearly unlimited number of options. There are so many mathematical functions to choose from that it can be difficult to know where to begin. The appropriate choice depends on the data, but a few guidelines can help you make a sensible choice.

A curve with an asymptote

The best advice we can offer is to keep things simple, unless the data suggest otherwise. Figure 17.8-1 illustrates this principle. The left panel shows a nonlinear regression model fitted to data on the population growth rate of a species of phytoplankton and the concentration of iron, a limiting nutrient. The mathematical function fit to the data passes through every data point. The resulting $MS_{residual}$ is zero.

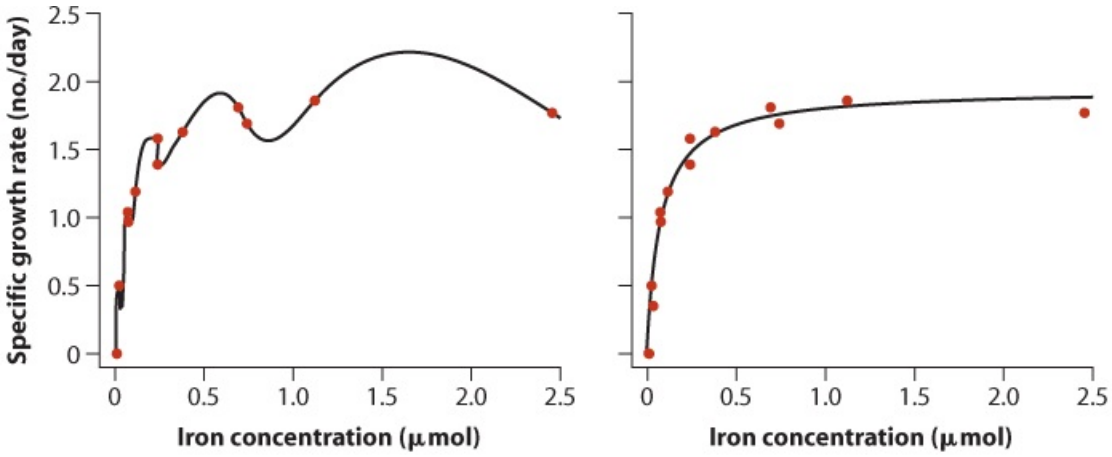


Figure 17.8-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.8-1 Population growth rate of a species of phytoplankton in culture in relation to the concentration of iron in the medium (data from Sunda and Huntsman 1997). The curve in the left panel is an arbitrarily complex function that passes through all of the data points. The curve in the right panel is a Michaelis-Menten curve that fits the data more simply.

This is hardly the best possible outcome, however, even though each data point fits precisely. The problem with the curve in the left panel of Figure 17.8-1 is that it would probably do a terrible job of predicting any *new* observations obtained from the same population, because the curve does not describe the general trend. Such a complicated curve is

also difficult to justify biologically—is there good reason to think that all the peaks and dips in this curve truly reflect the effects of iron on the growth of phytoplankton?

Greater simplicity, as demonstrated by the fitted curve in the right panel of Figure 17.8-1, solves both of these problems. The data are the same as in the left panel, but this time we've fit the much simpler function,

Y=aXb+X.$Y = \frac{aX}{b+X}$.

This is the *Michaelis-Menten* equation used frequently in biochemistry. The curve rises from a *Y*-intercept at zero and increases at a declining rate with increasing *X*, eventually reaching a saturation point, or asymptote. The asymptote is represented in the formula by the constant *a*, whereas *b* determines how fast the curve rises to the asymptote. Reminiscent of linear regression, the Michaelis-Menten equation has only two parameters to estimate (the true asymptote *a* and the true rate parameter *β*). These parameters are very different, however, from those of linear regression. We obtained the curve on the right of Figure 17.8-1 with a statistics package on the computer that used least squares to find the best fit.

The virtue of linear regression is simplicity. We should strive to retain this property as we look at the wide range of nonlinear functions available.

Quadratic curves

The *quadratic* curve is often used in biology to fit a humped-shape curve to data, such as the relationship shown in Figure 17.8-2 between the number of plant species present in ponds (*Y*) and pond productivity (*X*). The curve is a symmetric parabola described by the quadratic (second-degree polynomial) equation,

Y=a+bX+cX2.$Y = a + bX + cX^2$.

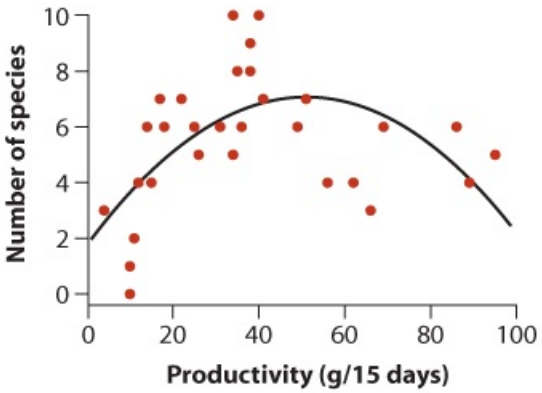


Figure 17.8-2
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 17.8-2 A quadratic curve fit to the relationship between the number of plant species present in ponds and pond productivity (Chase and Leibold 2002).

This equation is similar to the formula for a straight line except that one more term has been added for the square of *X*, and another regression coefficient *c* must be computed. When *c* is negative, the curve is humped, as in Figure 17.8-2. When *c* is positive, the parabola curves upward in a U-shape.

Asymptotic and quadratic curves are just two of several nonlinear functions commonly used in biology. The choice between them must depend on the data: Is the relationship asymptotic or humped? If humped, is the hump symmetric or does it fall more steeply on one

side than the other? If it falls more steeply on one side, then we must search for another function altogether. A good guide to a variety of curves used in biology can be found in Motulsky (1999).

Formula-free curve fitting

With the aid of a computer, it is possible to fit curves to data without specifying a formula. Often called **smoothing,** this approach gathers information from groups of nearby observations to estimate how the mean of Y changes with increasing values of X. There are several methods, including "kernel," "spline," and "loess" smoothing. We bypass the technical details here, but Example 17.8 illustrates the utility of the approach.

EXAMPLE 17.8 The incredible shrinking seal

Trites (1996) amassed a large set of measurements of the ages and sizes of northern fur seals (*Callorhinus ursinus*) in the Pacific Ocean, gathered over decades by many researchers. Most measurements were taken between spring and fall. In summer, females spend a lot of time on land giving birth and nursing young. The graph in Figure 17.8-3 shows the measurements for nearly 10,000 adult females. In a preliminary analysis, a line was fitted to the data, but the relationship appeared nonlinear. Average length increased with age but appeared to taper off by about 4500 days (i.e., about 12 years old).

© Andrew Trites

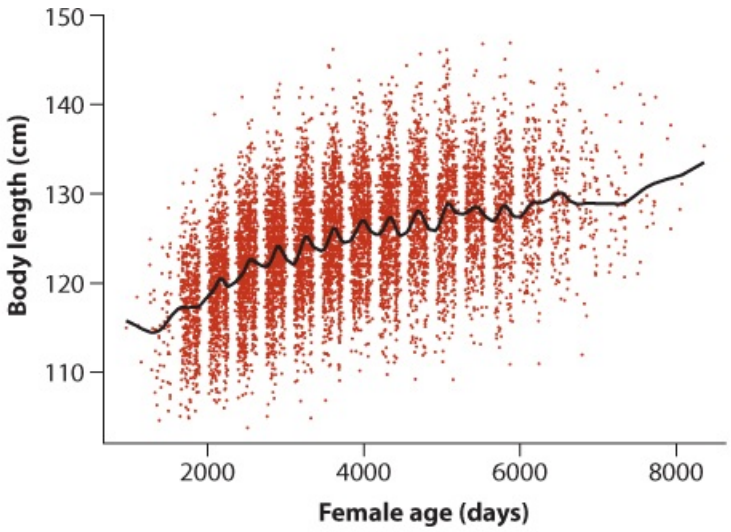


Figure 17.8-3
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 17.8-3 Measurements of body length as a function of age for female fur seals. The spline fit is in black.

To fit the data, we used a spline technique to calculate a smoothed fit of body length on female age. The result is plotted in Figure 17.8-3. Astonishingly, the curve indicates that average female body length does not rise steadily with age, but oscillates each year. Female fur seals become longer each summer and then shorter again by winter (keep in mind that these changes are in length, not weight). Elongation results in part because the seals are heavier on land than in water, and the added weight stretches the skeleton during summer breeding. The skeleton shrinks back after the seals return to the water in the winter. It would have been difficult to come up with a formula that captured this complex relationship.

The seal example illustrates that it is not always easy to anticipate or see the type of relationship present in the data by visual inspection alone. Smoothing techniques can help. Example 17.8 also demonstrates that we don't always need a mathematical formula if all we want to do is fit the data and use it to improve our understanding of a biological system.

The fit obtained by smoothing is controlled by a smoothing coefficient that determines how bumpy the curve is. You can adjust this coefficient in statistical computer programs so that you can explore its effects. A low value for the coefficient results in a bumpy curve that, in the extreme, would pass through all the data points. A larger value of the coefficient gives a smoother fit. Computer programs usually use rules of thumb to choose the best value for the smoothing parameter, but it is wise to try alternatives to see what effect varying the smoothing coefficient has on the curve.

Logistic regression: fitting a binary response variable

Logistic regression is a special type of nonlinear regression developed for a *binary* response variable—that is, when the *Y*-variable measured on independent units is either zero or one. A common use for logistic regression is to fit a dose-response curve, where mortality (or survival) of individuals (the "response") is plotted against the concentration of a drug, toxin, or other chemical (the "dose").[8] Here we provide a quick description with a minimum of calculations.

Our example data set is shown in Figure 17.9-1, from an experiment in which 160 guppies were exposed to cold temperatures (5°C) for different durations (3, 8, 12, or 18 minutes). Each treatment, or dose, was assigned to 40 fish. The study was by Pitkow (1960), who carried out the experiment to identify the physiological mechanism causing fish death at cold temperatures. Mortality is the binary response variable (*Y*) measured on each independent fish (1 = died, 0 = survived), and duration is the explanatory variable. We used logistic regression to fit the curve predicting the probability of fish mortality from duration of exposure (Figure 17.9-1). Because the data are binary, the curve describes the proportion of fish that die (the mean of *Y*) at levels of exposure *X*. Table 17.9-1 gives a frequency table of the data.

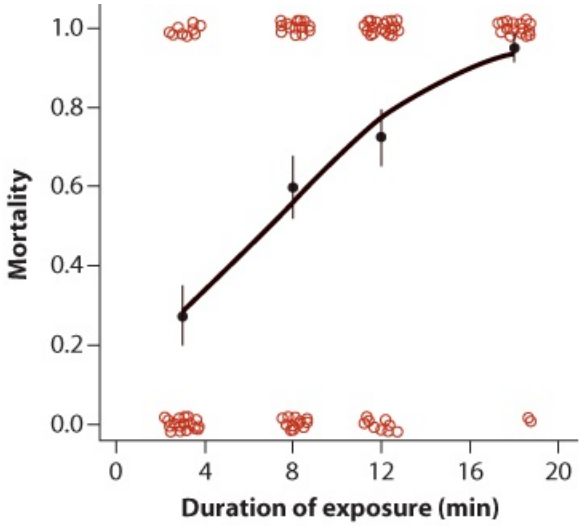


Figure 17.9-1
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 17.9-1 Mortality of guppies in relation to duration of exposure to a temperature of 5°C (data from Pitkow 1960). Treatments were 3, 8, 12, or 18 minutes of exposure, with 40 fish in each of the four treatments. Each point (red circle) indicates a different individual (points were offset using a random perturbation to reduce overlap). *Y* = 1 if the individual died, whereas *Y* = 0 if the individual survived. Black dots indicate the proportion of deaths (±1 SE) in each treatment. The curve is the logistic regression predicting the probability of death.

TABLE 17.9-1 Number of fish (out of 40) in two mortality groups at each of four cold-temperature treatments.

	Duration of exposure (min)			
	3	**8**	**12**	**18**

Died ($Y = 1$)	11	24	29	38
Survived ($Y = 0$)	29	16	11	2

Linear regression of Y on X is unsuitable for these data because the binary response variable violates three of its assumptions. The relationship between Y and X is not linear, because the predicted values $\hat{Y}$ cannot fall outside the interval between 0 and 1. The residuals $Y - \hat{Y}$ are not normally distributed; they are binary—each point is either $0 - \hat{Y}$ or $1 - \hat{Y}$. The variance of the residuals is not constant: variance is expected to be highest when the predicted Y is near 0.5, and lowest when the prediction of Y is close to zero or one. (This is because when the prediction is zero or one, most of the data are zeros or ones (respectively) and the variance among individuals is therefore small.) These problems are not fixed with a simple transformation of the data.

All three problems are solved by logistic regression. The method fits a curve constrained to lie between 0 and 1. Instead of the normal distribution, it assumes that outcomes at every X have a binomial distribution (Section 7.1) being either one or zero. The probability of an event (in this case, dying) is given by the corresponding predicted value on the regression curve. Finally, to correct for differences in the variance of residuals at different values of X, logistic regression weights each residual by its estimated variance obtained from the binomial distribution.

Logistic regression fits the following equation to binary data:

$$\text{log-odds}(Y) = \alpha + \beta X.$$

The log-odds refers to the natural log of the odds of Y (Section 9.2).[9] The right side of the equation $(\alpha + \beta X)$ is the formula for a straight line, with α the intercept and β the slope. In other words, an ordinary line is used to fit the log-odds of the proportion of individuals dying. The curve shown in Figure 17.9-1 is based on estimates of these two parameters: a for intercept and b for slope,

$$\text{log-odds}(Y) = a + bX.$$

Methods to calculate this regression curve from data are available in most computer statistical packages. When we analyzed the guppy data with such a statistics package, we obtained the output in Table 17.9-2, showing estimates of regression coefficients, and Table 17.9-3, showing the results of a test of the null hypothesis of zero slope (H_0: ($\beta = 0$). We will explain the contents of these two tables in turn.

TABLE 17.9-2 Logistic regression output. The values shown are the estimates for a and b of the intercept (α) and the slope (β) of the logistic regression curve for the cold-fish data. SE refers to the standard error of estimates.

	Estimate	SE
Intercept	−1.66	0.41
Slope	0.24	0.04
Number of iterations:	4	

TABLE 17.9-3 Analysis of deviance table, containing results of the log-likelihood ratio test for the cold-fish data.

Model	*df* Deviance	Residual *df*	Residual deviance	*P*

Null			159	209.55	
Duration	1	44.86	158	164.69	2.12×10^{-11}

From Table 17.9-2 we see that the best estimate for the intercept is given by $\alpha = -1.66 \pm 0.41$ SE, and the best estimate for the slope is $b = 0.24 \pm 0.04$ SE.

Computer programs might additionally provide Wald statistics (symbolized by z) and corresponding P-values for approximate tests of the null hypotheses H_0: $\alpha = 0$ and H_0: ($\beta = 0$. However, the Wald method is inaccurate, and we do not present these results. The log-likelihood ratio test (Table 17.9-3) should be used instead.

Remember: the estimates a and b are not intercept and slope of a linear regression of Y on X. Instead, they describe the linear relationship between X and the predicted log odds of Y [which we'll call log-odds($\hat{Y}$)]. To obtain the predicted values ($\hat{Y}$), we need to convert the log-odds to ordinary proportions:

$$\hat{Y} = \frac{e^{\text{log-odds}\left(\hat{Y}\right)}}{1+e^{\text{log-odds}\left(\hat{Y}\right)}}$$

Yˆ=elog-odds(Yˆ)1+elog-odds(Yˆ)

For example, to predict the proportion of fish dying for a cold-temperature duration of 10 minutes, we calculate

log-odds(Yˆ)=a+bX=-1.66+0.24(10)=0.74,Yˆ=e0.741+e0.74=0.68.

$$\text{log-odds}\left(\hat{Y}\right) = a + bX = -1.66 + 0.24\,(10) = 0.74,$$

$$\hat{Y} = \frac{e^{0.74}}{1+e^{0.74}} = 0.68.$$

In other words, a duration of 10 minutes at 5°C is predicted to cause 68% mortality.

Another useful quantity for the regression curve is the *LD50* (lethal dose 50), the estimated dose predicting 50% mortality:

LD50=-ab.$LD\,50 = -\frac{a}{b}.$

For the fish data,

LD50=--1.660.24=6.92 minutes.$LD\,50 = -\frac{-1.66}{0.24} = 6.92\,\text{minutes}.$

Computer output for logistic regression will typically also include the number of iterations used in the calculations (Table 17.9-2). Logistic regression uses maximum likelihood (Chapter 20) to fit the curve to the data, and no formula exists to calculate the estimates. Instead, the computer uses a series of iterations to search for the best-fit curve. The search ceases when there are no further improvements in the fit from successive iterations.

Table 17.9-3 is an *analysis of deviance* table, with the results of a log-likelihood ratio test (see Section 20.5) of the null hypothesis that there is no relationship between mortality and duration (H_0: $\beta = 0$ vs. H_A: $\beta \neq 0$). The analysis of deviance table is analogous to the ANOVA table in ordinary linear regression. The method fits two models to the data, one in which the variable duration is absent, and one in which it is present.

Null model:log-odds(Y)=αRegression model:log-odds(Y)=α+βX.

$$\text{Null model:} \quad \text{log-odds}\,(Y) = \alpha$$

$$\text{Regression model:} \quad \text{log-odds}\,(Y) = \alpha + \beta X.$$

The null model is a restatement of the null hypothesis that $\beta = 0$, whereas the regression model is a restatement of the alternative hypothesis that $\beta \neq 0$. Analysis of deviance compares the fit of the two models to the data. If the improvement in fit of the regression model over the null model is too large to be explained by chance, the null hypothesis is rejected.

The key quantity in the table is the *improvement* in fit when $\beta \neq 0$, which is here labeled simply “Deviance.” This quantity is the difference in fit between the two models, where “fit” is the discrepancy between the observed values of Y and the values predicted, $\hat{Y}$, from each model. For the fish data, the improvement in fit is 209.55 — 164.69 = 44.86. Residual deviance is analogous to residual mean square in ordinary linear regression.

Under the null hypothesis that $\beta = 0$, deviance has a χ^2 distribution with 1 *df*. The critical value for $\chi_{0.05,12}=3.84.$ $\chi^2_{0.05,1} = 3.84.$ Since 44.86 > 3.84, $P < 0.05$, and we reject the null hypothesis. An approximation to the exact *P*-value is given in the table.

0 Summary

- Regression is a method used to predict the value of a numerical response variable Y from the value of a numerical explanatory variable X.
- Linear regression fits a straight line through a scatter of points. The equation for the regression line is $Y = a + bX$, where b is the slope of the line and a is the intercept.
- The least squares regression line is found by minimizing the sum of the squared differences between the observed Y-values and the values predicted by the line.
- The residuals are the differences between the observed values of Y and the values predicted by the least-squares regression line, Y^.$\hat{Y}$. The variance of the residuals, $MS_{residual}$, measures the spread of points above and below the regression line.
- Linear regression calculated on a sample of points estimates the straight-line relationship between the two variables in the population. The formula for the population regression line is $Y = \alpha + \beta X$, where β is the slope of the line and a is the intercept.
- If the assumptions of regression are met, then the sampling distribution of b is normal with mean β and standard deviation estimated by the standard error of the estimate of the slope, SE_b.
- The confidence interval for the slope β is based on the t-distribution.
- There are two types of prediction in regression: the mean Y at a given X, and a single Y observation at X. Both predictions generate the same value Y^,$\hat{Y}$, but they have very different precision. Precision is lower when predicting an individual Y because it includes the variability between individuals having the same X-value.
- Confidence intervals for predicted mean Y-values at each X are represented by confidence bands. Analogous intervals for the predicted Y-values of a single individual are called prediction intervals.
- Extrapolation is the prediction of Y at values of X beyond the range of X-values in the sample. Extrapolation is problematic, though, because there is no way to ensure that the relationship between X and Y continues to be linear beyond the data.
- If the null hypothesis is correct that the slope b of a population regression line is zero, then the test statistic $t = b/SE_b$ has a t-distribution with $n - 2$ degrees of freedom.
- An ANOVA table and F-test can also be used to test the null hypothesis that the population slope $\beta = 0$.
- R^2 is a measure of the fit of a regression line to the data. It measures the “fraction of the variation in Y that is explained by X.”
- Regression toward the mean results when two imperfectly correlated variables are compared by regression. Individuals that are far from the mean for one of the measurements will on average lie closer to the mean for the other measurement.
- Methods for regression assume that the relationship between X and Y falls along a straight line, that the Y-measurements at each value of X are a random sample from a population of Y-values, that the distribution of Y-values at each value of X is normal, and that the

variance of Y-values is the same at all values of X.

- The scatter plot and the residual plot are graphical devices for detecting departures from the assumptions of linear regression.
- Transformations of X and/or Y can be used to render a nonlinear relationship linear and to correct violations of the assumption of equal variance of residuals at every X.
- The log-transformation is the most versatile transformation. It is useful when Y is related to X by a power function or by an exponential function.
- If transformations do not work, nonlinear regression is an option.
- Nonlinear regression curves should be kept as simple as the data warrant. Overly complex curves may be biologically unjustified and have low predictive power.
- Smoothing methods make it possible to fit nonlinear curves to data without specifying a formula.
- Logistic regression allows us to use a numerical variable to predict the probability that an individual has a particular value of a binary response variable.

1 Quick Formula Summary

Shortcuts

Sum of Products: ∑i(Xi−X¯)(Yi−Y¯)=∑i(XiYi)−(∑iXi)(∑iYi)n.

$$\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_i (X_i Y_i) - \frac{\left(\sum_i X_i\right)\left(\sum_i Y_i\right)}{n}.$$

Sum of squares for X: ∑i(Xi−X¯)2=∑i(Xi2)−(∑iXi)2n. $\sum_i (X_i - \overline{X})^2 = \sum_i (X_i^2) - \frac{\left(\sum_i X_i\right)^2}{n}.$

Regression slope

What is it for? Estimating the slope of the linear equation $Y = \alpha + \beta X$ between an explanatory variable X and a response variable Y.

What does it assume? The relationship between X and Y is linear; each Y-measurement at a given X is a random sample from a population of Y-measurements; the distribution of Y-values at each value of X is normal; and the variance of Y-values is the same at all values of X.

Parameter: β

Estimate: b

Formula: b=∑i(Xi−X¯)(Yi−Y¯)∑i(Xi−X¯)2 . $b = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2}.$

Standard error: SEb=MSresidual∑i(Xi−X¯)2, $\mathrm{SE}_b = \sqrt{\frac{\mathrm{MS}_{\text{residual}}}{\sum_i (X_i - \overline{X})^2}},$

where $\mathrm{MS}_{\text{residual}}$ is the mean squared residual (the estimated variance of the residuals);

MSresidual=∑i(Yi-Yˆi)2n-2. $\mathrm{MS}_{\text{residual}} = \frac{\sum_i \left(Y_i - \hat{Y}_i\right)^2}{n-2}.$

A quicker formula for $\mathrm{MS}_{\text{residual}}$, not requiring you to calculate the Y^$\hat{Y}$ first, is

MSresidual=∑i(Y-Y¯)2-b∑i(Xi-X¯)(Y-Y¯)n-2. $\mathrm{MS}_{\text{residual}} = \frac{\sum_i (Y - \overline{Y})^2 - b\sum_i (X_i - \overline{X})(Y - \overline{Y})}{n-2}.$

Regression intercept

What is it for? Estimating the intercept of the linear equation $Y = \alpha + \beta X$.

What does it assume? Same as the assumptions for the regression slope.

Parameter: α

Estimate: α

Formula: $a=\bar{Y}-b\bar{X}$. $a = \bar{Y} - b\bar{X}.$

Standard error: Set $X = 0$ in the formula for the standard error of the predicted mean Y at a given X [see "Confidence interval for the predicted mean Y at a given X (confidence bands)" on p. 574]. This is valid only if the value $X = 0$ falls within the range of X-values in the data.

Confidence interval for the regression slope

What is it for? An interval estimate of the population slope.

What does it assume? Same as the assumptions for the regression slope.

Statistic: b

Parameter: b

Formula: $b-t_{\alpha(2),df}SE_b<\beta<b+t_{\alpha(2),df}SE_b$, $b - t_{\alpha(2),df}\mathrm{SE}_b < \beta < b + t_{\alpha(2),df}\mathrm{SE}_b,$ where SE_b is the standard error of the slope (see the formula given previously with the regression slope), and $t_{\alpha(2),n-2}$ is the two-tailed critical value of the t-distribution having $df = n - 2$.

Confidence interval for the predicted mean *Y* at a given *X* (confidence bands)

What is it for? An interval estimate of the predicted *mean Y* of all individuals in the population having the given value of X.

What does it assume? Same as the assumptions for the regression slope.

Statistic: $\hat{Y}$

Formula: $\hat{Y}-t_{\alpha(2),n-2}SE[\hat{Y}]<\text{predicted}Y<\hat{Y}+t_{\alpha(2),n-2}SE[\hat{Y}]$, $\hat{Y} - t_{\alpha(2),n-2}\mathrm{SE}\left[\hat{Y}\right] < \text{predicted}Y < \hat{Y} + t_{\alpha(2),n-2}\mathrm{SE}\left[\hat{Y}\right],$

where $SE[\hat{Y}]=MS_{residual}\left(\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_i (X_i-\bar{X})^2}\right)$ $\mathrm{SE}\left[\hat{Y}\right] = \sqrt{\mathrm{MS}_{\text{residual}}\left(\frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum_i (X_i-\bar{X})^2}\right)}$ is the standard error of the

prediction, $MS_{residual}$ is the mean square residual (see the formula given previously under "Regression slope"), and $t_{\alpha(2)n-2}$ is the two-tailed critical value of the t-distribution having $df = n - 2$.

Confidence interval for the predicted individual *Y* at a given *X* (prediction intervals)

What is it for? An interval estimate of the predicted Y for a *single individual* having the given value of X.

What does it assume? Same as the assumptions for the regression slope.

Statistic: $\hat{Y}$

Formula: Yˆ−tα(2),n−2SE1[Yˆ]<predictedY<Yˆ+tα(2),n−2SE[Yˆ], $\hat{Y} - t_{\alpha(2),n-2}\mathrm{SE}_1\left[\hat{Y}\right] < \text{predicted}Y < \hat{Y} + t_{\alpha(2),n-2}\mathrm{SE}\left[\hat{Y}\right]$,

where SE1[Yˆ]=MSresidual(1+1n+(Xi-X¯)2∑i(Xi-X¯)2) $\mathrm{SE}_1\left[\hat{Y}\right] = \sqrt{\mathrm{MS}_{\text{residual}}\left(1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right)}$ is the standard error of

the prediction, $\mathrm{MS}_{\text{residual}}$ is the mean square residual (see the formula given previously under "Regression slope"), and $t_{\alpha(2)n-2}$ is the two-tailed critical value of the *t*-distribution having $df = n - 2$.

The *t*-test of a regression slope

What is it for? To test the null hypothesis that the population parameter β equals a null hypothesized value β_0.

What does it assume? Same as the assumptions for the regression slope.

Test statistic: *t*

Distribution under H_0: *t*-distributed with $n - 2$ degrees of freedom.

Formula: t=b-β0SEb $t = \frac{b - \beta_0}{\mathrm{SE}_b}$ is the standard error of *b* (see the formula given previously under "Regression slope").

The ANOVA method for testing zero slope

What is it for? To test the null hypothesis that the slope β equals zero, and to partition sources of variation.

What does it assume? Same as the assumptions for the regression slope.

Test statistic: *F*

Distribution under H_0: *F* distribution. *F* is compared with $F_{\alpha(1),1,n-2}$.

Source of variation	Sum of squares	*df*	Mean squares
Regression	SSregression=∑i(Yˆi-Yˆ)2 $\mathrm{SS}_{\text{regression}} = \sum_i \left(\hat{Y}_i - \hat{Y}\right)^2$	1	SSregressiondfregression $\frac{\mathrm{SS}_{\text{regression}}}{df_{\text{regression}}}$
Residual	SSresidual=∑i(Yi-Yˆi)2 $\mathrm{SS}_{\text{residual}} = \sum_i \left(Y_i - \hat{Y}_i\right)^2$	$n - 2$	SSresidualdfresidual $\frac{\mathrm{SS}_{\text{residual}}}{df_{\text{residual}}}$
Total	SStotal=∑i(Yi-Y¯)2 $\mathrm{SS}_{\text{total}} = \sum_i \left(Y_i - \overline{Y}\right)^2$	$n - 1$	

R squared (R^2)

What is it for? Measuring the fraction of the variation in *Y* that is “explained” by *X*.

Formula: R2=SSregressionSStotal, $R^2 = \frac{\mathrm{SS}_{\mathrm{regression}}}{\mathrm{SS}_{\mathrm{total}}}$,

where $\mathrm{SS}_{\mathrm{regression}}$ is the sum of squares for regression and $\mathrm{SS}_{\mathrm{total}}$ is the total sum of squares.

PRACTICE PROBLEMS

1. **Calculation problem: Regression lines.** Men's faces have higher width-to-height ratios than women's, on average. This turns out to reflect a difference in testosterone expression during puberty. Testosterone is also known to predict aggressive behavior. Does face shape predict aggression? To test this, Carré and McCormick (2008) compared the face width-to-height ratio of 21 university hockey players with the average number of penalty minutes awarded per game for aggressive infractions like fighting or cross-checking. Their data are below along with some partial calculations. We will calculate the equation for the line that best predicts penalty minutes from face width-to-height ratio.

Face width: height ratio (*X*)	Penalty minutes per game (*Y*)
1.59	0.44
1.67	1.43
1.65	1.57
1.72	0.14
1.79	0.27
1.77	0.35
1.74	0.85
1.74	1.13
1.77	1.47
1.78	1.51
1.76	1.99
1.81	1.06
1.83	1.20
1.83	1.23
1.84	0.80
1.87	2.53
1.92	1.23
1.95	1.10
1.98	1.61
1.99	1.95
2.07	2.95

a. Plot the data in a scatter plot.

b. Examine the graph. Based on this graph, do the assumptions of linear regression appear to be met?

c. Calculate the means of the two variables. (While you're doing so, record the sum of all X-values and the sum of all Y-values.)

d. Calculate the sum of X^2, the sum of Y^2, and the sum of the product XY.

e. Calculate the sum of products (∑i(Xi-X¯)(Y-Y¯)) $\left(\sum_i (X_i - \overline{X})(Y - \overline{Y})\right)$ and the sum of squares (∑i(Xi-X¯)2) $\left(\sum_i (X_i - \overline{X})^2\right)$ of the explanatory variable, face ratio.

f. How steeply does the number of penalty minutes increase per unit increase in face ratio? From the sum of products and sum of squares for face ratio, calculate the estimate *b* of the slope. Double-check that the sign of the slope matches your impression from the scatter plot.

g. Calculate the estimate of the intercept, *a*, from the variable means and *b*.

h. Write the result in the form of an equation for the line. Add your line to the graph in (a).

2. **Calculation problem: Standard error and confidence intervals of the slope.** How uncertain is our estimate of slope? Using the face ratio and hockey aggressive penalty data from Practice Problem 1, calculate the standard error and confidence interval of the slope of the linear regression.

a. Calculate the total sum of squares for the response variable, penalty minutes.

b. Calculate the residual mean square $MS_{residual}$, using the total sum of square for *Y*, the sum of products, and the slope *b*.

c. With the sum of squares for *X* and $MS_{residual}$, calculate the standard error of *b*.

d. How many degrees of freedom does this analysis of the slope have?

e. Find the two-tailed critical *t*-value for a 95% confidence interval ($\alpha = 0.05$) for the appropriate *df*.

f. Calculate the confidence interval of the population slope, β.

3. **Calculation problem: Testing the null hypothesis that the slope equals zero.** Can the relationship be explained by chance? Using the face ratio and hockey penalty data from Practice Problem 1, test the null hypothesis that the slope of the regression line is zero.

a. State the null and alternate hypotheses.

b. What is β_0 for this null hypothesis?

c. Calculate the test statistic *t* from *b*, β_0, and the standard error of *b*.

d. Find the critical value of *t* appropriate for the degrees of freedom, at $\alpha = 0.05$.

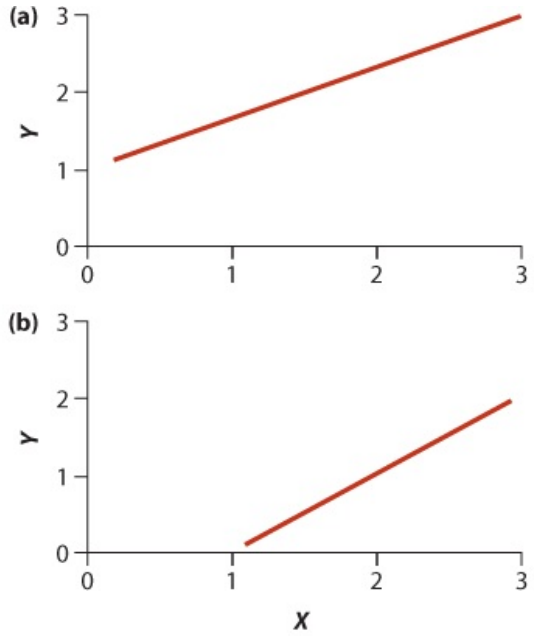
e. Is the absolute value of the *t* for this test greater than the critical value?

f. Using a computer or the statistical tables, be as precise as possible about the *P*-value for this test. Draw conclusions from the test.

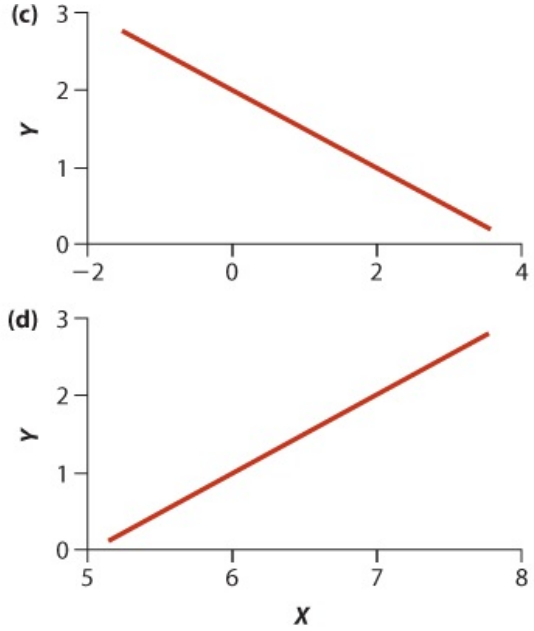
g. What fraction of the variation in average penalty minutes per game is accounted for by face ratio? Calculate the value of R^2.

4. Golomb et al. (2012) looked at whether higher chocolate consumption predicts higher body mass in humans. They fitted the data using a linear regression having chocolate consumption (number of times consumed per week) as the explanatory variable and the body mass index (BMI) as the response variable. BMI measures body mass relative to height, with high BMI typically meaning an overweight person. The slope of the regression was − 0.142, with a standard error of 0.053 and a *P*-value of 0.008. Is this evidence that people who eat more chocolate have higher BMI? Why or why not?

5. What is the formula for each of the following four regression lines?



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

6. Some species seem to thrive in captivity, whereas others are prone to health and behavior difficulties when caged. Maternal care problems in some captive species, for example, lead to high infant mortality. Can these differences be predicted? The following data are measurements of the infant mortality (percentage of births) of 20 carnivore species in captivity along with the log (base-10) of the minimal home-range sizes (in km^2) of the same species in the wild (Clubb and Mason 2003).

Log_{10} home-range size	Captive infant mortality (%)
−1.3	4
−0.5	22
−0.3	0
0.2	0
0.1	11

0.5	13
1.0	17
0.3	25
0.4	24
0.5	27
0.1	29
0.2	33
0.4	33
1.3	42
1.2	33
1.4	20
1.6	19
1.6	25
1.8	25
3.1	65

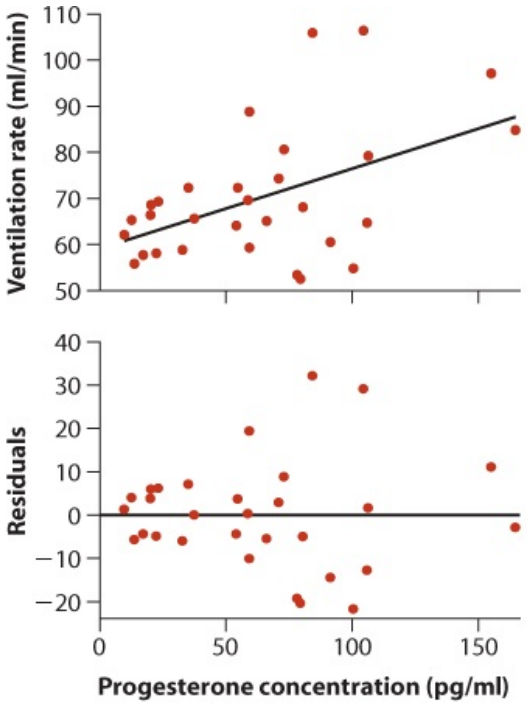
a. Draw a scatter plot of these data, with the log of home-range size as the explanatory variable. Describe the relationship between the two variables in words.

b. Estimate the slope and intercept of the least-squares regression line, with the log of home-range size as the explanatory variable. Add this line to your plot.

c. Does home-range size in the wild predict the mortality of captive carnivores? Carry out a formal test. Assume that the species data are independent.[10]

d. Outliers should be investigated because they might have a substantial effect on the estimates of the slope and intercept. Recalculate the slope and intercept of the regression line from part (c) after excluding the outlier at large home-range size (which corresponds to the polar bear). Add the new line to your plot. By how much did it change the slope?

7. The following two graphs display data gathered to test whether the exercise performance of women at high elevations depends on the stage of their menstrual cycle (Brutsaert et al. 2002). In the upper panel, the explanatory variable is the progesterone level and the response variable is the ventilation rate at submaximal exercise levels. The line is the least-squares regression. The lower panel is the corresponding residual plot.



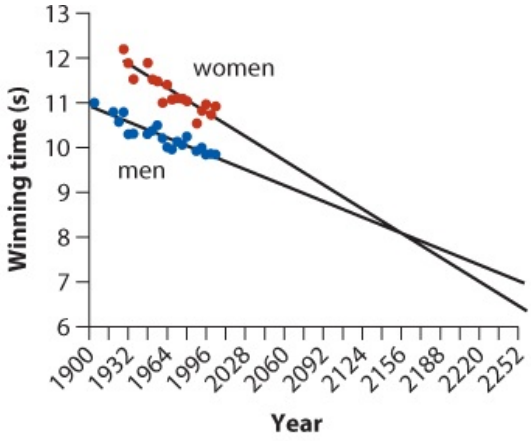
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

a. What is a "least-squares" regression line?

b. What are residuals?

c. Assume that the random sampling assumption is met. By viewing these plots, assess whether each of the three other main assumptions of linear regression is likely to be met in this study.

3. The slopes of the regression lines on the following graph show that the winning Olympic 100-m sprint times for men and women have been getting shorter and shorter over the years, with a steeper trend in women than in men (the graph is modified from Tatem et al. 2004). If trends continue, women are predicted to have a shorter winning time than men by the year 2156. What cautions should be applied to this conclusion? Explain.



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

4. In an analysis of the performance of Major League Baseball players, Schaal and Smith (2000) found that the batting scores of the top 10 players in the 1997 baseball season dropped on average in 1998. What is the best interpretation of this finding?

a. Players who did well in 1997 reduced their effort the following year, realizing that they didn't need to work as hard to get an above-average result.

b. Players performing above average in 1997 were older and more worn out by 1998.

c. Regression toward the mean.

d. Possibly (a) and (b), but (c) is likely and cannot be ruled out.

). Hybrid offspring of parents of different species are often sterile. How different must the parent species be, genetically, to produce this effect? The accompanying table (Moyle et al. 2004) lists the proportion of pollen grains that are sterile in hybrid offspring of crosses between pairs of species of *Silene* (bladder campions—see the photo on the first page of this chapter). Also listed is the genetic difference between the pair of species, as measured by DNA sequence divergence. Assume that different species pairs are independent.

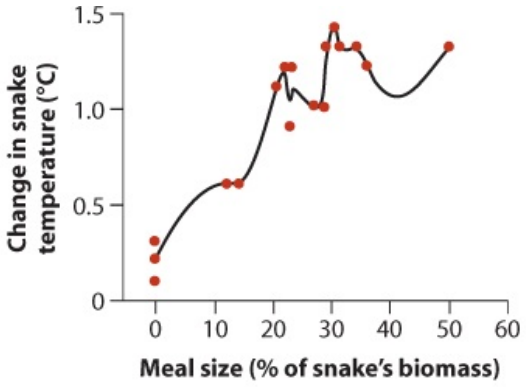
***Silene* species pair**	**Genetic distance**	**Proportion of pollen that is sterile**
1	0.00	0.02
2	0.00	0.06
3	0.00	0.14
4	0.00	0.24
5	0.00	0.30
6	0.03	0.62
7	0.02	0.28
8	0.03	0.23
9	0.04	0.15
10	0.04	0.45
11	0.05	0.84
12	0.11	0.65
13	0.12	0.77
14	0.12	1.00
15	0.13	1.00
16	0.13	0.93
17	0.13	0.91
18	0.14	0.93
19	0.13	0.96
20	0.13	1.00
21	0.15	1.00
22	0.16	0.97
23	0.18	1.00

a. We would like to predict the proportion of hybrid pollen that is sterile (Y) from the genetic distance between the species (X). Since the response variable is a proportion, what transformation would be your first choice to help meet the assumptions of linear regression?

b. Transform the proportions and then produce a scatter plot of the data. Estimate and draw the regression line.

c. Calculate the 95% confidence interval for the slope of the line.

l. Rattlesnakes often eat large meals that require significant increases in metabolism for efficient digestion. Snakes are known to adjust their thermoregulatory behavior after feeding, seeking out warmer spots to increase their metabolic rates. Can snakes increase body temperature, though, even without this behavior? Tattersall et al. (2004) measured the change in body temperature of snakes after meals of various sizes, and we have used their data in an inappropriate way in the following graph, fitting a nonlinear mathematical function indicated by the curve.



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

a. Why is the nonlinear fit shown inappropriate?

b. What alternative procedure would you recommend to achieve the goal of predicting snake body-temperature change from meal size?

12. Male lizards in the species *Crotaphytus collaris* use their jaws as weapons during territorial interactions. Lappin and Husak (2005) tested whether weapon performance (bite force) predicted territory size in this species. Their measurements for both variables are listed in the following table for 11 males.

Bite force (N)	Territory area (m^2)
28.2	437
33.9	589
29.5	871
39.8	977
41.7	1288
44.7	2138
46.8	2455
47.9	3548
36.3	2692
35.5	2042
33.9	3020

Matt Jeppson/Shutterstock

a. How rapidly does territory size increase with bite force? Estimate the slope of the regression line. Provide a standard error for your estimate.

b. How uncertain is our estimate of slope? Provide a 99% confidence interval for //.

c. Provide an interpretation for the 99% confidence interval in part (b). What does it measure?

d. Bite force is difficult to measure accurately, and so the values shown probably include some measurement error. Is the slope of the true regression line most likely to be underestimated, overestimated, or unaffected as a result?

e. Territory area is difficult to measure accurately, so the values shown probably include some measurement error. Is the slope of the true regression line most likely to be underestimated, overestimated, or unaffected as a result?

13. An ANOVA carried out to test the null hypothesis of zero slope for the regression of lizard territory area on bite force (see Practice Problem 12) yielded the following results.

Source of variation	Sum of squares	*df*	Mean squares	*F*-ratio
Regression	3758539	*1*		
Residual	7303662	9		
Total				

a. Complete the ANOVA table.

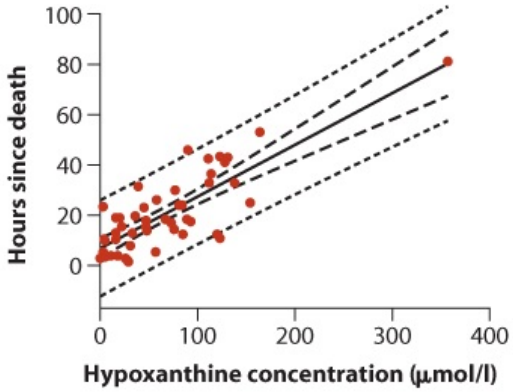
b. Using the *F*-statistic, test the null hypothesis of zero slope at the significance level $\alpha = 0.05$.

c. What are your assumptions in part (b)?

d. What does the $MS_{residual}$ measure?

e. Calculate the R^2 statistic. What does it measure?

14. James et al. (1997) demonstrated that the chemical hypoxanthine in the vitreous humour (the colorless jelly filling the eye) shows a postmortem linear increase in concentration with time since death. This suggests that hypox-anthine concentration might be useful in predicting time of death when it is unknown. The following graph shows measurements collected by the researchers on 48 subjects whose time of death was known. The regression line, the 95% confidence bands, and the 95% prediction interval are included on the graph.



Whitlock et al., *The Analysis of Biological Data*, 2e,

a. The data set depicted in the graph includes one conspicuous outlier on the far right. If you were advising the forensic scientists who gathered these data, how would you suggest they handle the outlier?

b. What do the confidence bands measure?

c. Are the inner dashed lines the confidence bands or the prediction interval?

d. If the regression depicted in the graph was to be used to predict the time of death in a murder case, which bands would provide the most relevant measure of uncertainty, the confidence bands or the prediction interval? Why?

15. Social spiders live together in kin groups that build communal webs and cooperate in gathering prey. The following web measurements were gathered on 17 colonies of the social spider *Cyrtophora citricola* in Gabon (Rypstra 1979).

Colony	Height of web aboveground (cm)	Number of spiders
1	90	17
2	150	32
3	270	96
4	320	195
5	180	372
6	380	135
7	200	83
8	120	36
9	240	85
10	120	20
11	210	82
12	250	95
13	140	59
14	300	89
15	290	152
16	180	62
17	280	64

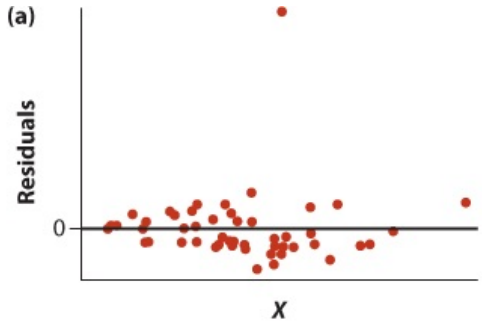
a. Use these data to draw a scatter plot of the relationship between the colony height

aboveground (explanatory variable) and the number of spiders in the colony (response variable).

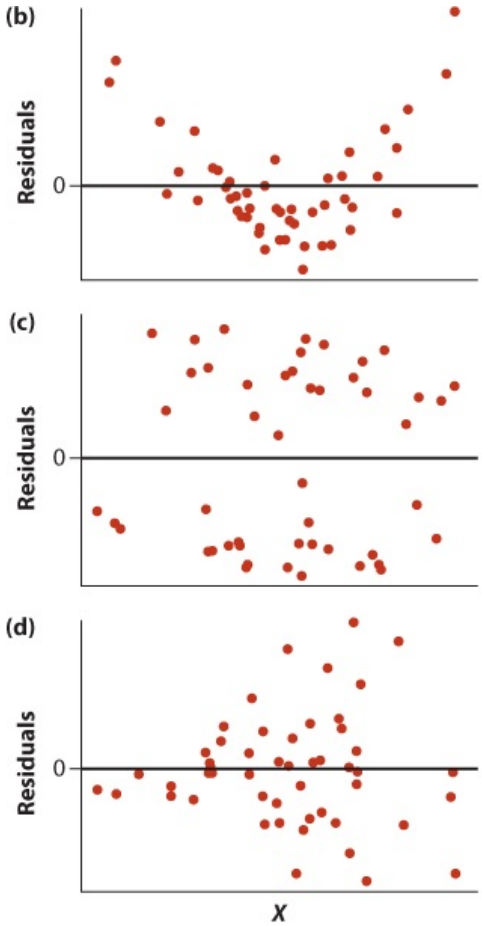
b. Examine the scatter plot and determine any impediments that might make it difficult to use linear regression to predict number of spiders in a colony from colony height.

c. In view of what you discerned in part (b), what method would you recommend to test whether colony height predicts the number of spiders?

16. Identify the assumption(s) of linear regression that is (are) violated in each of the following residual plots.



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

17. The forests of the northern United States and Canada have no native terrestrial earthworms, but many exotic species (including those used as bait when fishing) have been introduced. These immigrant species are dramatically changing the soil. The following data were gathered to predict the nitrogen content of mineral soils of 39 hardwood forest plots in Michigan's Upper Peninsula from the number of earthworm species found in those plots

(Gundale et al. 2005).

Earthworm species	Nitrogen content (%)
0	0.22, 0.19, 0.16, 0.08, 0.05
1	0.33, 0.30, 0.26, 0.24, 0.20, 0.18, 0.14, 0.13, 0.11, 0.09, 0.08
2	0.27, 0.24, 0.23, 0.18, 0.16, 0.13
3	0.32, 0.32, 0.29, 0.24, 0.22, 0.12, 0.40
4	0.34, 0.33, 0.23, 0.21, 0.18, 0.17, 0.15, 0.14
5	0.20, 0.54

a. Draw a scatter plot of these data, using the number of earthworm species as the explanatory variable.

b. Using the following intermediate calculations, calculate the regression line to predict the total nitrogen content of the soil from the number of earthworm species present. Add the line to your plot.

$X¯=2.205Y¯=0.215∑i(Xi-X¯)2=86.359∑i(Yi-Y¯)2=0.366∑i(Xi-X¯)(Yi-Y¯)=2.453.$

$\overline{X} = 2.205 \quad \overline{Y} = 0.215$

$$\sum_i (X_i - \overline{X})^2 = 86.359$$

$$\sum_i (Y_i - \overline{Y})^2 = 0.366$$

$$\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = 2.453.$$

c. What are the units of your estimate of slope, *b*?

d. What is the predicted nitrogen content of soil having five earthworm species?

e. Calculate a standard error of the slope.

f. Produce a 95% confidence interval for the slope.

3. Is the scaling of respiratory metabolism to body size in plants similar to that found in animals, where an approximate 3/4-power relation seems to hold? The data below are measurements of aboveground mass (in g) and respiration rate (in nmol/s) in 10 individuals of Japanese cypress trees (*Chamaecyparis obtusa*). They were obtained from a larger data set amassed by Reich et al. (2006). Respiratory metabolism (Y) is expected to depend on body mass (X) by the power law, $Y = \alpha X^{\beta}$, where β is the scaling exponent.

Aboveground mass (g)	Respiration rate (nmol/s)
453	666
1283	643
695	1512
1640	2198
1207	2535
2096	4176
2804	3196
3528	3494

5940	7386
10,000	10,363

a. Use linear regression to estimate p for Japanese cypress. Include a standard error of your estimate.

b. Plot your line and the data in a scatter plot.

c. Use the 95% confidence interval to determine the range of most-plausible values for β based on these data. Does this range include the value 3/4?

d. Carry out a formal test of the null hypothesis that $\beta = 3/4$.

e. It is a challenge to estimate mass and respiration rate of a living tree in the field, and both measurements are likely subject to measurement error. How is measurement error in each of these two traits likely to affect the estimate of the exponent?

ASSIGNMENT PROBLEMS

). You might think that increasing the resources available would elevate the number of plant species that an area could support, but the evidence suggests otherwise. The data in the accompanying table are from the Park Grass Experiment at Rothamsted Experimental Station in the U.K., where grassland field plots have been fertilized annually for the past 150 years (collated by Harpole and Tilman 2007). The number of plant species recorded in 10 plots is given in response to the number of different nutrient types added in the fertilizer treatment (nutrient types include nitrogen, phosphorus, potassium, and so on).

Plot	Number of nutrients added	Number of plant species
1	0	36
2	0	36
3	0	32
4	1	34
5	2	33
6	3	30
7	1	20
8	3	23
9	4	21
10	4	16

a. Draw a scatter plot of these data. Which variable should be the explanatory variable (X), and which should be the response variable (Y)?

b. What is the rate of change in the number of plant species supported per nutrient type added? Provide a standard error for your estimate.

c. Add the least-squares regression line to your scatter plot. What fraction of the variation in the number of plant species is "explained" by the number of nutrients added?

d. Test the null hypothesis of no treatment effect on the number of plant species.

). Heusner (1991) assembled the following data on the mass and basal metabolic rate of 17 species of primates, including the potto shown in the accompanying photo.

Species	Mass (g)	Basal metabolic rate (watts)
Alouatta palliata	4670.0	11.6
Aotus trivirgatus	1020.0	2.6
Arctocebus calabarensis	206.0	0.7
Callithrix jachus	190.0	0.9
Cebuella pygmaea	105.0	0.6
Cheirogaleus medius	300.0	1.1
Euoticus elegantilus	261.5	1.2

Galago crassicaudatus	1039.0	2.9
Galago demidovii	61.0	0.4
Galago elegantulus	261.5	1.2
Homo sapiens	70,000.0	82.8
Lemur fulvus	2330.0	4.2
Nycticebus coucang	1300.0	1.7
Papio anubis	9500.0	16.0
Perodicticus potto	1011.0	2.1
Saguinus geoffroyi	225.0	1.3
Saimiri sciureus	800.0	4.4

rod williams/Alamy

Previous research has indicated that basal metabolic rate (R) of mammal species depends on body mass (M) in the following way: $R = \alpha M^{\beta}$, where α and β are constants.

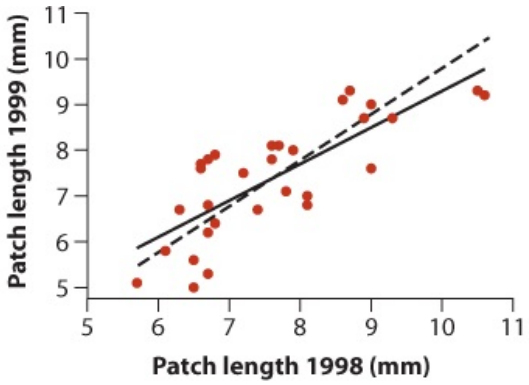
a. Use linear regression to estimate β for primates. Call your estimate b.

b. Plot your line and the data in a scatter plot.

c. How precise is the estimate of β? Provide a standard error for b and a 95% confidence interval for β. Assume that the species data are independent.

l. Previous evidence and some theory predict that the exponent β describing the relationship between metabolic rate and mass should equal 3/4. Using the data from Assignment Problem 20, test whether the exponent differs from the expected value of 3/4.

2. The white forehead patch of the male collared flycatcher is important in mate attraction. Griffith and Sheldon (2001) found that the length of the patch varied from year to year. They measured the forehead patch on a sample of 30 males in two consecutive years, 1998 and 1999, on the Swedish island of Gotland. The scatter plot provided gives the pair of measurements for each male. The solid regression line predicts the 1999 measurement from the 1998 measurement. The dashed line is drawn through the means for 1998 and 1999, but it has a slope of one. The difference between the two lines indicates that males with the longest patches in 1998 had smaller patches in 1999, relative to the other birds. Similarly, the males with the smallest patches in 1998 had larger patches, on average, in 1999, relative to other birds.



Whitlock et al., *The Analysis of Biological Data*, 2e,

a. The following table summarizes the data. Use these numbers to calculate the regression slope.

	Mean	Sum of squares	Sum of products
Patch length 1998	7.62	45.43	36.26
Patch length 1999	7.40	47.03	

b. Now let the patch length in 1998 be the response variable (Y). Use the patch length in 1999 to predict patch length in 1998. What is the slope of this new regression?

c. What is the most likely reason that the slope is less than one in both regressions?

3. Seedlings of understory trees in mature tropical rainforests must survive and grow using intermittent flecks of sunlight. How does the length of exposure to these flecks of sunlight (fleck duration) affect growth? Leakey et al. (2005) experimentally irradiated seedlings of the Southeast Asian rainforest tree *Shorea leprosula* with flecks of light of varying duration while maintaining the same total irradiance to all the seedlings. Their data for 21 seedlings are listed in the following table.

Tree	Mean fleck duration (min)	Relative growth rate (mm/mm/week)
1	3.4	0.013
2	3.2	0.008
3	3.0	0.007
4	2.7	0.005
5	2.8	0.003
6	3.2	0.003
7	2.2	0.005
8	2.2	0.003
9	2.4	0.000
10	4.4	0.009
11	5.1	0.010
12	6.3	0.009
13	7.3	0.009
14	6.0	0.016
15	5.9	0.025

16	7.1	0.021
17	8.8	0.024
18	7.4	0.019
19	7.5	0.016
20	7.5	0.014
21	7.9	0.014

$\bar{X}=5.062\ \bar{Y}=0.0111\ \sum_i(Xi-\bar{X})2=100.210\ \sum_i(Yi-\bar{Y})2=0.001024\ \sum_i(Xi-\bar{X})(Yi-\bar{Y})=0.2535.$

$$\bar{X} = 5.062 \quad \bar{Y} = 0.0111$$

$$\sum_i (X_i - \bar{X})^2 = 100.210$$

$$\sum_i (Y_i - \bar{Y})^2 = 0.001024$$

$$\sum_i (X_i - \bar{X})(Y_i - \bar{Y}) = 0.2535.$$

a. What is the rate of change in relative growth rate per minute of fleck duration? Provide a standard error for your estimate.

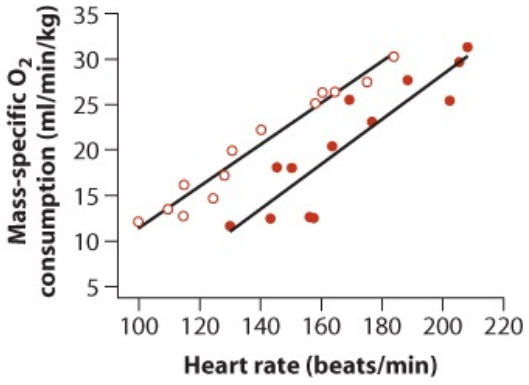
b. Using these data, test the hypothesis that fleck duration affects seedling growth rate.

c. Calculate a 99% confidence interval for the slope of the population regression.

d. What are your assumptions in parts (a)−(c)?

e. What is the main procedure you would employ to evaluate those assumptions?

l. How do we estimate a regression relationship when each subject is measured multiple times over a series of *X*-values? The easiest approach is to use a *summary* slope for each individual and then calculate the average slope. Green et al. (2001) dealt with exactly this type of problem in their study of macaroni penguins exercised on treadmills. Each penguin was exercised at a range of speeds, and its oxygen consumption was measured in relation to its heart rate (a proxy for metabolic rate). The graph provided shows the relationship for just two individual penguins.



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

The following table lists the estimated regression slopes for each of 24 penguins in three categories.

Group	Regression slope
Breeding males	0.31, 0.34, 0.30, 0.38, 0.35, 0.33, 0.32, 0.32, 0.37
Breeding females	0.30, 0.32, 0.23, 0.38, 0.31, 0.26, 0.42, 0.28, 0.35

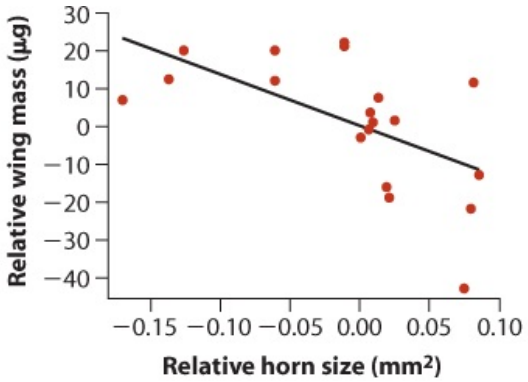
Molting females	0.25, 0.41, 0.32, 0.34, 0.27, 0.23

GenoEJSajko/Getty Images

a. Calculate the mean, standard deviation, and sample size of the slope for penguins in each of the three groups. Display your results in a table.

b. Test whether the means of the slopes are equal between the three groups.

5. Many species of beetle produce large horns that are used as weapons or shields. The resources required to build these horns, though, might be diverted from other useful structures. To test this, Emlen (2001) measured the sizes of wings and horns in 19 females of the beetle species *Onthophagus Sagittarius*. Both traits were scaled for body-size differences and hence are referred to as relative horn and wing sizes. Emlen's data are shown in the following scatter plot along with the least squares regression line ($Y = -0.13 - 132.6X$).



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

We used this regression line to predict the horn lengths at each of the 19 observed horn sizes. These are given in the following table along with the raw data.

Relative horn size (mm^2)	Relative wing mass (mg)	Predicted relative wing mass (mg)
0.074	−42.8	−9.9
0.079	−21.7	−10.6
0.019	− 18.8	−2.6
0.017	−16.0	−2.4
0.085	−12.8	−11.4
0.081	11.6	−10.9
0.011	7.6	−1.6
0.023	1.6	−3.2

0.005	3.7	−0.8
0.007	1.1	−1.1
0.004	−0.8	−0.7
−0.002	−2.9	0.1
−0.065	12.1	8.5
−0.065	20.1	8.5
−0.014	21.2	1.7
−0.014	22.2	1.7
−0.132	20.1	17.4
−0.143	12.5	18.8
−0.177	7.0	23.3

a. Use these results to calculate the residuals.

b. Use your results from part (a) to produce a residual plot.

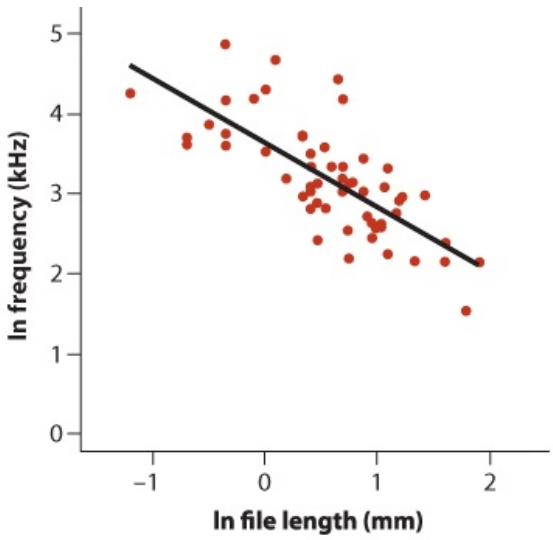
c. Use the graph provided and your residual plot to evaluate the main assumptions of linear regression.

d. In light of your conclusions in part (c), what steps should be taken?

ɜ. Can the songs of extinct species be predicted? Gua et al. (2012) used measurements of living species of katydid to predict the call frequency, or "pitch," of the extinct *Archaboilus musicus* based on a 165-million-year-old fossil. Male katydids call by stridulating—rubbing fore-wings together so that a scraper on one wing rubs against a "file" on the other. Call frequency is predicted by file length (see accompanying graph; the data are available at whitlockschluter.zoology.ubc.ca). File length of a single well-preserved fossil of the extinct *Archaboilus musicus* was 9.34 mm. What was its call frequency?

Summary for log-transformed data is as follows:
n=58,∑iXi=33.241,∑iYi=183.936,∑iXi2=42.615,∑iYi2=609.994,∑iXiYi=86.720.

$$n = 58, \sum_i X_i = 33.241, \sum_i Y_i = 183.936, \sum_i X_i^2 = 42.615, \sum_i Y_i^2 = 609.994, \sum_i X_i Y_i = 86.720.$$



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

a. Calculate the regression line from the summary numbers provided. Assume for the purpose of this exercise that the data points are independent.[11]

b. On the basis of this regression, what is the predicted log-transformed call frequency of *Archaboilus musicus?* The log file length for this species is 2.23.

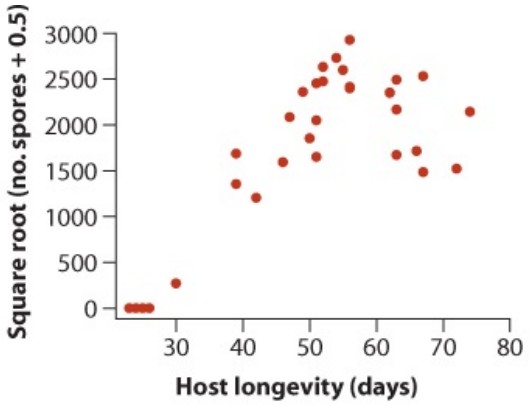
c. What is the most-plausible range of values for the stridulation frequency of the 165-million-year-old katydid? Give the appropriate 95% confidence interval or prediction interval to determine this.

d. Calls with a frequency above about 20 kHz [or ln(frequency) of about 3.0] are ultrasonic and inaudible by most humans. How confident can we be that the calls of *Archaboilus musicus* were audible to humans? Answer based on your confidence or prediction interval in part (c).

7. The parasitic bacterium *Pasteuria ramosa* castrates and later kills its host, the crustacean *Daphnia magna*. The length of time between infection and host death affects the number of spores eventually produced and released by the parasite, as the following scatter plot reveals. The *x*-axis measures age at death for 32 infected host individuals, and the response variable is the square-root-transformed number of spores produced by the infecting parasite (Jensen et al. 2006).

a. Describe the shape of the relationship between the number of spores and host longevity.

b. What equation would be best to try first if you wanted to carry out a nonlinear regression of *Y* on *X*?

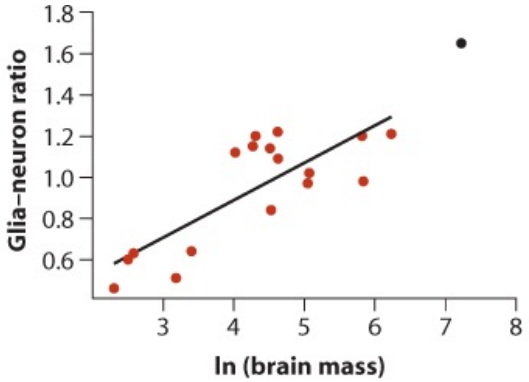


Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

8. Human brains have a large frontal cortex with excessive metabolic demands compared with the brains of other primates. However, the human brain is also three or more times the size of the brains of other primates. Is it possible that the metabolic demands of the human frontal cortex are just an expected consequence of greater brain size? Sherwood et al. (2006) investigated this question in a number of ways. Their data in the accompanying table and scatter plot shows the relationship between the glia-neuron ratio (an indirect measure of the metabolic requirements of brain neurons) and the log-transformed brain mass in nonhuman primates. A linear regression is drawn through these data.

Species	Brain mass (g)	ln (brain mass)	Glia-neuron ratio
Homo sapiens	1373.3	7.22	1.65
Pan troglodytes	336.2	5.82	1.20
Gorilla gorilla	509.2	6.23	1.21
Pongo pygmaeus	342.7	5.84	0.98

Hylobates muelleri	101.8	4.62	1.22
Papio anubis	155.8	5.05	0.97
Mandrillus sphinx	159.2	5.07	1.02
Macaca maura	92.6	4.63	1.09
Erythrocebus patas	102.3	4.53	0.84
Cercopithecus kandti	71.6	4.27	1.15
Colobus angolensis	74.4	4.31	1.20
Trachypithecus francoisi	91.2	4.51	1.14
Alouatta caraya	55.8	4.02	1.12
Saimire boliviensis	24.1	3.18	0.51
Aotus trivirgatus	13.2	2.58	0.63
Saguinus oedipus	10.0	2.30	0.46
Leontopithecus rosalia	12.2	2.50	0.60
Pithecia pithecia	30.0	3.40	0.64



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

a. Determine the equation of the regression line for nonhuman primates.

b. Using the nonhuman primate relationship, what is the predicted glia-neuron ratio for humans, given their brain mass?

c. Determine the most-plausible range of values for the prediction. Which confidence interval is relevant for your prediction of human glia-neuron ratio in (b): the confidence interval for the predicted mean glia-neuron ratio at the given brain mass, or the interval for the prediction of a single new observation?

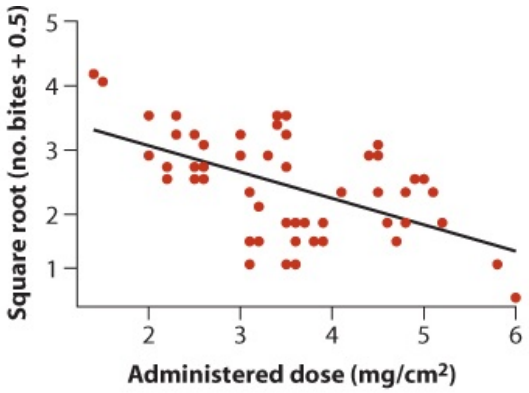
d. Carry out the calculation of the 95% confidence interval chosen in part (c). (See the Quick Formula Summary for the method.) Assume for the purpose of this exercise that the species data are independent.

e. On the basis of your result in part (d), does the human brain have an excessive glia-neu-ron ratio for its mass compared with other primates? Explain.

f. Considering the position of human data point relative to those data used to generate the regression line (see accompanying figure), what additional caution is warranted? Why?

). Golenda et al. (1999) carried out a human clinical trial to investigate the effectiveness of a formulation of DEET (*N*,*N*-diethyl-*m*-toluamide) in preventing mosquito bites. The study applied DEET to the underside of the left forearm of volunteers. Cages containing 15 fresh mosquitoes were then placed over the skin for five minutes, and the number of bites was

recorded. This was repeated four times at intervals of three hours. The scatter plot provided displays the total number of bites (square-root transformed) received by 52 women in the study in relation to the dose of DEET they received.



Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

a. What are the uses of the square-root transformation in linear regression?

b. What feature of this study justifies our calling it an *experimental* study rather than just an observational study?

c. Complete the ANOVA table for these data.

Source of variation	Sum of squares	*df*	Mean squares	*F*-ratio
Regression	9.97315			
Residual				
Total	32.0569			

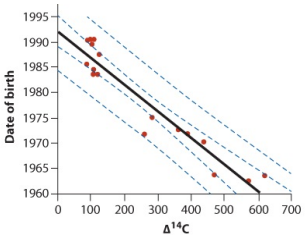
d. Use the *F*-statistic to test the null hypothesis of zero slope.

e. Calculate the R^2 statistic. What does it measure?

Everett Collection/Everett Collection

). Calculating the year of birth of cadavers is a tricky enterprise. One method proposed is based on the radioactivity of the enamel of the body's teeth. The proportion of the radioisotope ^{14}C in the atmosphere increased dramatically during the era of aboveground nuclear bomb testing between 1955 and 1963. Given that the enamel of a tooth is non-regenerating, measuring the ^{14}C content of a tooth tells when the tooth developed, and therefore the year of birth of its owner. Predictions based on this method seem quite accurate (Spalding et al.

2005), as shown in the accompanying graph. The *x*-axis is $\Delta^{14}C$, which measures the amount of ^{14}C relative to a standard (as a percentage).



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

There are three sets of lines on this graph. The solid line represents the least-squares regression line, predicting the actual year of birth from the estimate based on amount of ^{14}C. One pair of dashed lines shows the 95% confidence bands and the other shows the 95% prediction interval.

a. What is the approximate slope of the regression line?

b. Which pair of lines shows the confidence bands? What do these confidence bands tell us?

c. Which pair of lines shows the prediction interval? What does this prediction interval tell us?

l. A lot of attention has been paid recently to portion size in restaurants, and how it may affect obesity in North Americans. Portions have grown greatly over the last few decades. But is this phenomenon new? Wansink and Wansink (2010) looked at representations of the Last Supper in European paintings painted between about 1000 AD and 1700 AD. They scanned the images and measured the size of the food portions portrayed (relative to the sizes of heads in the painting). (For example, the painting reproduced here was painted by Ugolino di Nerio in 1234 AD.) They reported the year of the painting and the portion size as follows:

Portion size	Year
3.08	999
2.70	1004
2.14	1050
2.91	1098
3.69	1314
4.41	1314
3.51	1350
2.44	1309
3.21	1398
2.78	1400
3.39	1467
3.21	1458
3.17	1486
2.78	1494
2.57	1479
3.30	1527

3.81	1525
3.99	1520
4.24	1542
4.80	1515
5.40	1522
5.27	1568
5.44	1554
5.44	1544
5.70	1544
3.04	1561
3.30	1573
3.47	1618
3.56	1626
5.87	1707
1.93	1153
1.76	1434
1.84	1426

a. Calculate a regression line that best describes the relationship between year of painting and the portion size. What is the trend? How rapidly has portion size changed in paintings?

b. What is the most-plausible range of values for the slope of this relationship? Calculate a 95% confidence interval.

c. Test for a change in relative portion size painted in these works of art with the year in which they were painted.[12]

d. Draw a residual plot of these data and examine it carefully. Can you see any cause for concern about using a linear regression? Suggest an approach that could be tried to address the problem.

2. Scarlet king snakes (left photo) are relatively harmless snakes from the southeastern United States. Most individuals have a conspicuous color pattern very similar to the extremely venomous coral snake (right photo). The king snake mimics are thought to gain a survival advantage when coral snakes are present, because predators have learned to avoid coral snakes. However, king snakes also live well outside the range of coral snakes, where the conspicuous colors of these mimics should make them more vulnerable than non-mimic king snakes, because the predators have not learned to avoid coral snakes. To test this, Harper and Pfennig (2008) compared predation rates on mimic and non-mimic king snake color patterns at locations with varying distance from the boundary of the range of coral snakes. The results are given in the table. The first variable is the distance in km between each study location and the boundary of the area where coral snakes are present. Negative numbers mean locations inside the range of coral snakes, and positive numbers mean locations outside the range. At each location, plasticine dummies of king snakes were set out in the habitat, with half the dummies painted to look like mimics and the other half like less-conspicuous non-mimics. The second variable in the table is the proportion of attacks by predators on the mimics at each location.

Distance from boundary	Proportion of attacks on mimics
−97	0
−47	0.01
−33	0
−23	0
−72	0.33
−23	0.5
152	0.4
−15	0.67
97	0.66
113	0.66
105	1
80	1
138	1
148	1
152	1
49	0.4
48	0

a. Give the equation for the line that best predicts proportion of attacks on mimics from the distance to the boundary. What is the trend? Assume that the relationship is linear over the range of the data (being a proportion, the true relationship cannot extend below zero or above one).

b. Test the hypothesis that distance to the boundary predicts the proportion of attacks on mimics.

3. The warm temperatures of spring and summer arrive earlier now at high latitudes than they did in the past, as a result of human-caused climate change. One consequence is that many

organisms start breeding earlier in the year than in previous years, often at suboptimal times. For example, historically the great tit *Parus major* (a well-studied European bird; see Example 2.3A) laid its eggs on dates that resulted in the chicks hatching around the time that caterpillars, a major source of food, became abundant. Currently, a shift in breeding date has led to a mismatch between hatching date and the dates when the caterpillars appear. Does this mismatch affect the growth rate of the bird population? To test this, Reed et al. (2013) used multiple years of study to examine the average timing mismatch (X), in days, with the growth rate of the bird population (Y), expressed as log of the ratio of the number of birds in one year over the number of birds in the previous year. A growth rate greater than zero indicates that the population is increasing, whereas a negative value indicates that the population is declining. Their data are summarized below (available at whitlockschluter.zoology.ubc.ca).

$\sum iXi=4.923885, \sum iYi=41.12394, n=38, \sum iXi2=2005.83430, \sum iYi2=50.15243, \sum iXiYi=-3.9752$

$$\sum_i X_i = 4.923885, \sum_i Y_i = 41.12394, n = 38,$$

$$\sum_i X_i^2 = 2005.83430, \sum_i Y_i^2 = 50.15243,$$

$$\sum_i X_i Y_i = -3.97524.$$

a. Find the formula of the line that best predicts population growth rate from mismatch. What is the trend in growth rate with timing mismatch?

b. What is the confidence interval for the slope of this line?

c. Are the data consistent with a "substantial" effect on population growth rate (where "substantial" refers to a decline of 0.1 or more in growth per 10-day mismatch, which would be enough to cause extinction with expected climate change)?

d. Is there a significant relationship between mismatch and growth rate of the population? Carry out a formal test.

l. Dads transmit many more new mutations than do mothers to their babies at conception. These mutations occur from copying errors during sperm production. There is increasing interest in the effect of father age on this process. As part of a larger study into the genetics of mental illness, Kong et al. (2012) used complete genome squencing of 21 father-child pairs to tally the total number of new mutations inherited from each father (in this particular sample, all the offspring were afflicted with schizophrenia). These counts are listed in the following table along with fathers' ages at offspring conception.

Age of father (years)	Number of new mutations
16	39
18	41
20	39
19	49
22	50
24	54
24	55
24	61
25	57

28	52
29	54
30	57
32	61
37	67
36	70
34	77
30	83
29	67
33	68
26	54
33	65

a. Graph the relationship between number of new mutations (Y) and father's age (X). Add the regression line to your plot.

b. Based on these data, how rapidly does the number of new mutations increase with father's age? Provide a standard error for your estimate.

c. What is the predicted mean number of new mutations from fathers 36 years of age? How does this compare with the predicted number for fathers only 18 years old?

d. Use the ANOVA approach to test the null hypothesis of no relationship between father's age and number of new mutations. Include an ANOVA table with your results.

e. What fraction of the variation among fathers in the number of new mutations is explained by father's age?

6. The threat of bioterrorism makes it necessary to quantify the risk of exposure to infectious agents such as anthrax (*Bacillus anthracis*). Hans (2002) measured the mortality of rhesus monkeys in an exposure chamber to aerosolized anthrax spores of varying concentration. The data are available at whitlockschluter.zoology.ubc.ca and are tabulated below.

Anthrax concentration (spores/l)	Survived	Died
29,300	7	1
32,100	4	4
45,300	3	5
57,300	2	6
64,800	3	5
67,000	5	3
100,000	0	8
125,000	1	7
166,000	0	8

a. Graph the relationship between mortality (Y) and anthrax concentration (X).

b. We would like to use these data to predict the probability of death based on anthrax concentration. Which assumptions of linear regression are violated by these data? Explain.

c. Which method could be used instead to predict mortality from anthrax concentration?

d. An analysis of these data using the method in part (c) yielded the following results. Using these results, what is the predicted mortality from a concentration of 100,000 spores/l?

	Estimate	SE
Intercept	−1.7445	0.6206
Slope	0.00003643	0.00001119

Model	*df*	Deviance	Residual *df*	Residual deviance
Null			71	92.982
Duration	1	19.02	70	73.962

e. Based on these results, add the regression curve to your plot in (a).

f. Based on these data, what is the concentration predicting a 50% mortality (include units)?

g. Using the results in part (d), test the null hypothesis of zero slope.

h. Do individual differences in stress physiology influence survival or reproduction in natural populations? Blas et al. (2007) investigated this question in a Spanish population of European white stork (*Ciconia ciconia).* The accompanying data display stress-induced corticosterone levels circulating in the blood of 34 storks, measured once when they were nestlings, and their survival over the subsequent five years of study. “Stress” involved restraining each stork for 45 minutes and then taking a blood sample.

Corticosterone (ng/ml)	Survival
26	1
28.2	1
29.8	1
34.9	1
34.9	1
35.9	1
37.4	1
37.6	1
38.3	1
39.9	1
41.6	1
42.3	1
52	1
26.6	0
27	0
27.9	0
31.1	0
31.2	0
34.9	0
35.9	0

41.8	0
43	0
45.1	0
46.8	0
46.8	0
47.4	0
47.4	0
47.7	0
47.8	0
50.7	0
51.6	0
56.4	0
57.6	0
61.1	0

a. Graph the relationship between survival (Y) and stress-induced corticosterone levels (X).

b. Give three reasons that linear regression would not be suitable for these data.

c. What regression method could be used instead to predict stork survival from corticosterone levels? How does the method overcome the problems noted in part (b)?

d. An analysis of these data using the method in (c) yielded the following results. Based on these results, add the regression curve to your plot in part (a).

	Estimate	SE
Intercept	2.70304	1.74725
Slope	−0.07980	0.04368

Model	*df*	Deviance	Residual *df*	Residual deviance
Null			33	45.234
Duration	1	3.84	32	41.396

e. Based on these data, what is the estimated concentration predicting a 50% mortality (include units)?

f. Using the results in part (d), test the null hypothesis of zero slope.

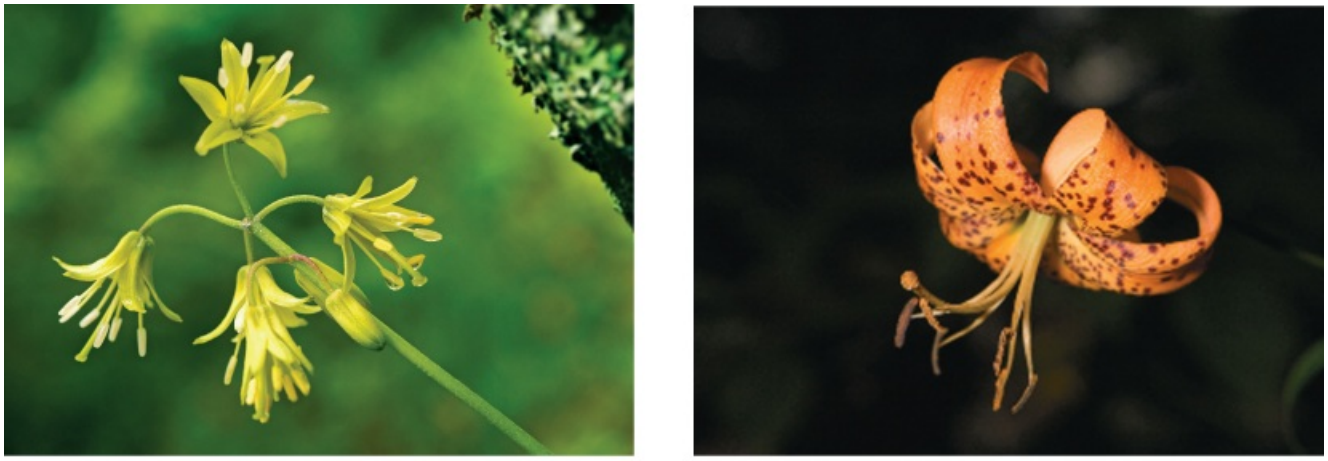
11 INTERLEAF

Using species as data points

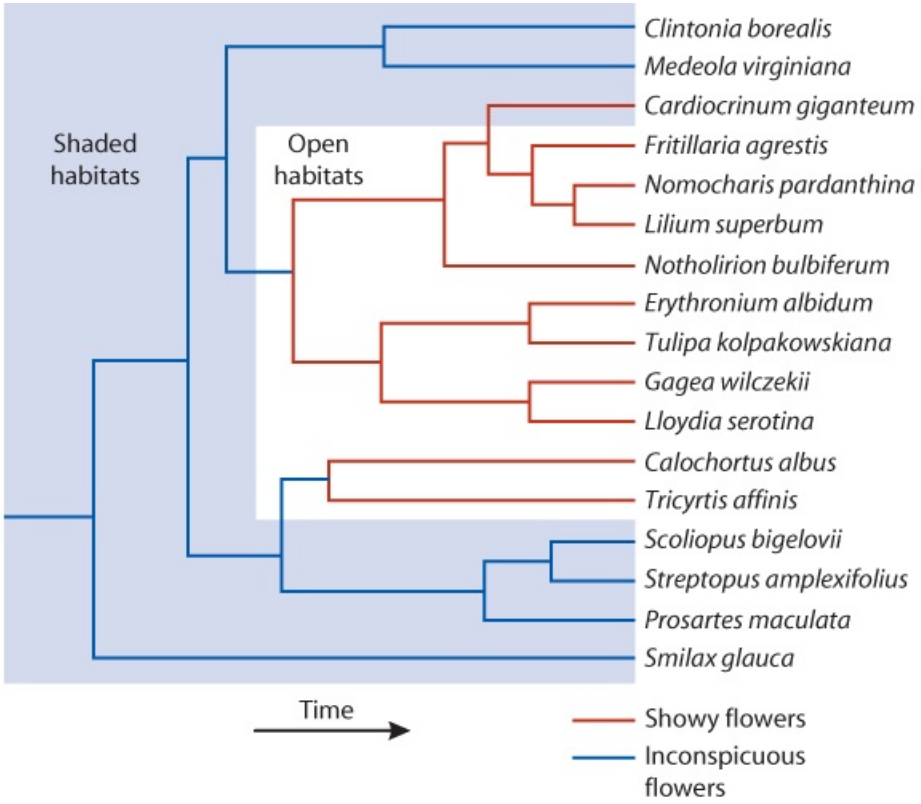
Many types of studies in biology use species measurements as data points. We've encountered several cases in this book. In Chapter 2, for instance, we looked at the frequency distribution of bird abundances using the abundance measurements of different bird species as data points (Example 2.2B). In Chapter 16, we illustrated the association between brain and body mass in mammals by using measurements of different mammal species as data points. What we haven't told you yet is how tricky it can be to analyze such data. The trouble is that species data are not usually *independent*. The reason is that species share an evolutionary history. Here we explain the situation and what can be done about it.

The following study of lilies illustrates the problem created by shared evolutionary history. Patterson and Givnish (2002) found that lily species flowering in the low-light environment of the forest understory, such as the bluebead lily (*Clintonia borealis;* below left), tend to have small and inconspicuous flowers that are whitish or greenish in color. Lilies that live in sunny, open habitats, or that live in deciduous woods but flower before the tree leaves come out, such as the Turk's-cap lily (*Lilium superbum;* below right), tend to have large, showy flowers. Data from 17 lily species, shown in the branching figure on the next page, indicate an almost perfect association between habitat and flower type. All 10 species flowering in open habitats had large and showy flowers. Six of the seven species flowering in shaded habitats had relatively small and inconspicuous flowers. A χ^2 contingency test with these data soundly rejects the null hypothesis of no association ($\chi^2 = 13.24$, $df = 1$, $P = 0.0003$).

However, this contingency test assumes that the data from the 17 species of lilies are independent. The figure indicates that this assumption is likely false. The branching tree in the figure is a *phylogeny*, indicating the ancestor-descendant relationships among the 17 lily species in the data set, which are at the tips of the tree. Branching points, or nodes, in the tree represent points in history when a single ancestor split into two descendant species. Two lily species at the tips of the tree are relatively closely related if they have a recent ancestor in common, such as *Nomocharis pardanthina* and *Lilium super-bum*. Two species are more distantly related if their common ancestor is deeper in the tree, such as *Lilium superbum* and *Prosartes mac-ulata*. The color of the branches in the figure indicates flower type, and the background shading indicates habitat type. (We can only guess about the habitat types and flower types of the ancestors, because they are not alive any more. The colors and shading used in the figure represent just one of the more likely possible scenarios for the transitions in habitat and flower type through history.)



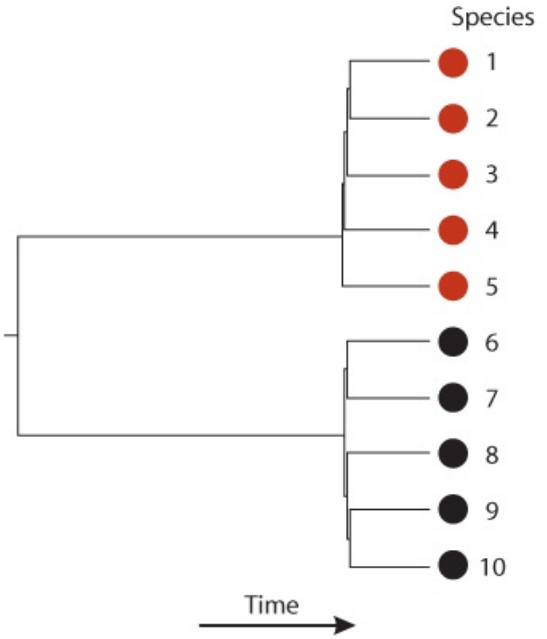
Left Photo: Dale Wilson/Masterfile; Right Photo: mhare2000/Getty Images



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

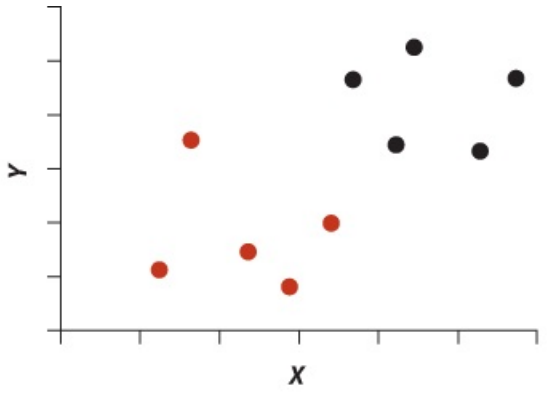
The crucial insight from the tree is that closely related lily species tend to have the same flower type. They also tend to have the same habitat type. Both attributes were likely inherited from their common ancestor. In all, *closely related species are more similar on average than species picked at random*. This means that the species data points are not independent. The situation is like that confronted when pollsters interview more than one person from the same household, or when a bird researcher measures more than one chick from the same nest. However, in the present case, the non-independence is not the fault of the way the species were sampled by the researcher. It is generated by the process of evolution. This is a problem unique to biology.

The preceding example is about discrete variables, but the same issue arises when species data are numerical. The following extreme case makes the point. The branching tree in the figure above shows a phylogeny for 10 hypothetical species. In this example, the species fall into two lineages that split from a common ancestor a long time ago (at the first branching point at the far left of the figure). The lineages have been evolving separately ever since. More recently, each lineage split into five new species more or les; simultaneously. We've used distinct colors to represent species in the two groups.



Whitlock et al., *The Analysis of Biological Data*, 2e,

Now consider two numerical variables, *X* and *Y*, measured on all 10 hypothetical species. It is often the case that species in the same group (indicated by red or black) will be more similar to each other in these traits than to species in the alternative group, just because they share a more recent common ancestor. The scatter plot at bottom left shows example measurements for the 10 species. The symbols in the scatter plot indicate the main lineage to which the species belong.



Whitlock et al., *The Analysis of Biological Data*, 2e,

If we paid no attention to the evolutionary relationships among the 10 species represented by the different symbols, and if we assumed that the species data were independent, we would conclude that *X* and *Y* were positively correlated ($r = 0.69$, $df = 8$, $P = 0.016$). However, it is clear from the scatter plot that closely related species have similar values of *X* and of Y, a likely outcome of the common history they shared until very recently. Within the two groups, there appears to be no correlation between *X* and *Y*.

What can be done about the non-independence of species data resulting from shared evolutionary history? Happily, methods have been developed that correct for the problem. They make it possible to test whether an association is present and to put confidence limits on the strength of the association. The most widely used method for analyzing associations between continuously varying species traits is known as **phylogenetically independent contrasts,** invented by Felsenstein (1985). His explanation of how and why it works is very clear, and we refer you to his original paper for details. Analogous methods have been developed for

categorical species traits.[1]

None of the methods that correct for the problem of non-independence of species traits are foolproof, because all make assumptions that can be difficult to verify. For example, they all assume that the process of trait evolution through time can be adequately mimicked by a simple mathematical model of a "random walk." If the mathematical model badly describes the process of evolution in a specific instance, then using the method can be even worse than ignoring the problem of shared history altogether and just using conventional statistical methods. Biologists nowadays tend to cover all bases and analyze their data both ways. Another strategy is to begin an analysis of species data by examining whether closely related species really are more similar on average than species picked at random. If not, then conventional statistical methods are adequate. If so, then the more specialized methods are used, often along with the results from conventional statistical methods, so that the outcomes can be compared.

1. Specialized computer programs are available to carry out phylogenetic comparative methods for continuously varying traits (such as phylogenetically independent contrasts) and discrete traits, such as MESQUITE (Maddison and Maddison 2011). Several contributed packages are available for the R statistical computing language (see the topic "Trait Evolution" at http:cran.r-project.org/web/views/Phylogenetics.html).

Review Problems 3

1. The early movies by Eadweard Muybridge in the late 19th century showed for the first time the exact positions and movements of the legs during walking by horses and other large mammals. How much has this scientific analysis affected the representation of such animals in art? And how well do modern images of quadrupeds depict walking compared to images made by prehistoric humans? Horvath et al. (2012) examined a large number of images of horses and other animals in art created after Muybridge, in art made by modern humans before Muybridge, and in art from prehistoric humans depicted in cave paintings. For each image, they assessed whether the animal was presented in a biologically realistic posture. The data are at the bottom of the page.
   a. Draw a graph of these data. What is the pattern?
   b. Is there a statistically significant difference in modern images before and after Muybridge in the probability of getting the posture correct?

A Horse, c.15,000–10,000BC (pigments on stone) /Prehistoric/VISUAL ARTS LIBRARY/Caves of Lascaux, Dordogne, France/Bridgeman Images

TABLE FOR PROBLEM 1

Period	Correct walking posture	Incorrect walking posture
Prehistoric	21	18
Modern (pre-Muybridge)	45	227
Modern (post-Muybridge)	289	397

   c. What assumptions are you making in (b)?
   d. Calculate a confidence interval for the proportion of prehistoric paintings that depicted walking posture correctly.

2. Assume that the few remaining hairs on a balding man's head occur independently of each other and with equal probability for each square centimeter (cm) of scalp. Imagine that this man has 2.3 hairs per square cm on average. What is the probability that a randomly chosen square cm of his scalp has exactly four hairs?

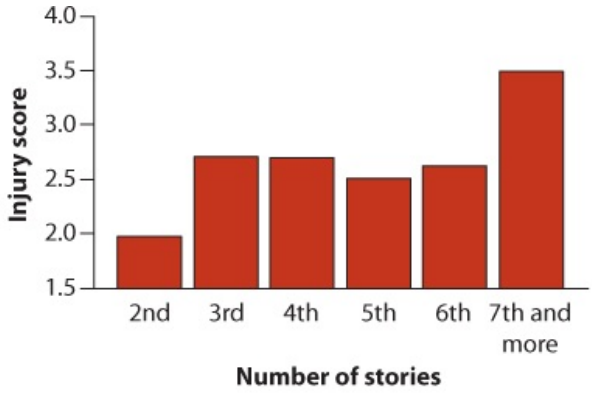
3. Many species have "assortative mating," meaning that a female is more likely to mate with a

male that is similar to her in some particular feature, such as body size, than with a dissimilar male. Imagine a female butterfly weighing 0.4 g in a population where the weight of males is normally distributed with mean 0.3 g and standard deviation 0.08 g. Assume that the female encounters males independently of his body weight.

a. What is the probability that the first male she encounters is within 0.1 g of her own weight?

b. What is the probability that the first five males she encounters are all more than 0.1 g different from her in body weight?

l. *Spot the flaw*. In their more recent study of "high-rise syndrome" (see Chapter 1), Vnuk et al. (2004) reported injury scores (0–4) of 119 fallen cats brought to a veterinary clinic in Zagreb, Croatia. The graph illustrates the average injury scores by the number of stories fallen. Identify the two principles of good graph design that are violated in this figure.



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

ɔ. Most of us would like science to find ways to extend our lives. Genetic research has found some promising variants of genes in other organisms that greatly increase life span. Some mutations in the gene *daf-2* cause the worm *Caenorhabditis elegans*[1] to live almost three times as long as normal worms. But does this greater life span come at a cost? Below are data for the average life span (in days) of worms having one of 14 different mutations at the *daf-2* gene, along with data on the number of offspring produced during their lives (expressed as a percentage of number of offspring of normal unmutated worms) (Gems et al. 1998). Here we wish to calculate the correlation between these variables.

***daf-2* mutation**	**Life span (days)**	**Relative number of offspring**
e1365	28.2	101
m577	25.8	95
sa193	33.5	96
e1371	31.8	99
e1368	33.2	85
m41	27.0	87
m212	48.4	95
e1369	52.8	88
m120	32.3	93
e1370	33.8	70
m596	36.8	75

m579	44.3	73
e1391	63.4	61
e979	50.3	70

a. Both variables have positive skew. Find an appropriate transformation to reduce the skew (you will find that skew can be reduced but not altogether eliminated).

b. Include a plot of the transformed data. What trend is suggested?

c. Is life span significantly correlated with number of offspring? Answer using the transformed data.

6. Mendel famously discovered the basics of genetics with garden peas. He proposed a law of independent assortment, that the inheritance of different genes should be independent. We now know that this "law" is erroneous because genes that are linked on the same chromosome tend to be inherited together. Mendel (1866) used the following data from a cross of peas to test the predictions made by independent assortment.

Yellow smooth: 315

Yellow wrinkled: 101

Green smooth: 108

Green wrinkled: 32

The traits yellow/green and smooth/wrinkled are determined by different genes. If independent assortment were true, then the traits of the offspring of the cross should have the following proportions: 9/16 yellow smooth peas, 3/16 yellow wrinkly peas, 3/16 green smooth peas, and 1/16 green wrinkly pea. Test whether Mendel's prediction about the proportions is consistent with these data.

7. Van Hylckama Vlieg et al. (2009) investigated the relationship between oral contraceptive use and thrombosis in women. In a sample of 1524 adult female patients who had thrombosis, 103 had taken oral contraceptives regularly. In a second sample of 1760 women from the same population who did not have thrombosis, 658 had taken oral contraceptives regularly.

a. What type of study design was used?

b. Graph the data. Which treatment condition (oral contraceptive use) had the higher proportion of women with thrombosis?

c. Test for an association between oral contraceptive use and thrombosis.

d. What is the odds ratio of thrombosis in women taking oral contraceptives compared to women not taking oral contraceptive? Include a confidence interval for the population odds ratio.

e. Under what circumstances can we say that the odds ratio estimated in part (d) is a reasonable estimate of the relative risk of thrombosis?

8. Studying the influence of metabolic differences among individuals on survival or reproductive success in nature requires that an individual's metabolism doesn't vary too much from time point to time point. To investigate, Hayes and O'Connor (1999) measured repeatability of thermogenic capacity (ability to generate heat) by recording maximal rate of oxygen consumption (VO_2max) in high-altitude deer mice exposed to cold temperatures in a wind tunnel. A sample of 34 mice were measured twice about 68 days apart. The two

measurements on each mouse, in ml/min, are given below.

Mouse	VO_2max
1	5.62, 5.84
2	5.13, 4.75
3	5.00, 5.65
4	5.76, 6.07
5	6.10, 5.22
6	5.11, 5.68
7	5.63, 6.10
8	5.40, 6.20
9	4.62, 5.54
10	5.06, 5.56
11	4.29, 5.41
12	5.24, 5.58
13	4.81, 5.04
14	5.84, 5.69
15	5.34, 5.66
16	5.53, 5.89
17	5.59, 4.80
18	5.75, 5.92
19	4.41, 5.15
20	4.63, 4.82
21	5.59, 6.42
22	5.14, 5.31
23	5.18, 5.30
24	5.13, 5.21
25	4.80, 4.64
26	5.69, 6.17
27	5.00, 3.70
28	4.98, 5.32
29	5.33, 5.86
30	5.15, 5.50
31	4.79, 5.46
32	5.95, 5.85
33	5.91, 6.04
34	4.78, 4.83

a. Calculate the variance components of VO_2max within and among deer mice.

b. What is the repeatability of thermogenic capacity, as measured using VO_2max under cold exposure?

c. What are your assumptions in parts (a) and (b)?

). Collins and Bell (2004) investigated the impacts of elevated carbon dioxide (CO_2) concentrations on plant evolution. They raised separate lines of the unicellular algae *Chlamydomonas* under normal and high CO_2 levels. After 1000 generations, they measured the growth rate of all of the experimental lines in a high CO_2 environment. The results for 14 experimental lines are presented in the following table. Growth rate is measured relative to the starting strain and has no units. Use these data to test whether the mean growth rate is associated with the CO_2 treatment.

CO_2 treatment	Growth rate
Normal	2.31
Normal	1.95
Normal	1.86
Normal	1.59
Normal	1.55
Normal	1.30
Normal	1.07
High	2.37
High	1.89
High	1.55
High	1.49
High	1.26
High	1.20
High	0.98

). To investigate whether subcutaneous fat provides insulation in humans, Sloan and Keatinge (1973) measured the rate of heat loss by boys swimming for up to 40 min in water at 20.3°C and expending energy at about 4.8 kcal/min. Heat loss was measured by the change in body temperature, recorded using a thermometer under the tongue, divided by time spent swimming, in minutes. The authors measured an index of body "leanness" on each boy as the reciprocal of the skin-fold thickness adjusted for total skin surface area (in meters squared) and body mass (in kg). Their data are listed in the following table.

Body leanness (m/kg)	Heat-loss rate (°C/min)
7.0	0.103
7.0	0.097
6.2	0.090
5.0	0.091
4.4	0.071
3.3	0.024
3.6	0.014
2.8	0.041

2.4	0.031
2.1	0.010
2.1	0.006
1.7	0.002

a. Draw a scatter plot of these data, showing the relationship.

b. Does body leanness predict heat-loss rate? Using the following intermediate calculations, calculate the regression line and add it to your plot in part (a). Carry out a formal test.

$\bar{X}=3.96667 \quad \bar{Y}=0.04833$

$\sum_i (X_i-\bar{X})^2 = 41.14667$

$\sum_i (Y_i-\bar{Y})^2 = 0.01696$

$\sum_i (X_i-\bar{X})(Y_i-\bar{Y}) = 0.78053.$

c. How uncertain is the estimate of slope? Calculate a 95% confidence interval.

d. What are your assumptions in parts (b) and (c)?

e. What fraction of the variation in heat-loss rate is predictable from body leanness?

l. For each of the following scenarios, state what statistic would be used to estimate the effect of interest.

a. How different are the two hospitals X and Y in the frequency of doctors who wash and do not wash their hands before medical procedures?

b. How different is the number of bacteria on hands between people who wash for one minute and people who do not wash their hands?

c. How different are athletes and nonathletes in the mean number of mitochondria per muscle cell?

d. How different is the number of mitochondria per cell between the muscles of people's dominant arm and the muscles in their other arm?

e. What fraction of individuals in an elephant population are male?

f. How different are the frequencies of males in two populations of elephants?

g. How much variation in weight is there among individuals in an elephant population?

h. How strong is the association between the number of mitochondria per cell in arm muscles and leg muscles?

2. For each of the following scenarios or questions, say which method for hypothesis testing would be most appropriate to best answer the scientific question. Unless otherwise stated, make any necessary assumptions. Be as specific as possible. (*Do not try to answer the biological question posed; just say what statistical technique would be best.*)

a. Do Hospital A and Hospital B differ in the frequency of doctors who wash their hands before medical procedures?

b. Does the mean rate at which doctors wash their hands before medical procedures vary among the three hospitals X, Y, and Z?

c. Does the rate of hand washing at hospitals predict the proportion of patients catching infections?

d. Does washing hands for five minutes leave a different number of bacteria on people's

hands than washing for one minute, on average?

e. Which group washes hands for the greatest mean number of minutes each time, doctors or nurses?

f. Does whether or not doctors wash their hands before the first examination of patients have an effect on the lengths of patient stays in the hospital?

g. Do athletes have more mitochondria per muscle cell on average than nonathletes? (Assume that mitochondria per cell is normally distributed among individuals.)

h. Do athletes have a greater mean number of mitochondria per cell than nonathletes? (Assume that the number of mitochondria is not normally distributed but they have the same shape of distribution in the two groups.)

i. Is the mean number of mitochondria per cell in the muscles of people's dominant arm different from the number in their other arm?

j. Is the proportion of males in an elephant population equal to 0.50?

k. Do two populations of elephants have the same proportion of males?

l. Are the left tusks of elephants on average longer than their right tusks?

m. Are elephants spread out over the savanna independently and with equal probability everywhere?

n. Is the length of elephants' trunks normally distributed?

o. Are male elephants more variable in weight than are females?

p. Do male and female elephants differ in their mean growth rates? (Assume that elephant growth rate is not normally distributed but males and females have the same shape of distribution.)

q. Does the thickness of an elephant's first left molar predict its age?

r. Does the thickness of an elephant's first left molar predict whether it lives at least to 5 years of age?

3. The naked mole rat is a very unusual creature. For one thing, it is the only known mammal that is eusocial, with most individuals forgoing reproduction and instead helping to raise the offspring of the "king" and "queen" of their colony. They also live many times longer than other animals their size, and even up to twice as long as their two closely related species, the blind mole rat and the Damaraland mole rat (Edrey et al. 2012). It is possible that this difference is in part caused by differential expression of a transcription factor called HIF1-α, which regulates proteins called neuregulins (neural growth factors thought to be involved in maintaining nerve function). The data below give measurements of HIF1-α expression in several individuals from each of the three species (expressed as a percentage of the expression of actin, a common protein used as a reference). Does the expression of HIF1-α a differ among these three species?

Ron Austing/FLPA

Naked mole rat: 3.5, 3.8, 5.6, 12.9, 13.9, 28.2

Blind mole rat: 5.2, 8.7, 8.9, 11.4, 12.6

Damaraland mole rat: 4.3, 5.2, 8.4, 10.2, 10.2, 20.6

a. Show the data in a graph. What trend is suggested?

b. Do the species differ significantly in their mean amount of HIF1-α? (Use a log transformation to improve the fit to assumptions.)

l. Previous studies have shown that the antibody titers in obese people are lower after vaccination than in people of normal weight. One suggested reason is that the vaccines may not effectively penetrate the layer of subcutaneous fat in obese individuals. To test this, Middleman et al. (2010) compared the response to hepatitis B virus vaccine in obese participants in two different groups. The researchers vaccinated one group of 10 individuals with standard 1-inch (2.5 cm) needles. They used 1.5-inch (3.8 cm) needles instead for a second group of 14 individuals. They later measured the antibody titers (in units of mIU/ml) of each participant. Greater numbers indicate a more successful response to the vaccine. These results are as follows.

Short-needle group: 51.6, 87.4, 143.6, 144.6, 189.7, 189.8, 208.9, 324.7, 368, 383.9

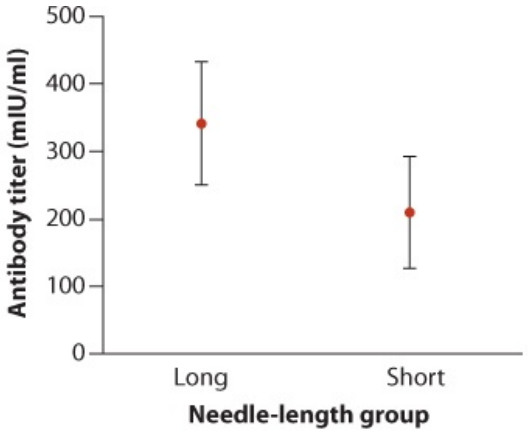
Long-needle group: 28.0, 181.6, 203.9, 243, 249.6, 274.3, 341.2, 349.6, 393.0, 429.2, 464.2, 473.1, 492.9, 647.0

a. What is the most-plausible range of values for the difference in mean antibody titers between the long- and short-needle groups? Use the 95% confidence interval to answer this question.

b. Use an appropriate hypothesis test to compare the means of the two groups. What can you conclude about the effectiveness of the vaccine as a function of the length of the needle?

c. What is the 95% confidence interval for antibody titer in the long-needle group?

ɔ. *Spot the flaw*. Refer to Problem 14. Grover, a student who skipped reading Chapter 2 of this book, made the following graph when presenting the results of the needle-length study to his epidemiology class. The points are means, and the error bars are 95% confidence intervals. What is the biggest weakness of the graph?

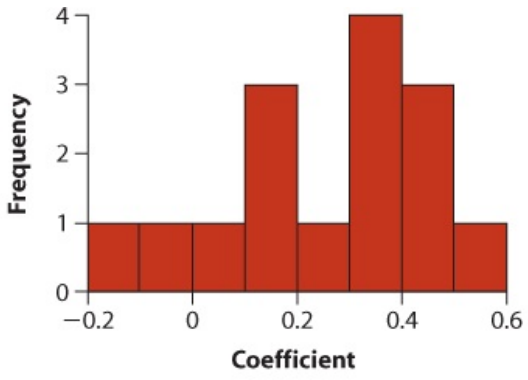


Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

6. We are accustomed to thinking that the proportion of males at birth is fixed by genetics in birds and mammals to be close to 50%. Some have suggested, however, that the sex of offspring can be adjusted by females, such as in response to the quality of her mate or the number of helpers she has. West and Sheldon (2002) found a total of 15 studies, all done on birds, that have measured changes in the sex ratio in response to such social factors, and each has expressed its results in terms of a coefficient ranging from −1 to 1. The coefficient is positive if the change in the sex ratio of offspring is in agreement with evolutionary theory, and negative if the data disagree with the theory. A coefficient of zero indicates no association with social factors in the data. The measures are as follows:

−0.160, −0.037, 0.034, 0.144, 0.137, 0.118, 0.395, 0.363, 0.350, 0.376, 0.253, 0.440, 0.453, 0.460, 0.563.

The frequency distribution of the coefficients is shown in the following histogram.



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

a. What method could be used to test whether these data are consistent with a mean or median coefficient of zero? Discuss why you would use this method in contrast to specific other methods.

b. Apply the best method to these data to test whether the mean coefficient differs from zero.

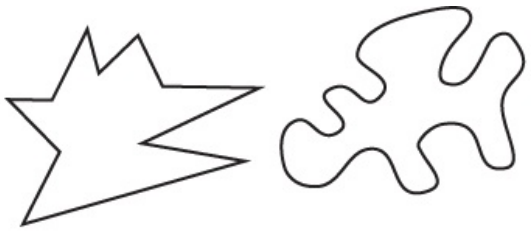
7. Cellulose from the butts of smoked cigarettes is commonly used by urban birds in nest construction. In an observational study of the house finches in Mexico City, Suarez-Rodnguez et al. (2013) discovered that nests with more cellulose from smoked cigarette butts contained fewer nest-dwelling ectoparasites of birds (such as mites) than nests with less cellulose from smoked cigarettes. In a separate experimental study, the researchers placed thermal traps in 28 active house finch nests (the parasites hiding in the nest are drawn to the warmth and become trapped). Smoked Marlboro cigarette butts[2] were placed in the trap in

about half the nests, randomly chosen. In the other nests, filters from unsmoked cigarettes (lacking tobacco residues) were used as control. At the end of the experiment, the researchers counted the number of ectoparasites caught in each trap. The traps containing smoked butts had fewer ectoparasites than traps with unsmoked filters.

a. Which study provided the stronger evidence that the chemical contents of smoked cigarette butts deters ectoparasites: the observational study or the experimental study? Explain your reasoning.

b. Which of the six commonly used components of experimental design were not incorporated in the experimental study described above? What benefit might result from including them?

8. We often assume that the mapping between words and their meanings is completely arbitrary. Maurer et al. (2006) tested whether this was completely true. College students were shown the following two shapes, and asked to say which was "bouba" and which was "kiki."



From *The Shape of Boubas: Sound-Shape Correspondences in Toddlers and Adults*, Maurer, D., Pathman, T., and Mondloch, Catherine J. Developmental Science 9:3 (2006), Fig. 1, p 318. © 2006 The Authors. Journal compilation © 2006 Blackwell Publishing Ltd. Reprinted with permission.

Eighteen of 20 students called the angular shape on the left kiki, while the other two called that shape bouba.

a. Calculate a confidence interval for the proportion of adults who would call the left shape kiki and the other bouba.

b. Test whether kiki and bouba are used with equal probability in the student population.

9. Sex, with its many benefits, also brings risk. For example, individuals that are more promiscuous are exposed to more sexually transmitted diseases. This is true for other primates as well as for our own species. Different species of primates vary widely in the mean number of sexual partners per individual, and this raises the question, are the immune systems of more promiscuous species different from those of less promiscuous species? Researchers approached this question by comparing pairs of closely related primate species, in which one species of the pair was more promiscuous and the other less promiscuous (Nunn et al. 2000). They measured the mean white blood cell (WBC) count in cells per nanoliter for each species. The results are listed in the following table.

WBC count: Less promiscuous species	WBC count: More promiscuous species
5.7	10.4
7.2	10.4
7.4	9.9

8.1	9.1
8.4	9.2
9.2	11.9
9.1	9.3
9.1	8.9
10.6	12.5

a. What is the mean difference in WBC count between less and more promiscuous species? Which type of species (more promiscuous or less promiscuous) has the higher WBC count on average?

b. What is the 99% confidence interval for this difference?

c. Test the null hypothesis that there is no mean difference in WBC count between more and less promiscuous species.

d. What are your assumptions in (b) and (c)?

1. The lowly worm *C. elegans* is a popular study organism in aging research. See also Chapter 15, Assignment Problem 20. The *daf-2* protein is an insulin receptor.
2. Cigarettes were consumed using an artificial smoking machine and contained chemical residues from the smoked tobacco.