Chapter 5

Introduction to linear regression

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where graphs with straight lines are overlaid on scatterplots. Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.

Figure 5.1 shows two variables whose relationship can be modeled perfectly with a straight line. The equation for the line is

$$y = 5 + 57.49x$$

Imagine what a perfect linear relationship would mean: you would know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process. For example, if we took family income x, this value would provide some useful information about how much financial support y a college may offer a prospective student. However, there would still be variability in financial support, even when comparing students whose families have similar financial backgrounds.

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

$$y = \beta_0 + \beta_1 x \tag{5.1}$$

β_0, β_1
Linear model parameters

where β_0 and β_1 represent two model parameters (β is the Greek letter *beta*). These parameters are estimated using data, and we write their point estimates as b_0 and b_1. When we use x to predict y, we usually call x the explanatory or **predictor** variable, and we call y the response.

It is rare for all of the data to fall on a straight line, as seen in the three scatterplots in Figure 5.2. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. The second plot shows an upward trend that, while evident, is not as strong as the first. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it. In each of these examples, we will have some uncertainty regarding our estimates of the model parameters, β_0 and β_1. For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less?

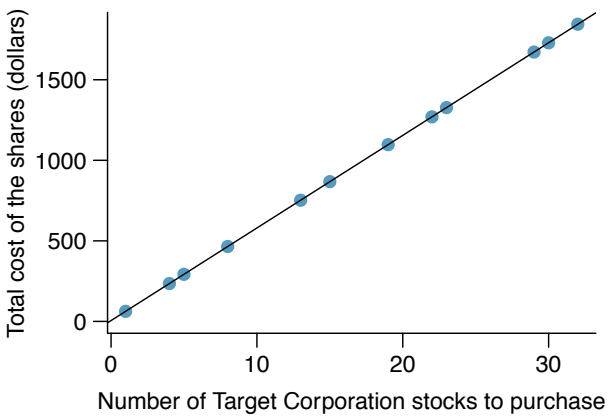

Figure 5.1: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker `TGT`, April 26th, 2012), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear fit is perfect.

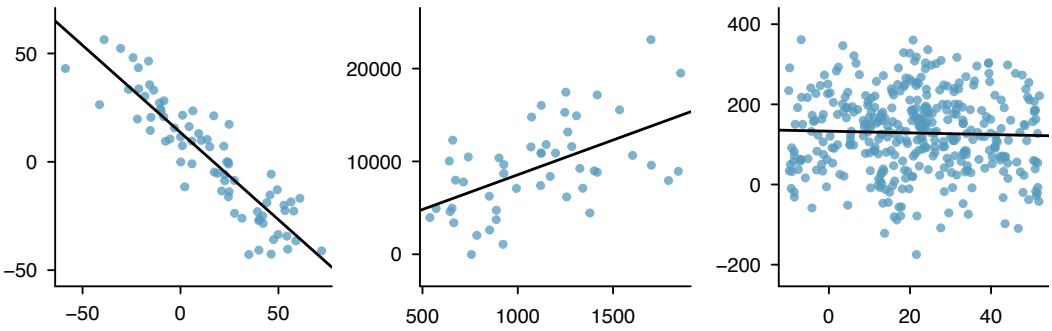

Figure 5.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

As we move forward in this chapter, we will learn different criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.

We will also see examples in this chapter where fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown in Figure 5.3 where there is a very strong relationship between the variables even though the trend is not linear. We will discuss nonlinear trends in this chapter and the next, but the details of fitting nonlinear models are saved for a later course.

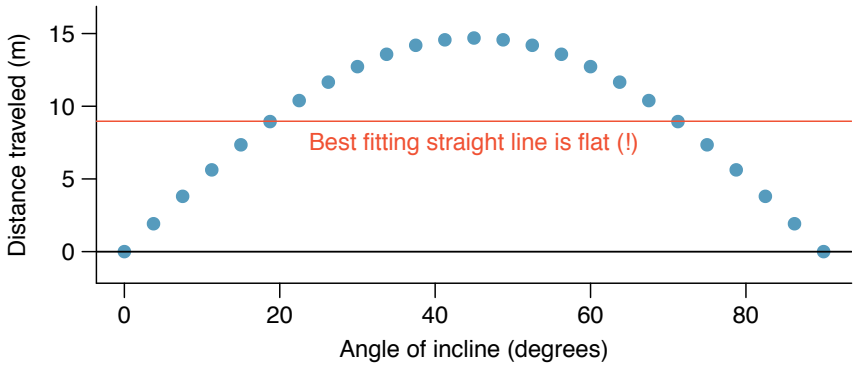

Figure 5.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

5.1 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*.

5.1.1 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 5.4 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 5.6. The right panel of Figure 5.6 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 5.6, more advanced techniques should be used.

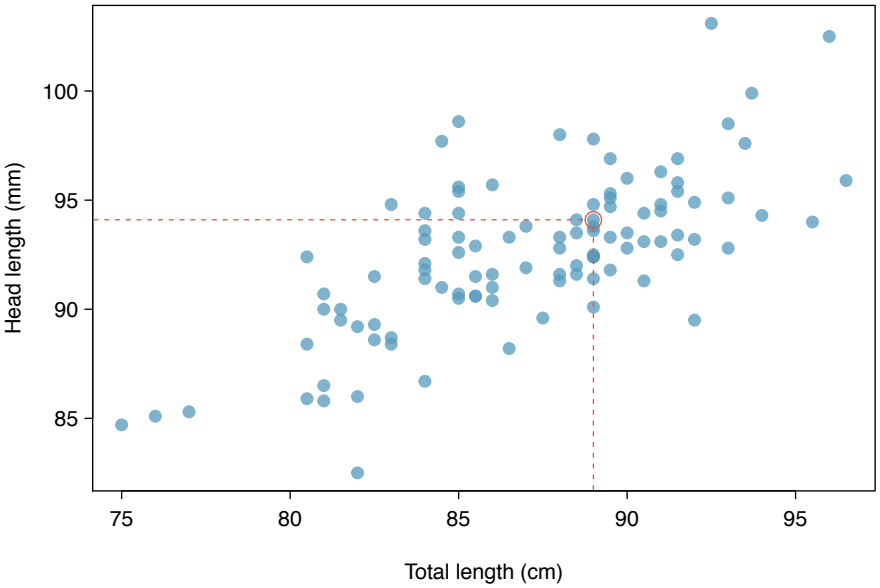

Figure 5.4: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

Figure 5.5: The common brushtail possum of Australia.

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

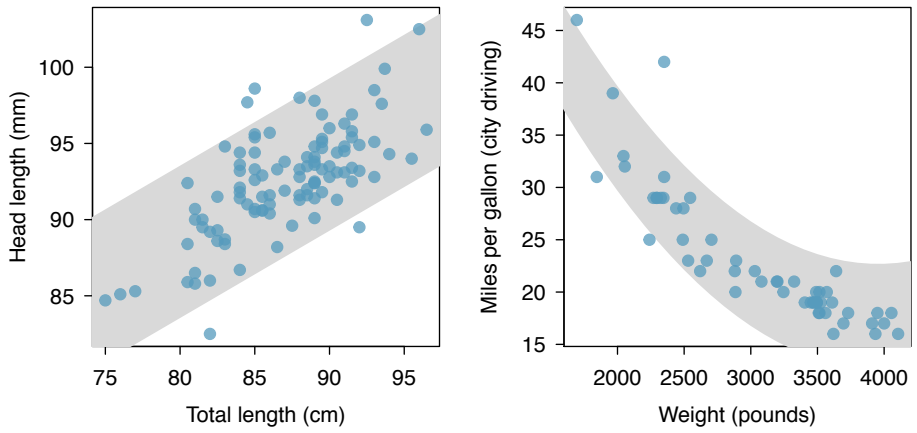

Figure 5.6: The figure on the left shows head length versus total length, and reveals that many of the points could be captured by a straight band. On the right, we see that a curved band is more appropriate in the scatterplot for `weight` and `mpgCity` from the `cars` data set.

5.1.2 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 5.7. The equation for this line is

$$\hat{y} = 41 + 0.59x \tag{5.2}$$

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.

5.1.3 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.

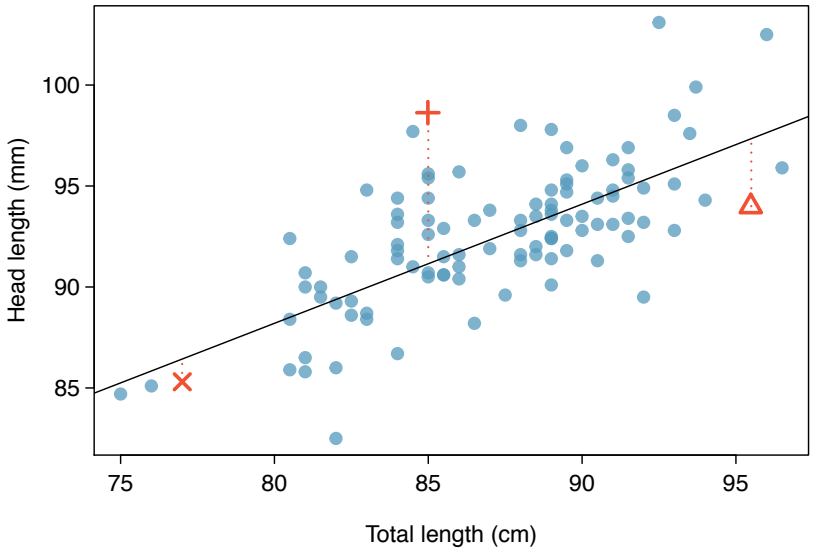

Figure 5.7: A reasonable linear model was fit to represent the relationship between head length and total length.

Three observations are noted specially in Figure 5.7. 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 $(\hat{y}_i)$:

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

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

Example 5.3 The linear fit shown in Figure 5.7 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:

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

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

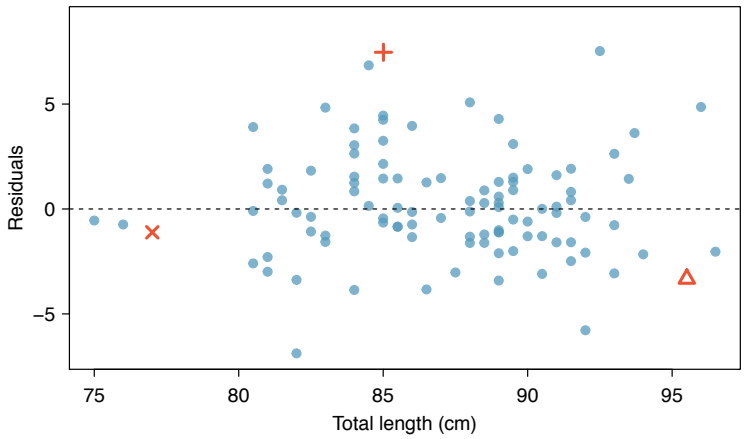

Figure 5.8: Residual plot for the model in Figure 5.7.

⊙ **Guided Practice 5.4** If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[1]

⊙ **Guided Practice 5.5** 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 5.8 for the regression line in Figure 5.7. 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.

● **Example 5.6** One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 5.9 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.

[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

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

This was close to the earlier estimate of 7.

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

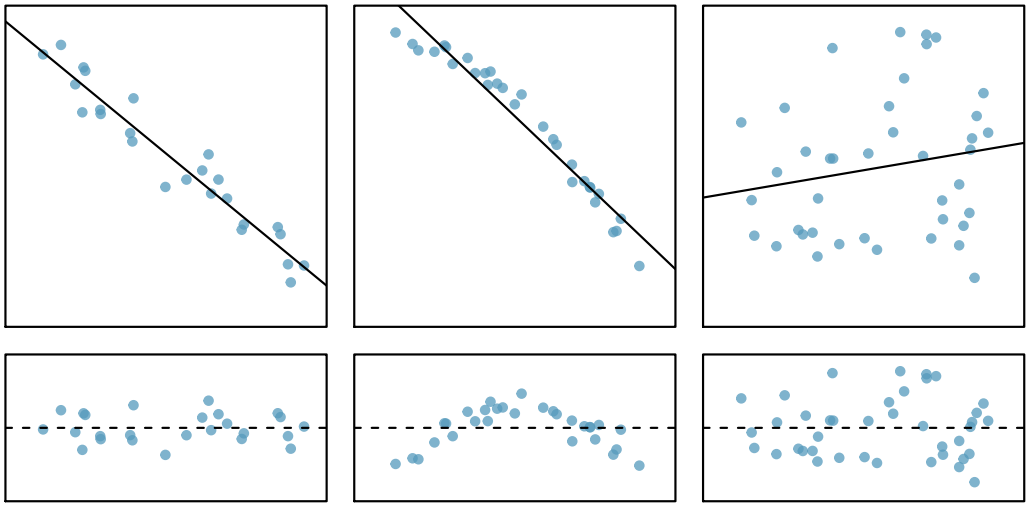

Figure 5.9: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

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 5.4.

5.1.4 Describing linear relationships with correlation

Correlation: strength of a linear relationship

R
correlation

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R.

We can compute the correlation 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. Figure 5.10 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,

[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.

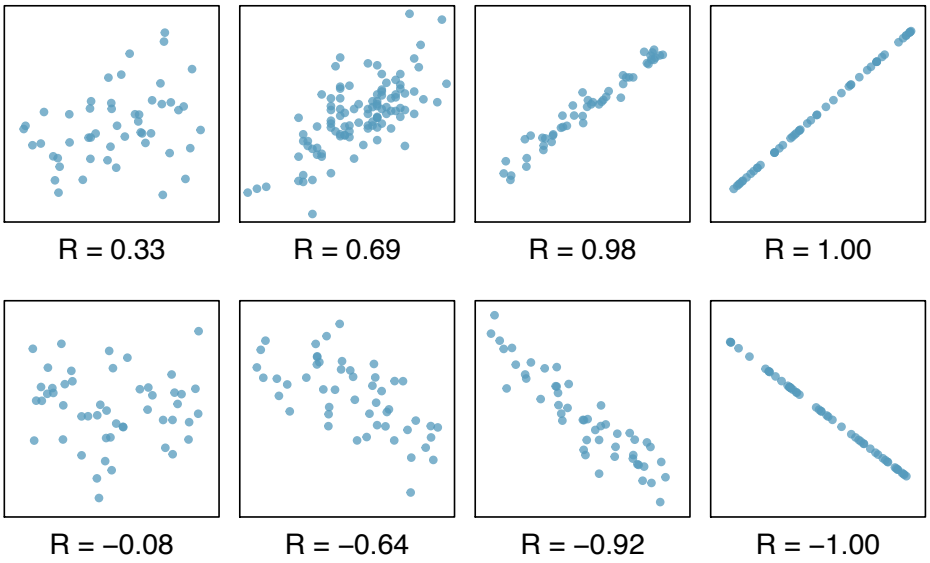

Figure 5.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 5.11.

⊙ **Guided Practice 5.7** It appears no straight line would fit any of the datasets represented in Figure 5.11. Instead, try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[4]

5.2 Fitting a line by least squares regression

Fitting linear models by eye is open to criticism since it is based on an individual preference. In this section, we use *least squares regression* as a more rigorous approach.

This section considers family income and gift aid data from a random sample of fifty students in the 2011 freshman class of Elmhurst College in Illinois.[5] Gift aid is financial aid that does not need to be paid back, as opposed to a loan. A scatterplot of the data is shown in Figure 5.12 along with two linear fits. The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university.

⊙ **Guided Practice 5.8** Is the correlation positive or negative in Figure 5.12?[6]

[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.

[5] These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435

[6] Larger family incomes are associated with lower amounts of aid, so the correlation will be negative. Using a computer, the correlation can be computed: -0.499.