values (x_i, y_i), $i = 1, \ldots, n$, of (X, Y) observed on each of n units or *cases*. In any particular problem, both X and Y will have other names such as *Temperature* or *Concentration* that are more descriptive of the data that is to be analyzed. The goal of regression is to understand how the values of Y change as X is varied over its range of possible values. A first look at how Y changes as X is varied is available from a scatterplot.

Inheritance of Height

One of the first uses of regression was to study inheritance of traits from generation to generation. During the period 1893–1898, E. S. Pearson organized the collection of $n = 1375$ heights of mothers in the United Kingdom under the age of 65 and one of their adult daughters over the age of 18. Pearson and Lee (1903) published the data, and we shall use these data to examine inheritance. The data are given in the data file `heights.txt`[1].

Our interest is in inheritance *from* the mother *to* the daughter, so we view the mother's height, called *Mheight*, as the predictor variable and the daughter's height, *Dheight*, as the response variable. Do taller mothers tend to have taller daughters? Do shorter mothers tend to have shorter daughters?

A scatterplot of *Dheight* versus *Mheight* helps us answer these questions. The scatterplot is a graph of each of the n points with the response *Dheight* on the vertical axis and predictor *Mheight* on the horizontal axis. This plot is shown in Figure 1.1. For regression problems with one predictor X and a response Y, we call the scatterplot of Y versus X a *summary graph*.

Here are some important characteristics of Figure 1.1:

1. The range of heights appears to be about the same for mothers and for daughters. Because of this, we draw the plot so that the lengths of the horizontal and vertical axes are the same, and the scales are the same. If all mothers and daughters had *exactly* the same height, then all the points would fall exactly on a 45° line. Some computer programs for drawing a scatterplot are not smart enough to figure out that the lengths of the axes should be the same, so you might need to resize the plot or to draw it several times.
2. The original data that went into this scatterplot was rounded so each of the heights was given to the nearest inch. If we were to plot the original data, we would have substantial *overplotting* with many points at exactly the same location. This is undesirable because we will not know if one point represents one case or many cases, and this can be very misleading. The easiest solution is to use *jittering*, in which a small uniform random number is added to each value. In Figure 1.1, we used a uniform random number on the range from -0.5 to $+0.5$, so the jittered values would round to the numbers given in the original source.
3. One important function of the scatterplot is to decide if we might reasonably assume that the response on the vertical axis is *independent* of the predictor

[1] See Appendix A.1 for instructions for getting data files from the Internet.

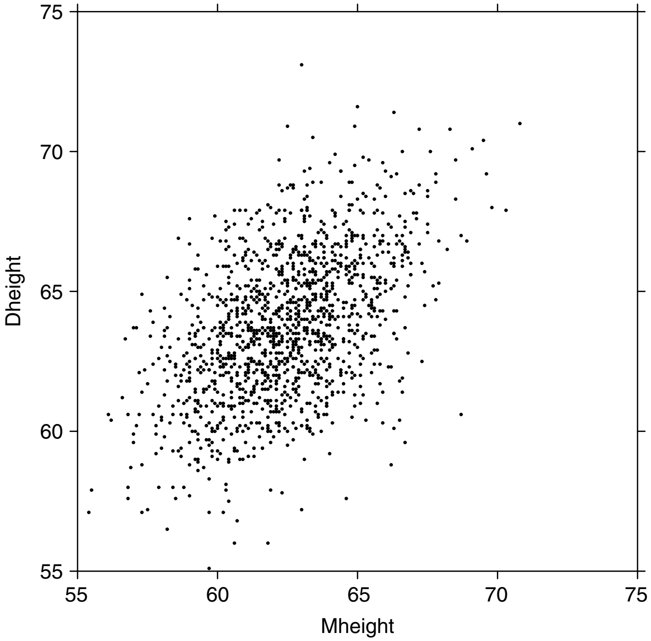

FIG. 1.1 Scatterplot of mothers' and daughters' heights in the Pearson and Lee data. The original data have been jittered to avoid overplotting, but if rounded to the nearest inch would return the original data provided by Pearson and Lee.

on the horizontal axis. This is clearly not the case here since as we move across Figure 1.1 from left to right, the scatter of points is different for each value of the predictor. What we mean by this is shown in Figure 1.2, in which we show only points corresponding to mother–daughter pairs with *Mheight* rounding to either 58, 64 or 68 inches. We see that within each of these three strips or *slices*, even though the number of points is different within each slice, (a) the mean of *Dheight* is increasing from left to right, and (b) the vertical variability in *Dheight* seems to be more or less the same for each of the fixed values of *Mheight*.

4. The scatter of points in the graph appears to be more or less elliptically shaped, with the axis of the ellipse tilted upward. We will see in Section 4.3 that summary graphs that look like this one suggest use of the simple linear regression model that will be discussed in Chapter 2.

5. Scatterplots are also important for finding *separated points*, which are either points with values on the horizontal axis that are well separated from the other points or points with values on the vertical axis that, given the value on the horizontal axis, are either much too large or too small. In terms of this example, this would mean looking for very tall or short mothers or, alternatively, for daughters who are very tall or short, given the height of their mother.

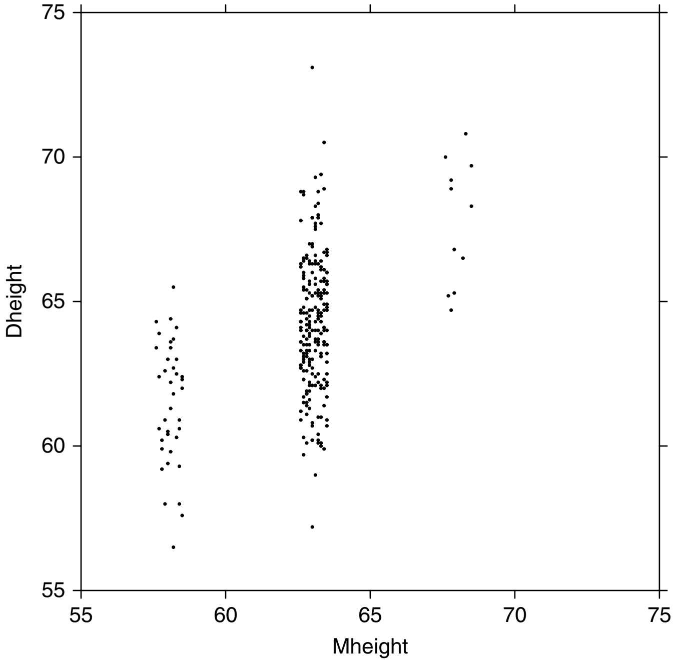

FIG. 1.2 Scatterplot showing only pairs with mother's height that rounds to 58, 64 or 68 inches.

These two types of separated points have different names and roles in a regression problem. Extreme values on the left and right of the horizontal axis are points that are likely to be important in fitting regression models and are called *leverage* points. The separated points on the vertical axis, here unusually tall or short daughters give their mother's height, are potentially *outliers*, cases that are somehow different from the others in the data.

While the data in Figure 1.1 do include a few tall and a few short mothers and a few tall and short daughters, given the height of the mothers, none appears worthy of special treatment, mostly because in a sample size this large we expect to see some fairly unusual mother–daughter pairs.

We will continue with this example later.

Forbes' Data

In an 1857 article, a Scottish physicist named James D. Forbes discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water. He knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. In the middle of the nineteenth century, barometers were fragile instruments, and Forbes wondered if a simpler measurement of the boiling point of water could substitute for a direct reading of barometric pressure. Forbes

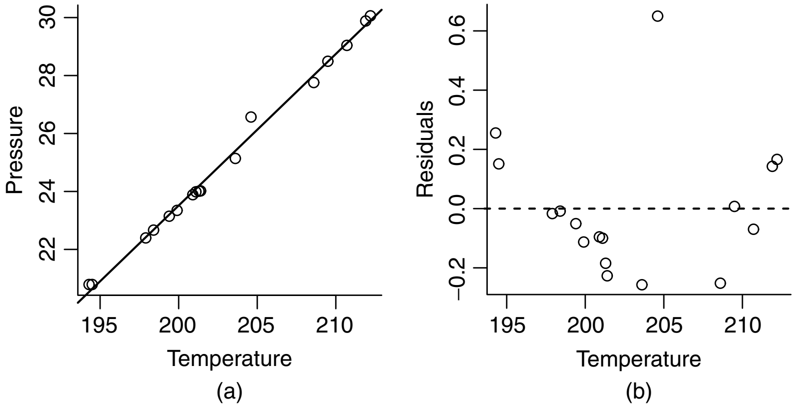

FIG. 1.3 Forbes data. (a) *Pressure* versus *Temp*; (b) Residuals versus *Temp*.

collected data in the Alps and in Scotland. He measured at each location pressure in inches of mercury with a barometer and boiling point in degrees Fahrenheit using a thermometer. Boiling point measurements were adjusted for the difference between the ambient air temperature when he took the measurements and a standard temperature. The data for $n = 17$ locales are reproduced in the file `forbes.txt`.

The scatterplot of *Pressure* versus *Temp* is shown in Figure 1.3a. The general appearance of this plot is very different from the summary graph for the heights data. First, the sample size is only 17, as compared to over 1300 for the heights data. Second, apart from one point, all the points fall almost exactly on a smooth curve. This means that the variability in pressure for a given temperature is extremely small.

The points in Figure 1.3a appear to fall very close to the straight line shown on the plot, and so we might be encouraged to think that the mean of pressure given temperature could be modelled by a straight line. Look closely at the graph, and you will see that there is a small systematic error with the straight line: apart from the one point that does not fit at all, the points in the middle of the graph fall below the line, and those at the highest and lowest temperatures fall above the line. This is much easier to see in Figure 1.3b, which is obtained by removing the linear trend from Figure 1.3a, so the plotted points on the vertical axis are given for each value of *Temp* by

$$\mathit{Residual} = \mathit{Pressure} - \text{ point on the line}$$

This allows us to gain resolution in the plot since the range on the vertical axis in Figure 1.3a is about 10 inches of mercury while the range in Figure 1.3b is about 0.8 inches of mercury. To get the same resolution in Figure 1.3a, we would need a graph that is $10/0.8 = 12.5$ as big as Figure 1.3b. Again ignoring the one point that clearly does not match the others, the curvature in the plot is clearly visible in Figure 1.3b.

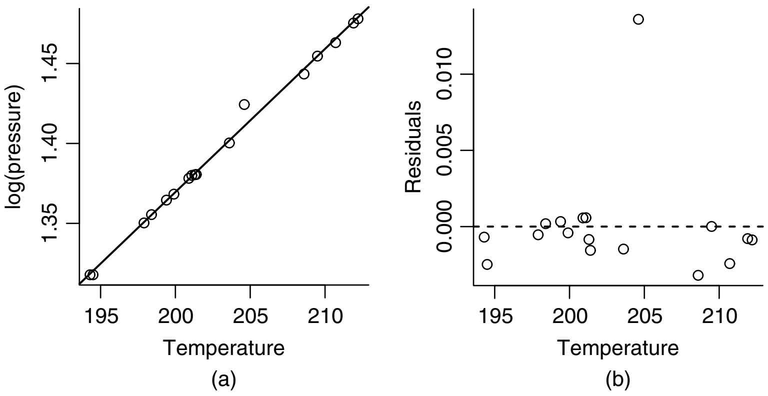

FIG. 1.4 (a) Scatterplot of Forbes' data. The line shown is the OLS line for the regression of log(*Pressure*) on *Temp*. (b) Residuals versus *Temp*.

While there is nothing at all wrong with curvature, the methods we will be studying in this book work best when the plot can be summarized by a straight line. Sometimes we can get a straight line by transforming one or both of the plotted quantities. Forbes had a physical theory that suggested that log(*Pressure*) is linearly related to *Temp*. Forbes (1857) contains what may be the first published summary graph corresponding to his physical model. His figure is redrawn in Figure 1.4. Following Forbes, we use base ten common logs in this example, although in most of the examples in this book we will use base-two logarithms. The choice of base has no material effect on the appearance of the graph or on fitted regression models, but interpretation of parameters can depend on the choice of base, and using base-two often leads to a simpler interpretation for parameters.

The key feature of Figure 1.4a is that apart from one point the data appear to fall very close to the straight line shown on the figure, and the residual plot in Figure 1.4b confirms that the deviations from the straight line are not systematic the way they were in Figure 1.3b. All this is evidence that the straight line is a reasonable summary of these data.

Length at Age for Smallmouth Bass

The smallmouth bass is a favorite game fish in inland lakes. Many smallmouth bass populations are managed through stocking, fishing regulations, and other means, with a goal to maintain a healthy population.

One tool in the study of fish populations is to understand the growth pattern of fish such as the dependence of a measure of size like fish length on age of the fish. Managers could compare these relationships between different populations with dissimilar management plans to learn how management impacts fish growth.

Figure 1.5 displays the *Length* at capture in mm versus *Age* at capture for $n =$ 439 small mouth bass measured in West Bearskin Lake in Northeastern Minnesota in 1991. Only fish of age seven or less are included in this graph. The data were provided by the Minnesota Department of Natural Resources and are given in the

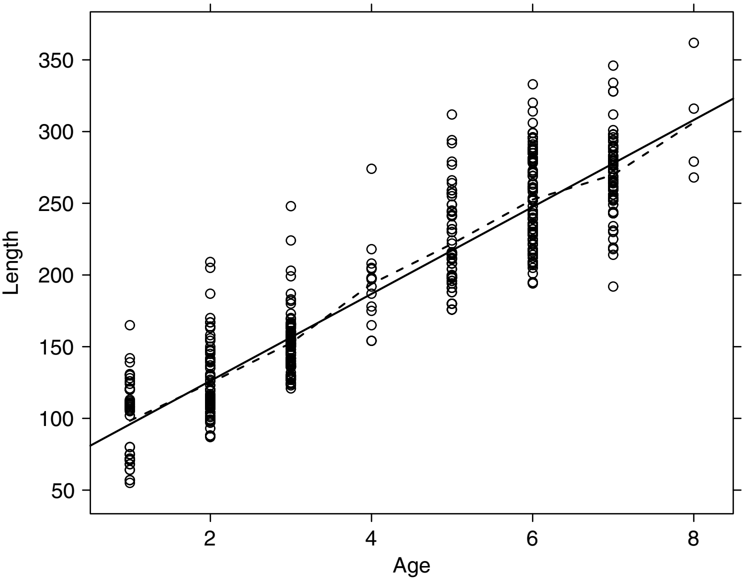

FIG. 1.5 *Length* (mm) versus *Age* for West Bearskin Lake smallmouth bass. The solid line shown was estimated using ordinary least squares or OLS. The dashed line joins the average observed length at each age.

file `wblake.txt`. Fish scales have annular rings like trees, and these can be counted to determine the age of a fish. These data are *cross-sectional*, meaning that all the observations were taken at the same time. In a *longitudinal* study, the same fish would be measured each year, possibly requiring many years of taking measurements. The data file gives the *Length* in mm, *Age* in years, and the *Scale* radius, also in mm.

The appearance of this graph is different from the summary plots shown for last two examples. The predictor *Age* can only take on integer values corresponding to the number of annular rings on the scale, so we are really plotting seven distinct populations of fish. As might be expected, length generally increases with age, but the longest fish at age-one fish exceeds the length of the shortest age-four fish, so knowing the age of a fish will not allow us to predict its length exactly; see Problem 2.5.

Predicting the Weather

Can early season snowfall from September 1 until December 31 predict snowfall in the remainder of the year, from January 1 to June 30? Figure 1.6, using data from the data file `ftcollinssnow.txt`, gives a plot of *Late* season snowfall from January 1 to June 30 versus *Early* season snowfall for the period September 1 to December 31 of the previous year, both measured in inches at Ft. Collins, Colorado[2]. If *Late* is related to *Early*, the relationship is considerably weaker than

[2] The data are from the public domain source http://www.ulysses.atmos.colostate.edu.

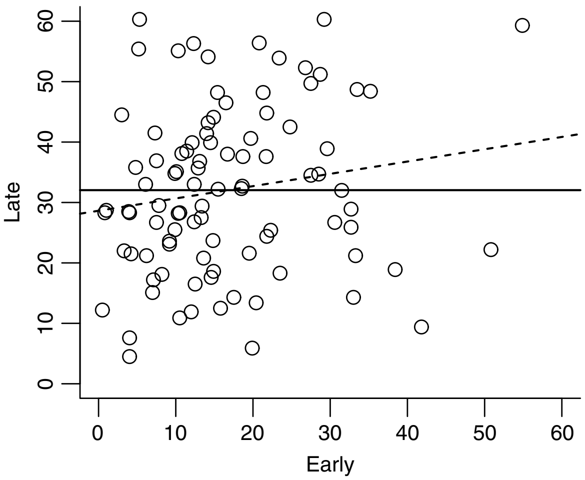

FIG. 1.6 Plot of snowfall for 93 years from 1900 to 1992 in inches. The solid horizontal line is drawn at the average late season snowfall. The dashed line is the best fitting (ordinary least squares) line of arbitrary slope.

in the previous examples, and the graph suggests that early winter snowfall and late winter snowfall may be completely unrelated, or *uncorrelated*. Interest in this regression problem will therefore be in testing the hypothesis that the two variables are uncorrelated versus the alternative that they are not uncorrelated, essentially comparing the fit of the two lines shown in Figure 1.6. Fitting models will be helpful here.

Turkey Growth

This example is from an experiment on the growth of turkeys (Noll, Weibel, Cook, and Witmer, 1984). Pens of turkeys were grown with an identical diet, except that each pen was supplemented with a *Dose* of the amino acid methionine as a percentage of the total diet of the birds. The methionine was provided using either a standard source or one of two experimental sources. The response is average weight gain in grams of all the turkeys in the pen.

Figure 1.7 provides a summary graph based on the data in the file `turkey.txt`. Except at $Dose = 0$, each point in the graph is the average response of five pens of turkeys; at $Dose = 0$, there were ten pens of turkeys. Because averages are plotted, the graph does not display the variation between pens treated alike. At each value of $Dose > 0$, there are three points shown, with different symbols corresponding to the three sources of methionine, so the variation between points at a given *Dose* is really the variation between sources. At $Dose = 0$, the point has been arbitrarily labelled with the symbol for the first group, since $Dose = 0$ is the same treatment for all sources.

For now, ignore the three sources and examine Figure 1.7 in the way we have been examining the other summary graphs in this chapter. Weight gain seems

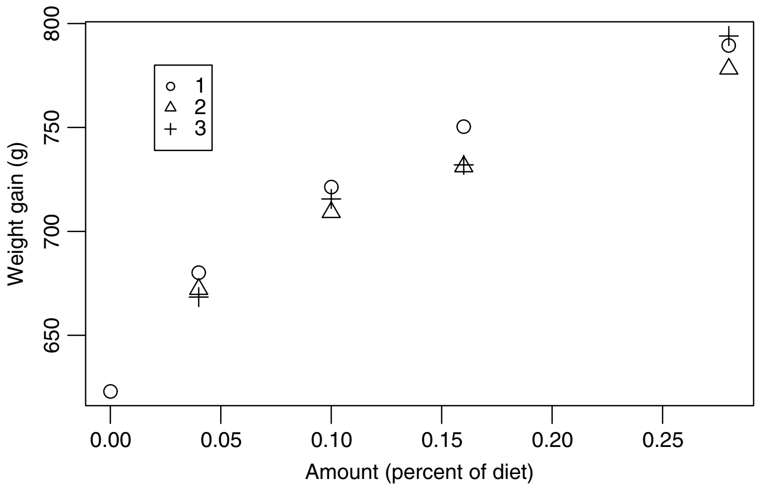

FIG. 1.7 Weight gain versus *Dose* of methionine for turkeys. The three symbols for the points refer to three different sources of methionine.

to increase with increasing *Dose*, but the increase does not appear to be linear, meaning that a straight line does not seem to be a reasonable representation of the average dependence of the response on the predictor. This leads to study of mean functions.

1.2 MEAN FUNCTIONS

Imagine a generic summary plot of Y versus X. Our interest centers on how the distribution of Y changes as X is varied. One important aspect of this distribution is the *mean function*, which we define by

$$\mathrm{E}(Y|X=x) = \text{a function that depends on the value of } x \qquad (1.1)$$

We read the left side of this equation as "the expected value of the response when the predictor is fixed at the value $X = x$;" if the notation "E( )" for expectations and "Var( )" for variances is unfamiliar, please read Appendix A.2. The right side of (1.1) depends on the problem. For example, in the heights data in Example 1.1, we might believe that

$$\mathrm{E}(\mathit{Dheight}|\mathit{Mheight}=x) = \beta_0 + \beta_1 x \qquad (1.2)$$

that is, the mean function is a straight line. This particular mean function has two *parameters*, an intercept β_0 and a slope β_1. If we knew the values of the βs, then the mean function would be completely specified, but usually the βs need to be estimated from data.

Figure 1.8 shows two possibilities for βs in the straight-line mean function (1.2) for the heights data. For the dashed line, $\beta_0 = 0$ and $\beta_1 = 1$. This mean function