13.2: Line Fitting, Residuals, and Correlation

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 13.2.1 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 discussed elsewhere. In this section, we examine criteria for identifying a linear model and introduce a new statistic, correlation.

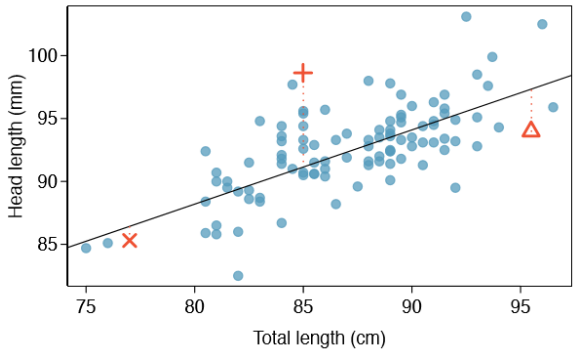

Figure 13.2.1: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

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 13.2.2 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.

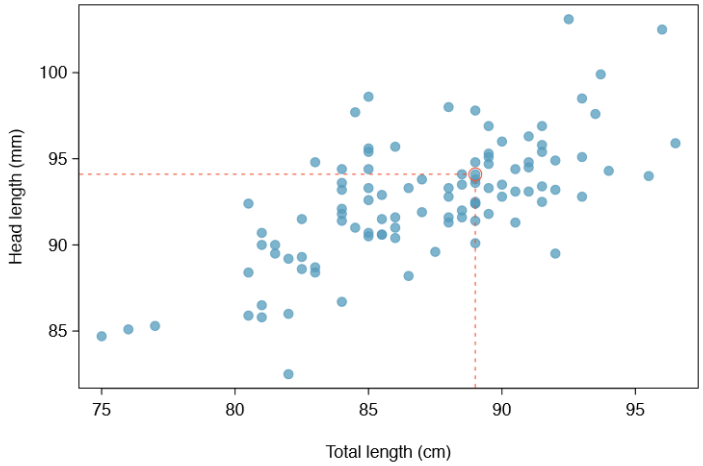

Figure 13.2.2: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89 cm is highlighted.

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.

Figure 13.2.3: The common brushtail possum of Australia. *Photo by wollombi on Flickr: www.ickr.com/photos/wollombi/58499575*

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 13.2.4. The right panel of Figure 13.2.4 shows a case where a curved line would be more useful in understanding the relationship between the two variables.

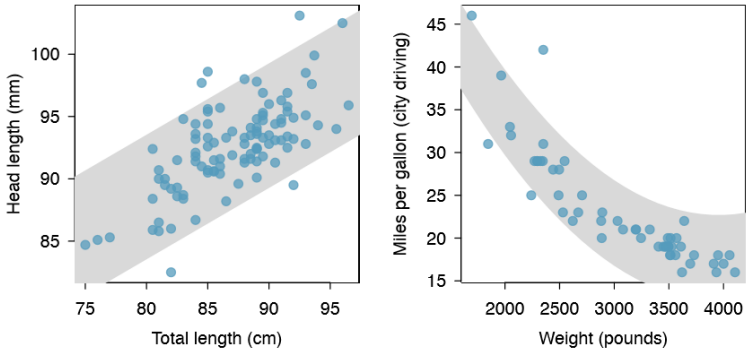

Figure 13.2.4: The figure on the left shows head length versus total length, and reveals that many of the points could be captured by a straight band. On the right, we see that a curved band is more appropriate in the scatterplot for weight and mpgCity from the cars data set.

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 13.2.4, more advanced techniques should be used.

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 13.2.5. The equation for this line is

$$\hat{y} = 41 + 0.59x \tag{7.2}$$

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

$$\hat{y} = 41 + 0.59 \times 80 \tag{13.2.1}$$

$$= 88.2 \tag{13.2.2}$$

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.

Residuals

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

$$\text{Data} = \text{Fit} + \text{Residual} \tag{13.2.3}$$

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.

Three observations are noted specially in Figure 13.2.5 The observation marked by an "X" has a small, negative residual of about -1; the observation marked by "+" has a large residual of about +7; and the observation marked by Δ 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 Δ is larger than that of "X" because | - 4| is larger than | - 1|.

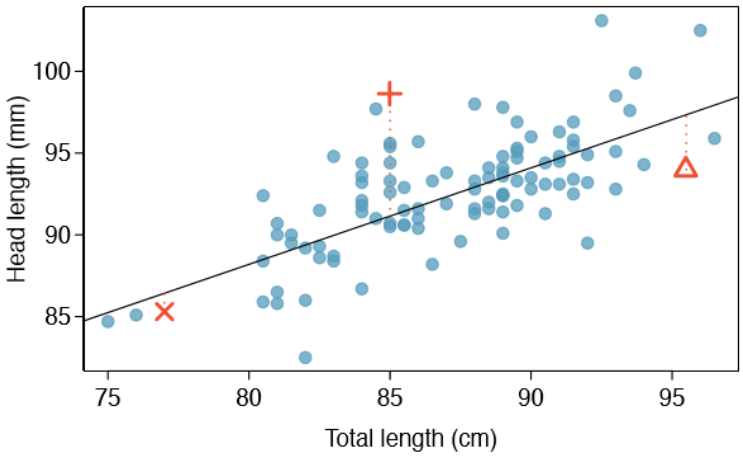

Figure 13.2.5: A reasonable linear model was to represent the relationship between head length and total length.

Residual: difference between observed and expected

The residual of the fifth 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 $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i \tag{13.2.4}$$

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

Example 13.2.1

The linear fit shown in Figure 13.2.5 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 "X" on the plot. Check it against the earlier visual estimate, -1.

Solution

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

$$\hat{y} = 41 + 0.59x_x = 41 + 0.59 \times 77.0 = 86.4 \tag{13.2.5}$$

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

$$e_x = y_x - \hat{y}_x = 85.3 - 86.4 = -1.1 \tag{13.2.6}$$

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

Exercise 13.2.1A

If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?

Answer

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.

Exercise 13.2.1B

Compute the residuals for the observations (85.0, 98.6) ("+" in Figure 13.2.5) and (95.5, 94.0) ("Δ") using the linear relationship

$$\hat{y} = 41 + 0.59x. \tag{13.2.7}$$

Answer

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

$$\hat{y}_+ = 41 + 0.59x_+ = 41 + 0.59 \times 85.0 = 91.15 \tag{13.2.8}$$

Then the residual is given by

$$e_+ = y_+ - \hat{y}_+ = 98.6 - 91.15 = 7.45 \tag{13.2.9}$$

This was close to the earlier estimate of 7.

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

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 13.2.6 for the regression line in Figure 13.2.5. 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.

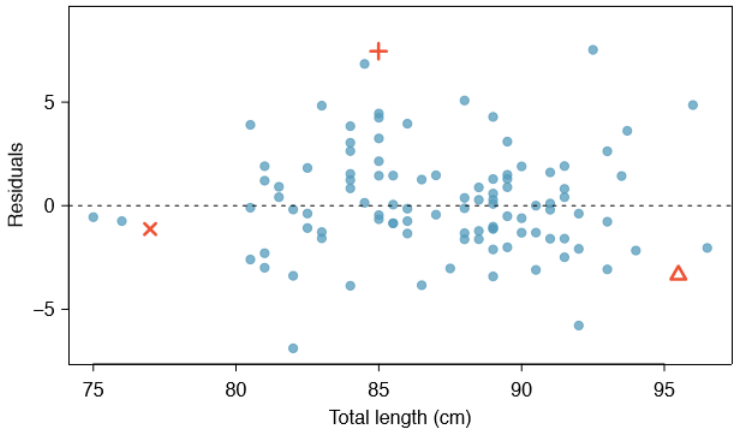

Figure 13.2.5.

Example 13.2.1

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 13.2.7 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?

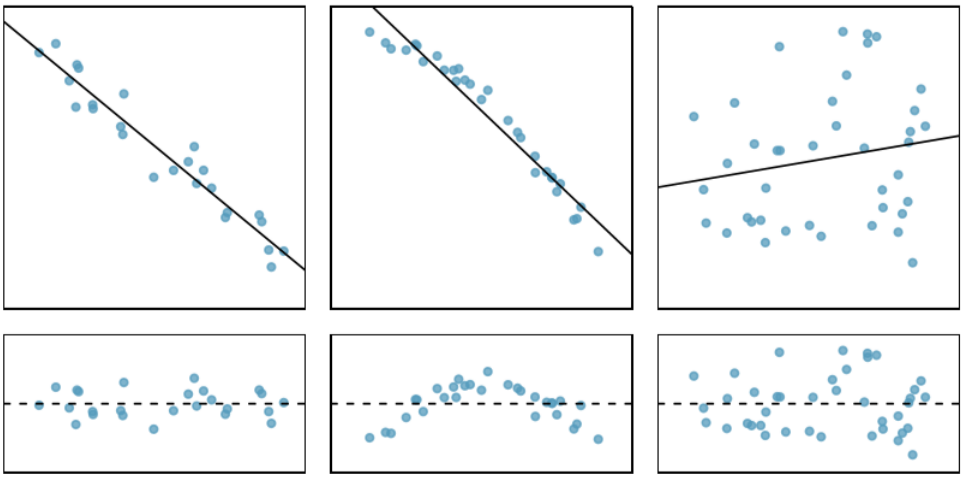

Figure 13.2.7: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

Solution

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 7.4.

Describing Linear Relationships with Correlation

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, so we generally perform the calculations on a computer or calculator. Figure 13.2.8 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.

Formally, we can compute the correlation for observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ using the formula

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y} \tag{13.2.10}$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

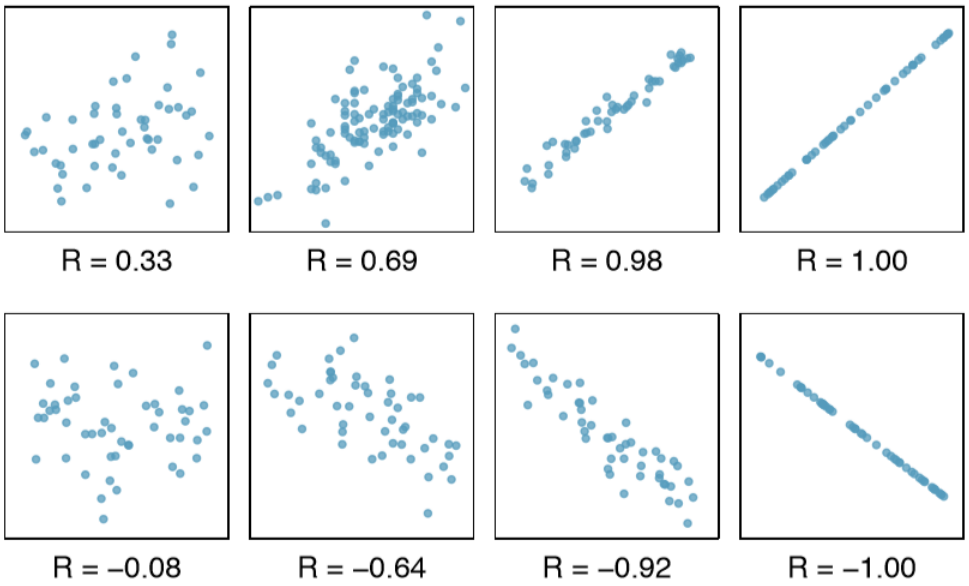

Figure 13.2.8: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

Correlation: strength of a linear relationship

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.

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 13.2.9.

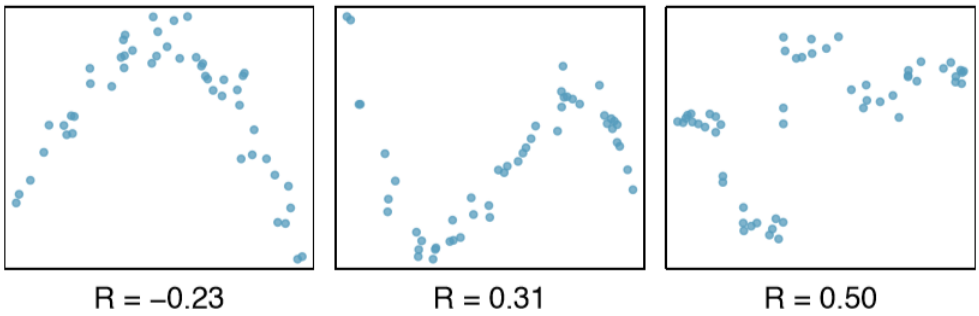

Figure 13.2.9: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

Exercise 13.2.1

It appears no straight line would fit any of the datasets represented in Figure 13.2.9. Try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[4]

Answer

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.

platform.

- **7.2: Line Fitting, Residuals, and Correlation** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.