Chapter 2

Summarizing Data

After collecting data, the next stage in the investigative process is to summarize the data. Graphical displays allow us to visualize and better understand the important features of a data set.

2.1 Examining numerical data

In this section we will focus on numerical variables. The `email50` and `county` data sets from Section 1.2 provide rich opportunities for examples. Recall that outcomes of numerical variables are numbers on which it is reasonable to perform basic arithmetic operations. For example, the `pop2010` variable, which represents the populations of counties in 2010, is numerical since we can sensibly discuss the difference or ratio of the populations in two counties. On the other hand, area codes and zip codes are not numerical, but rather they are categorical variables.

2.1.1 Scatterplots for paired data

Sometimes researchers wish to see the relationship between two variables. When we talk of a relationship or an association between variables, we are interested in how one variable behaves as the other variable increases or decreases.

A **scatterplot** provides a case-by-case view of data that illustrates the relationship between two numerical variables. In Figure 1.8 on page 8, a scatterplot was used to examine how federal spending and poverty were related in the `county` data set. Another scatterplot is shown in Figure 2.1, comparing the number of line breaks (`line_breaks`) and number of characters (`num_char`) in emails for the `email50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `email50`, there are 50 points in Figure 2.1.

Example 2.1 A scatterplot requires paired data. What does **paired data** mean?

We say observations are *paired* when the two observations correspond to each other. In unpaired data, there is no such correspondence. Here the two observations correspond to a particular email.

The variable that is suspected to be the response variable is plotted on the vertical axis and the variable that is suspected to be the explanatory variable is plotted on the

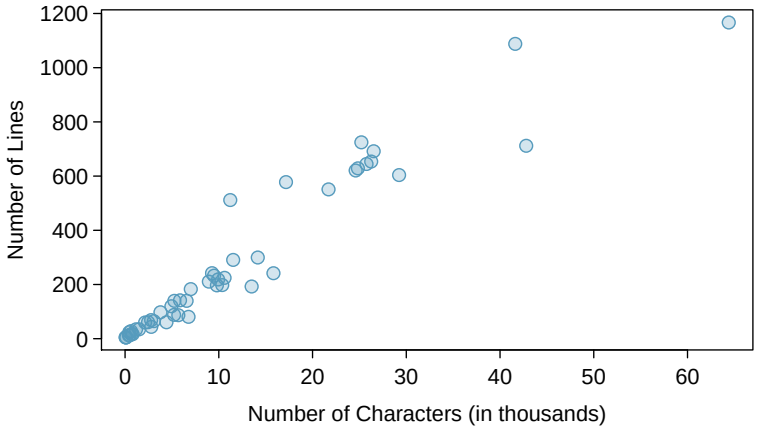

Figure 2.1: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

horizontal axis. In this example, the variables could be switched since either variable could reasonably serve as the explanatory variable or the response variable.

TIP: Drawing scatterplots
(1) Draw the axes and add scales to each. (2) Label each axis. (3) Plot the dots.

The association between two variables can be **positive** or **negative**, or there can be no association. Positive association means that larger values of the first variable are associated with larger values of the second variable. Additionally, the association can follow a linear trend or a curved (nonlinear) trend.

⊙ **Guided Practice 2.2** What would it mean for two variables to have a *negative* association? What about *no* association?[1]

⊙ **Guided Practice 2.3** What does the scatterplot in Figure 2.1 reveal about the email data?[2]

● **Example 2.4** Consider a new data set of 54 cars with two variables: vehicle price and weight.[3] A scatterplot of vehicle price versus weight is shown in Figure 2.2. What can be said about the relationship between these variables?

The relationship is evidently nonlinear, as highlighted by the dashed line. This is different from previous scatterplots we've seen, such as Figure 1.8 on page 8 and Figure 2.1, which show relationships that are very linear.

[1] Negative association implies that larger values of the first variable are associated with smaller values of the second variable. No association implies that the values of the second variable tend to be independent of changes in the first variable.

[2] The association between the number of characters in an email and the number of lines in an email is positive (when one is larger, the other tends to be larger as well). As the number of characters increases, number of lines increases is an approximately linear fashion.

[3] Subset of data from http://www.amstat.org/publications/jse/v1n1/datasets.lock.html

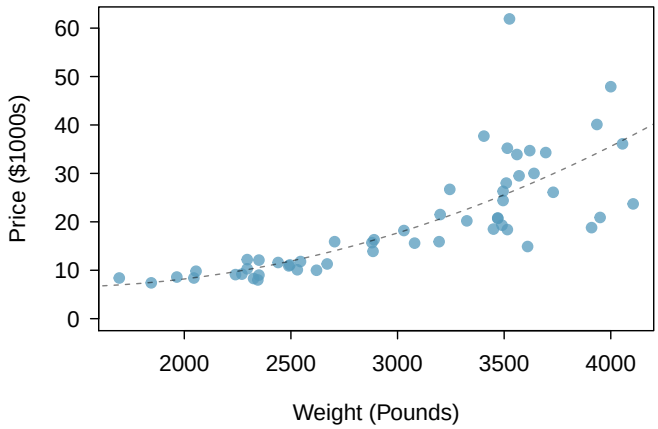

Figure 2.2: A scatterplot of `price` versus `weight` for 54 cars.

⊙ **Guided Practice 2.5** Describe two variables that would have a horseshoe shaped (i.e. "U"-shaped) association in a scatterplot.[4]

[4]Consider the case where your vertical axis represents something "good" and your horizontal axis represents something that is only good in moderation. Health and water consumption fit this description since water becomes toxic when consumed in excessive quantities.

Chapter 8

Introduction to linear regression

8.1 Line fitting, residuals, and correlation

It is helpful to think deeply about the line fitting process. In this section, we examine criteria for identifying a linear model and introduce a new statistic, *correlation*.

8.1.1 Beginning with straight lines

Scatterplots were introduced in Chapter 1 as a graphical technique to present two numerical variables simultaneously. Such plots permit the relationship between the variables to be examined with ease. Figure 8.4 shows a scatterplot for the head length and total length of 104 brushtail possums from Australia. Each point represents a single possum from the data.

The head and total length variables are associated. Possums with an above average total length also tend to have above average head lengths. While the relationship is not perfectly linear, it could be helpful to partially explain the connection between these variables with a straight line.

Straight lines should only be used when the data appear to have a linear relationship, such as the case shown in the left panel of Figure 8.6. The right panel of Figure 8.6 shows a case where a curved line would be more useful in understanding the relationship between the two variables.

Caution: Watch out for curved trends
We only consider models based on straight lines in this chapter. If data show a nonlinear trend, like that in the right panel of Figure 8.6, more advanced techniques should be used.

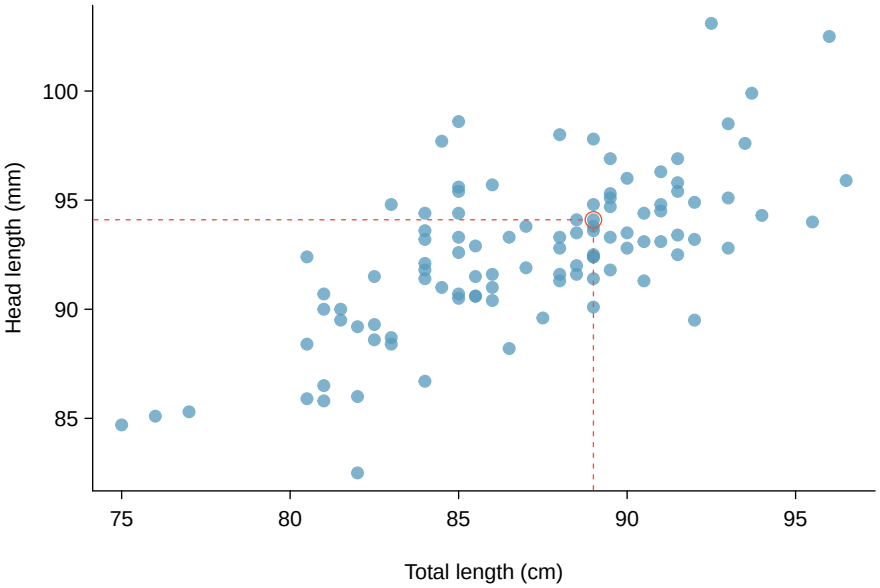

Figure 8.4: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

Figure 8.5: The common brushtail possum of Australia.

Photo by wollombi on Flickr: www.flickr.com/photos/wollombi/58499575

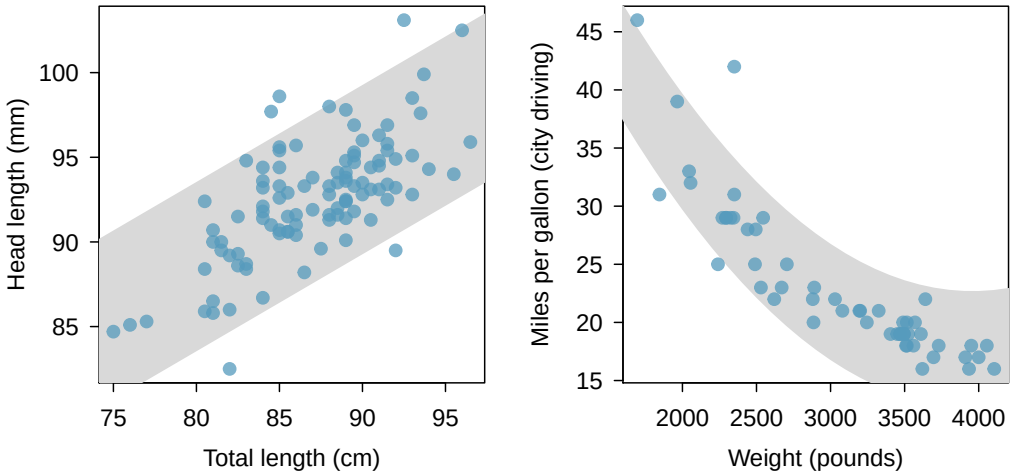

Figure 8.6: The figure on the left shows head length versus total length, and reveals that many of the points could be captured by a straight band. On the right, we see that a curved band is more appropriate in the scatterplot for `weight` and `mpgCity` from the `cars` data set.

8.1.2 Fitting a line by eye

We want to describe the relationship between the head length and total length variables in the possum data set using a line. In this example, we will use the total length as the predictor variable, x, to predict a possum's head length, y. We could fit the linear relationship by eye, as in Figure 8.7. The equation for this line is

$$\hat{y} = 41 + 0.59x \tag{8.2}$$

We can use this line to discuss properties of possums. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of

$$\begin{aligned}\hat{y} &= 41 + 0.59 \times 80 \\ &= 88.2\end{aligned}$$

A "hat" on y is used to signify that this is an estimate. This estimate may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. Absent further information about an 80 cm possum, the prediction for head length that uses the average is a reasonable estimate.

8.1.3 Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

Each observation will have a residual. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

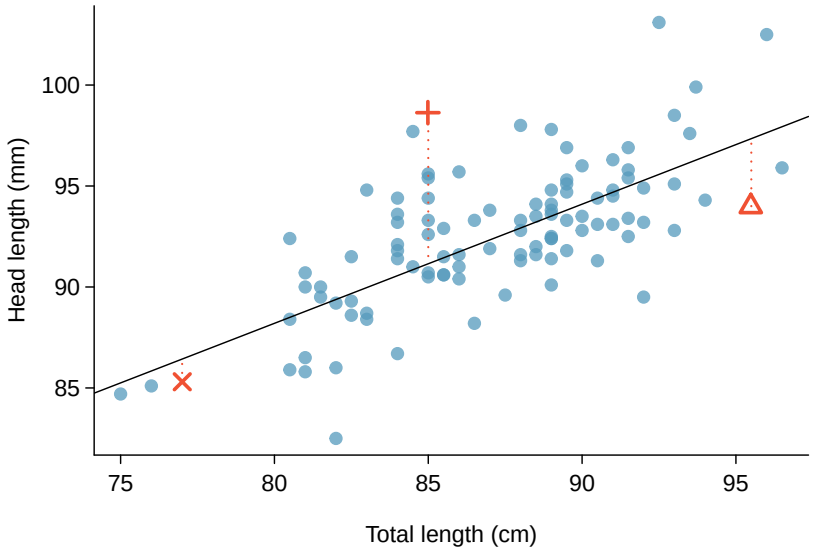

Figure 8.7: A reasonable linear model was fit to represent the relationship between head length and total length.

Three observations are noted specially in Figure 8.7. The observation marked by an "×" has a small, negative residual of about -1; the observation marked by "+" has a large residual of about +7; and the observation marked by "△" has a moderate residual of about -4. The size of a residual is usually discussed in terms of its absolute value. For example, the residual for "△" is larger than that of "×" because $|-4|$ is larger than $|-1|$.

Residual: difference between observed and expected

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$:

$$\text{residual}_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging x_i into the model.

Example 8.3 The linear fit shown in Figure 8.7 is given as $\hat{y} = 41 + 0.59x$. Based on this line, formally compute the residual of the observation $(77.0, 85.3)$. This observation is denoted by "×" on the plot. Check it against the earlier visual estimate, -1.

We first compute the predicted value of point "×" based on the model:

$$\hat{y}_\times = 41 + 0.59x_\times = 41 + 0.59 \times 77.0 = 86.4$$

Next we compute the difference of the actual head length and the predicted head length:

$$residual_\times = y_\times - \hat{y}_\times = 85.3 - 86.4 = -1.1$$

This is very close to the visual estimate of -1.

⊙ **Guided Practice 8.4** If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[1]

⊙ **Guided Practice 8.5** Compute the residuals for the observations (85.0, 98.6) ("+" in the figure) and (95.5, 94.0) ("△") using the linear relationship $\hat{y} = 41 + 0.59x$. [2]

Residuals are helpful in evaluating how well a linear model fits a data set. We often display them in a **residual plot** such as the one shown in Figure 8.8 for the regression line in Figure 8.7. The residuals are plotted at their original horizontal locations but with the vertical coordinate as the residual. For instance, the point $(85.0, 98.6)_+$ had a residual of 7.45, so in the residual plot it is placed at (85.0, 7.45). Creating a residual plot is sort of like tipping the scatterplot over so the regression line is horizontal.

From the residual plot, we can better estimate the **standard deviation of the residuals**, often denoted by the letter s. The standard deviation of the residuals tells us the average size of the residuals. As such, it is a measure of the average deviation between the y values and the regression line. In other words, it tells us the average prediction error using the linear model.

● **Example 8.6** Estimate the standard deviation of the residuals for predicting head length from total length using the regression line. Also, interpret the quantity in context.

To estimate this graphically, we use the residual plot. The approximate 68, 95 rule for standard deviations applies. Approximately 2/3 of the points are within ± 2.5 and approximately 95% of the points are within ± 5, so 2.5 is a good estimate for the standard deviation of the residuals. On average, the prediction of head length is off by about 2.5 cm.

Standard deviation of the residuals
The standard deviation of the residuals, often denoted by the letter s, tells us the average error in the predictions using the regression model. It can be estimated from a residual plot.

[1] If a model underestimates an observation, then the model estimate is below the actual. The residual, which is the actual observation value minus the model estimate, must then be positive. The opposite is true when the model overestimates the observation: the residual is negative.

[2] (+) First compute the predicted value based on the model:

$$\hat{y}_+ = 41 + 0.59x_+ = 41 + 0.59 \times 85.0 = 91.15$$

Then the residual is given by

$$residual_+ = y_+ - \hat{y}_+ = 98.6 - 91.15 = 7.45$$

This was close to the earlier estimate of 7.

(△) $\hat{y}_\triangle = 41 + 0.59x_\triangle = 97.3$. $residual_\triangle = y_\triangle - \hat{y}_\triangle = -3.3$, close to the estimate of -4.

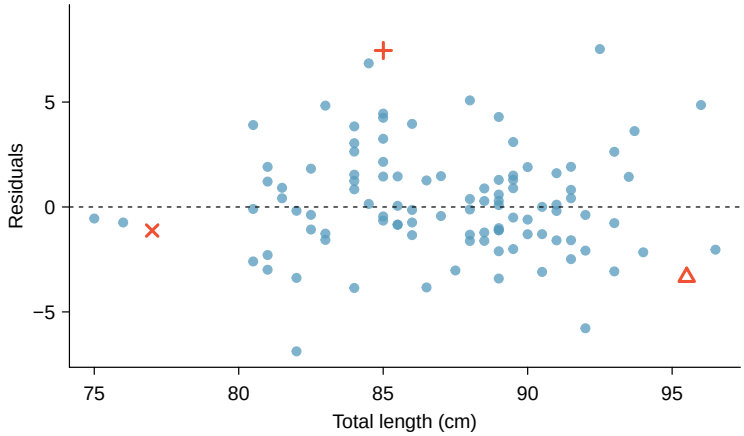

Figure 8.8: Residual plot for the model in Figure 8.7.

Example 8.7 One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 8.9 shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns remaining in the residuals?

In the first data set (first column), the residuals show no obvious patterns. The residuals appear to be scattered randomly around the dashed line that represents 0.

The second data set shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used.

The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is statistically significant evidence that the slope parameter is different from zero. The point estimate of the slope parameter, labeled b_1, is not zero, but we might wonder if this could just be due to chance. We will address this sort of scenario in Section 8.4.

8.1.4 Describing linear relationships with correlation

Correlation coefficient, r, measures the strength of a linear relationship

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. It can be strong, moderate, or weak.

R
correlation

We can compute the correlation coefficient (or just correlation for short) using a formula, just as we did with the sample mean and standard deviation. However, this formula is rather complex,[3] so we generally perform the calculations on a computer or calculator.

[3]Formally, we can compute the correlation for observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) using the formula $r = \frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i-\bar{x}}{s_x}\frac{y_i-\bar{y}}{s_y}$, where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

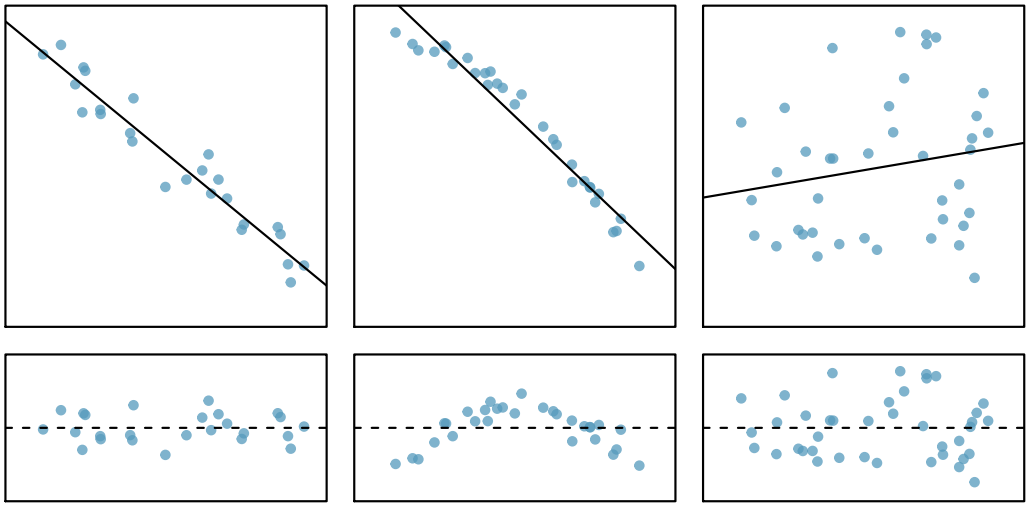

Figure 8.9: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

Figure 8.10 shows eight plots and their corresponding correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.

The correlation is intended to quantify the strength of a linear trend. Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 8.11.

⊙ **Guided Practice 8.8** It appears no straight line would fit any of the datasets represented in Figure 8.11. Try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[4]

● **Example 8.9** Take a look at Figure 8.7. How would this correlation change if head length were measured in cm rather than mm? What if head length were measure in inches rather than mm?

Here, changing the units of y corresponds to multiplying all the y values by a certain number. This would change the mean and the standard deviation of y, but it would not change the correlation. To see this, imagine dividing every number on the vertical axes by 10. The units of y are now cm rather than mm, but the graph has remain exactly the same

Changing units of x and y does not affect r.
The correlation between two variables should not be dependent upon the units in which the variables are recorded. Adding a constant, subtracting a constant, or multiplying a *positive* constant to all values of x or y does not affect the correlation.

[4] We'll leave it to you to draw the lines. In general, the lines you draw should be close to most points and reflect overall trends in the data.

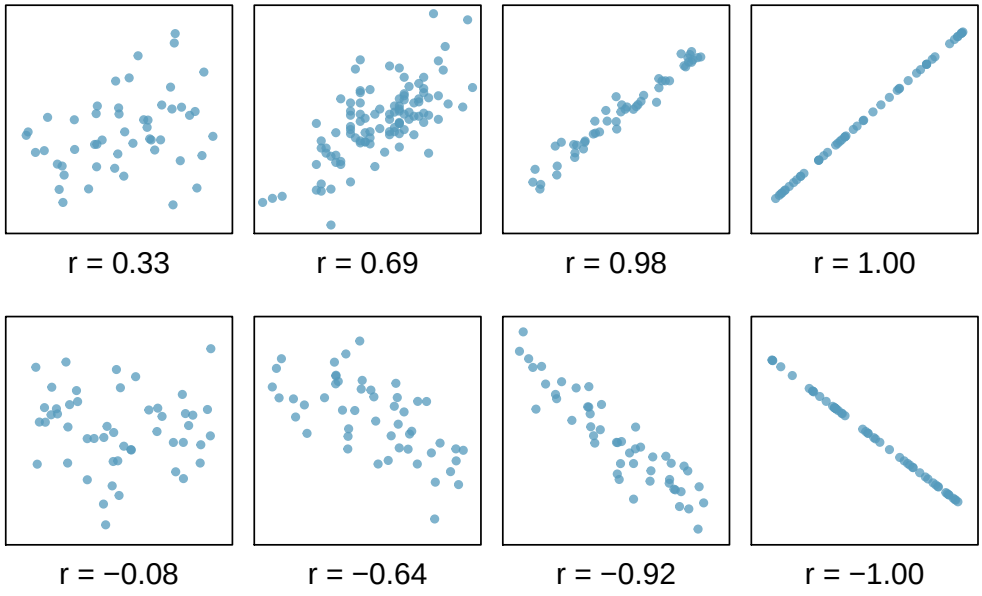

Figure 8.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

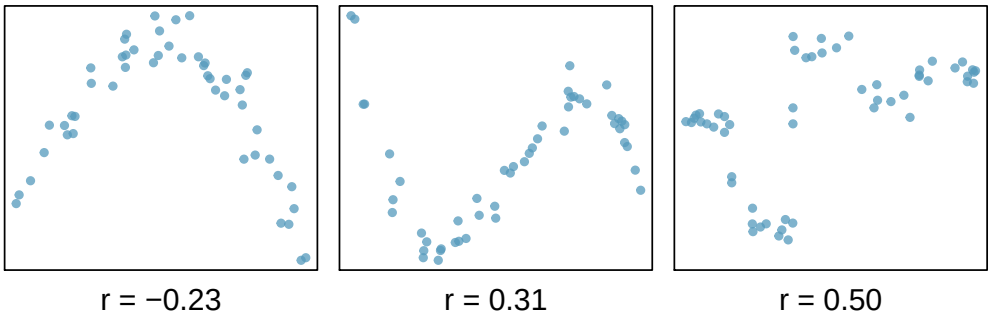

Figure 8.11: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

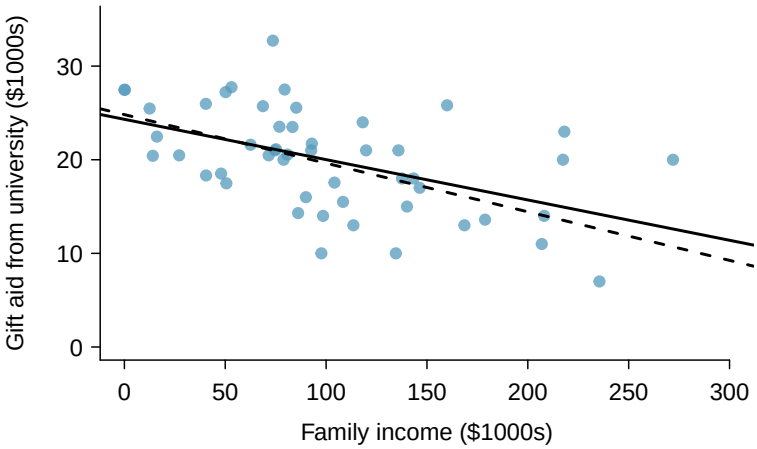

Figure 8.12: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

Appendix A

End of chapter exercise solutions

2 Summarizing data

2.1 (a) Positive association: mammals with longer gestation periods tend to live longer as well. (b) Association would still be positive. (c) No, they are not independent. See part (a).

8 Introduction to linear regression

8.1 (a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant. (b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.