

Scatterplots and Regression

Regression is the study of dependence. It is used to answer interesting questions about how one or more predictors influence a response. Here are a few typical questions that may be answered using regression:

- Are daughters taller than their mothers?
- Does changing class size affect success of students?
- Can we predict the time of the next eruption of Old Faithful Geyser from the length of the most recent eruption?
- Do changes in diet result in changes in cholesterol level, and if so, do the results depend on other characteristics such as age, sex, and amount of exercise?
- Do countries with higher per person income have lower birth rates than countries with lower income?
- Are highway design characteristics associated with highway accident rates? Can accident rates be lowered by changing design characteristics?
- Is water usage increasing over time?
- Do conservation easements on agricultural property lower land value?

In most of this book, we study the important instance of regression methodology called *linear regression*. This method is the most commonly used in regression, and virtually all other regression methods build upon an understanding of how linear regression works.

As with most statistical analyses, the goal of regression is to summarize observed data as simply, usefully, and elegantly as possible. A theory may be available in some problems that specifies how the response varies as the values

of the predictors change. If theory is lacking, we may need to use the data to help us decide on how to proceed. In either case, an essential first step in regression analysis is to draw appropriate graphs of the data.

We begin in this chapter with the fundamental graphical tools for studying dependence. In regression problems with one predictor and one response, the *scatterplot* of the response versus the predictor is the starting point for regression analysis. In problems with many predictors, several simple graphs will be required at the beginning of an analysis. A *scatterplot matrix* is a convenient way to organize looking at many scatterplots at once. We will look at several examples to introduce the main tools for looking at scatterplots and scatterplot matrices and extracting information from them. We will also introduce notation that will be used throughout the book.

1.1 SCATTERPLOTS

We begin with a regression problem with one predictor, which we will generically call X , and one response variable, which we will call Y .¹ Data consist of values (x_i, y_i) , $i = 1, \dots, n$, of (X, Y) observed on each of n units or *cases*. In any particular problem, both X and Y will have other names that will be displayed in this book using typewriter font, such as temperature or concentration, that are more descriptive of the data that are 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, Karl Pearson (1857–1936) 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`.²

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

¹In some disciplines, predictors are called independent variables, and the response is called a dependent variable, terms not used in this book.

²See Appendix A.1 for instructions for getting data files from the Internet.

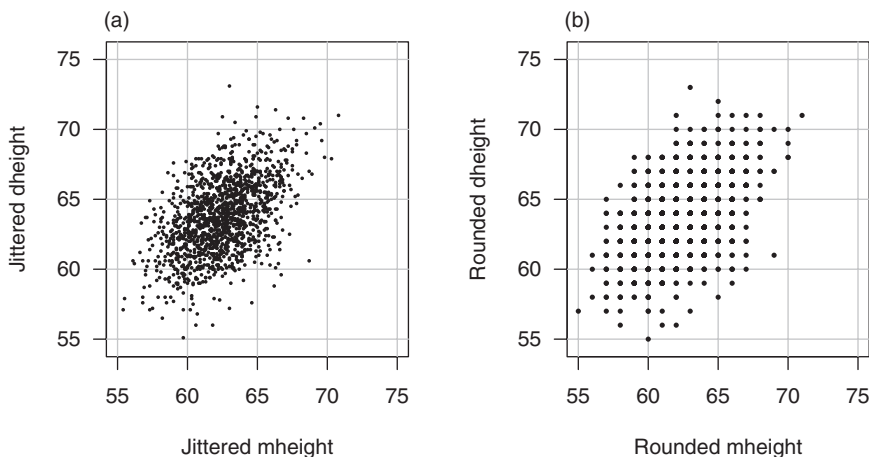


Figure 1.1 Scatterplot of mothers' and daughters' heights in the Pearson and Lee data. The original data have been jittered to avoid overplotting in (a). Plot (b) shows the original data, so each point in the plot refers to one or more mother–daughter pairs.

shown in Figure 1.1a. 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 this scatterplot:

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 pairs 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 were rounded so each of the heights was given to the nearest inch. The original data are plotted in Figure 1.1b. This plot exhibits substantial *overplotting* with many points at exactly the same location. This is undesirable because one point on the plot can correspond to many cases. The easiest solution is to use *jittering*, in which a small uniform random number is added to each value. In Figure 1.1a, 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 on the horizontal axis. This is clearly not the case here since as we move across Figure 1.1a from left to right, the scatter of points is

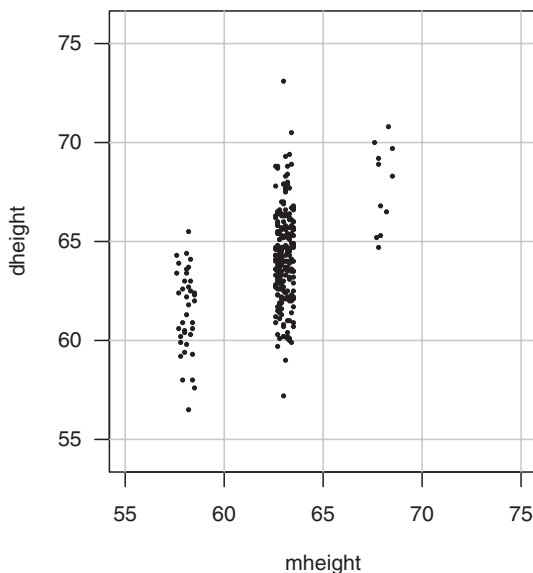


Figure 1.2 Scatterplot showing only pairs with mother's height that rounds to 58, 64, or 68 inches.

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*, the number of points is different, and the mean of `dheight` is increasing from left to right. The vertical variability in `dheight` seems to be more or less the same for each of the fixed values of `mheight`.

4. In Figure 1.1a the scatter of points appears to be more or less elliptically shaped, with the major axis of the ellipse tilted upward, and with more points near the center of the ellipse rather than on the edges. We will see in Section 1.4 that summary graphs that look like this one suggest the use of the simple linear regression model that will be discussed in Chapter 2.
5. Scatterplots are also important for finding *separated points*. Horizontal separation would occur for a value on the horizontal axis `mheight` that is either unusually small or unusually large relative to the other values of `mheight`. Vertical separation would occur for a daughter with `dheight` either relatively large or small compared with the other daughters with about the same value for `mheight`.

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. Outliers are more easily discovered in residual plots, as illustrated in the next example.

While the data in Figure 1.1a 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.

Forbes’s Data

In an 1857 article, the Scottish physicist James D. Forbes (1809–1868) 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. Barometers in the middle of the nineteenth century 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 collected data in the Alps and in Scotland. He measured at each location the atmospheric pressure `pres` in inches of mercury with a barometer and boiling point `bp` 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`.

The scatterplot of `pres` versus `bp` 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 with over 1,300 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 boiling point is extremely small.

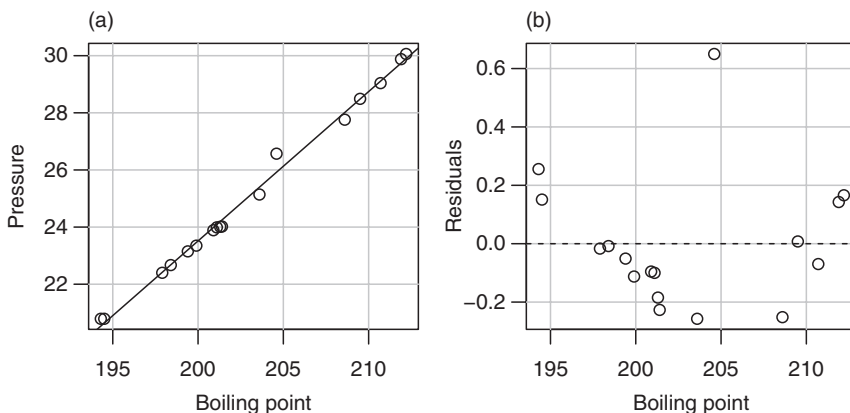


Figure 1.3 Forbes data: (a) `pres` versus `bp`; (b) residuals versus `bp`.

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 boiling point could be modeled by a straight line. Look closely at the graph, and you will see that there is a small systematic deviation from 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 boiling points 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 bp by

$$\text{residual} = \text{pres} - \text{point on the line}$$

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.

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 $\log(\text{pres})$ is linearly related to bp . Forbes (1857) contains what may be the first published summary graph based on 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 natural logarithms. The choice

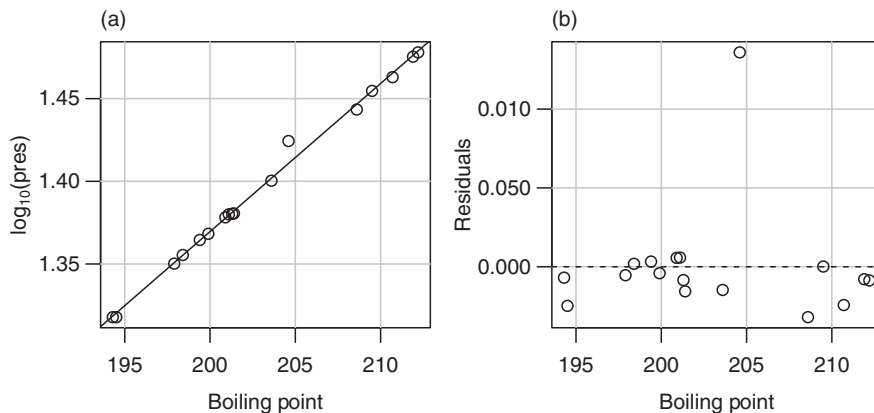


Figure 1.4 (a) Scatterplot of Forbes's data. The line shown is the OLS line for the regression of $\log(\text{pres})$ on bp . (b) Residuals versus bp .

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.

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 that are managed differently to learn how management impacts fish growth.

Figure 1.5 displays the Length at capture in mm versus Age at capture for $n = 439$ smallmouth bass measured in West Bearskin Lake in Northeastern Minnesota in 1991. Only fish of age 8 or less are included in this graph. The data were provided by the Minnesota Department of Natural Resources and are given in the file `wblake`. Similar to trees, the scales of many fish species have annular rings, and these can be counted to determine the age of a fish.

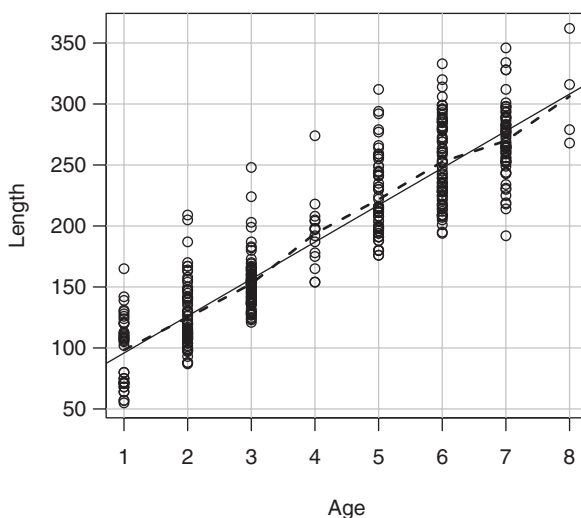


Figure 1.5 Length (mm) versus Age for West Bearskin Lake smallmouth bass. The solid line shown was estimated using ordinary least squares or ols. The dashed line joins the average observed length at each age.

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 appearance of this graph is different from the summary graphs shown for the 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 eight distinct populations of fish. As might be expected, length generally increases with age, but the length of the longest fish at age 1 exceeds the length of the shortest fish at age 4, so knowing the age of a fish will not allow us to predict its length exactly; see Problem 2.15.

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`, 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 (Colorado Climate Center, 2012). If Late is related to Early, the relationship is considerably weaker than 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

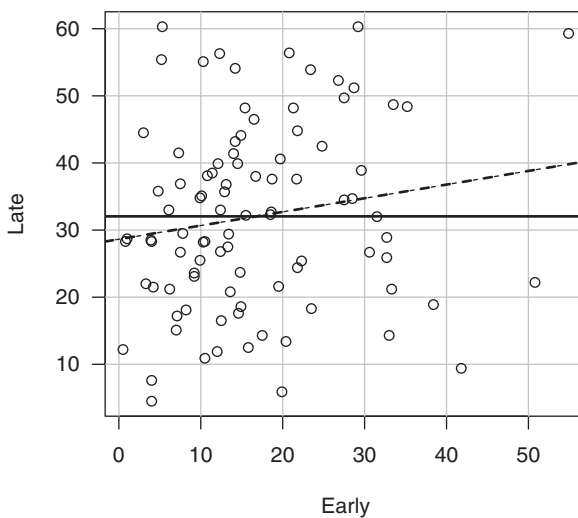


Figure 1.6 Plot of snowfall for 93 years from 1900 to 1992 in inches. The solid horizontal line is drawn at the average late season snowfall. The dashed line is the OLS line.

comparing the fit of the two lines shown in Figure 1.6. Fitting models and performing tests will be helpful here.

Turkey Growth

This example is from an experiment on the growth of turkeys (Noll et al., 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`. Except at `Dose = 0`, each point in the graph is the average response of five pens of turkeys; at `Dose = 0`, there were 10 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 labeled 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 is seen 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

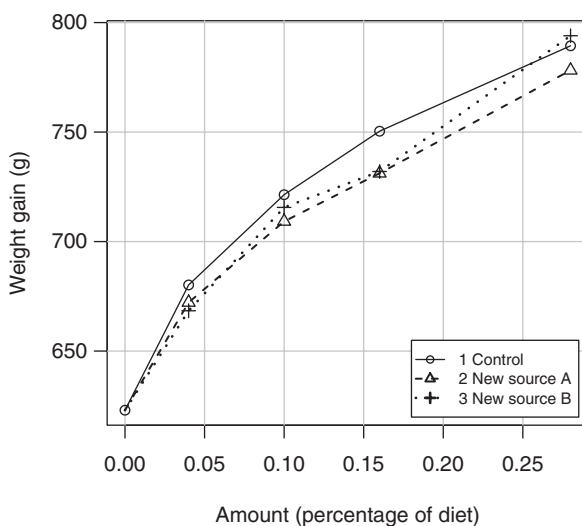


Figure 1.7 Weight gain versus `Dose` of methionine for turkeys. The three symbols for the points refer to sources of methionine. The lines on the plot join the means within a source.

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

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

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 “ $\text{Var}(\)$ ” for variances is unfamiliar, refer to 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

$$E(\text{dheight}|\text{mheight} = x) = \beta_0 + \beta_1 x \quad (1.2)$$

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 β s, then the mean function would be completely specified, but usually the β s need to be estimated from data. These parameters are discussed more fully in the next chapter.

Figure 1.8 shows two possibilities for the β s 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 would suggest that daughters have the same height as their mothers on the average for mothers of any height. The second line is estimated using ordinary least squares, or OLS, the estimation method that will be described in the next chapter. The OLS line has slope less than 1, meaning that tall mothers tend to have daughters who are taller than average because the slope is positive, but shorter than themselves because the slope is less than 1. Similarly, short mothers tend to have short daughters but taller than themselves. This is perhaps a surprising result and is the origin of the term *regression*, since extreme values in one generation tend to revert or regress toward the population mean in the next generation (Galton, 1886).

Two lines are shown in Figure 1.5 for the smallmouth bass data. The dashed line joins the average length at each age. It provides an estimate of the mean function $E(\text{Length}|\text{Age})$ without actually specifying any functional form for the mean function. We will call this a *nonparametric* estimated mean function; sometimes we will call it a *smoother*. The solid line is the OLS estimated straight line (1.1) for the mean function. Perhaps surprisingly, the straight line and the dashed lines that join the within-age means appear to agree very closely, and we might be encouraged to use the straight-line mean function to describe

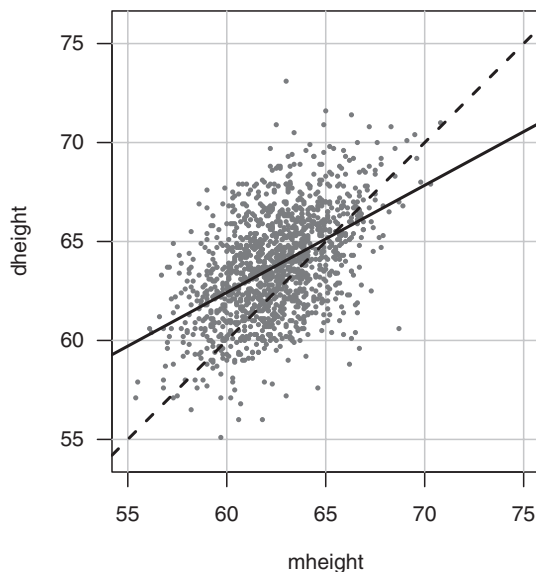


Figure 1.8 The heights data. The dashed line is for $E(dheight|mheight) = mheight$, and the solid line is estimated by OLS.

these data. The increase in length per year is modeled to be the same for all ages. We cannot expect this to be true if we were to include older-aged fish because eventually the growth must slow down. For the range of ages here, the approximation seems to be adequate.

For the Ft. Collins weather data, we might expect the straight-line mean function (1.1) to be appropriate but with $\beta_1 = 0$. If the slope is 0, then the mean function is parallel to the horizontal axis, as shown in Figure 1.6. We will eventually test for independence of `Early` and `Late` by testing the hypothesis that $\beta_1 = 0$ against the alternative hypothesis that $\beta_1 \neq 0$.

Not all summary graphs will have a straight-line mean function. In Forbes's data, to achieve linearity we have replaced the measured value of `pres` by $\log(\text{pres})$. Transformation of variables will be a key tool in extending the usefulness of linear regression models. In the turkey data and other growth models, a nonlinear mean function might be more appropriate, such as

$$E(Y|\text{Dose} = x) = \beta_0 + \beta_1[1 - \exp(-\beta_2 x)] \quad (1.3)$$

The β s in (1.3) have a useful interpretation, and they can be used to summarize the experiment. When `Dose` = 0, $E(Y|\text{Dose} = 0) = \beta_0$, so β_0 is the baseline growth without supplementation. Assuming $\beta_2 > 0$, when the `Dose` is large, $\exp(-\beta_2 \text{Dose})$ is small, and so $E(Y|\text{Dose})$ approaches $\beta_0 + \beta_1$ for larger values of `Dose`. We think of $\beta_0 + \beta_1$ as the limit to growth with this additive. The rate

parameter β_2 determines how quickly maximum growth is achieved. This three-parameter mean function will be considered in Chapter 11.

1.3 VARIANCE FUNCTIONS

Another characteristic of the distribution of the response given the predictor is the *variance function*, defined by the symbol $\text{Var}(Y|X = x)$ and in words as the variance of the response given that the predictor is fixed at $X = x$. For example, in Figure 1.2 we can see that the variance function for `dheight|lmheight` is approximately the same for each of the three values of `mheight` shown in the graph. In the `smallmouth bass` data in Figure 1.5, an assumption that the variance is constant across the plot is plausible, even if it is not certain (see Problem 1.2). In the `turkey` data, we cannot say much about the variance function from the summary plot because we have plotted treatment means rather than the actual pen values, so the graph does not display the information about the variability between pens that have a fixed value of `Dose`.

A frequent assumption in fitting linear regression models is that the variance function is the same for every value of x . This is usually written as

$$\text{Var}(Y|X = x) = \sigma^2 \quad (1.4)$$

where σ^2 (read “sigma squared”) is a generally unknown positive constant. Chapter 7 presents more general variance models.

1.4 SUMMARY GRAPH

In all the examples except the `snowfall` data, there is a clear dependence of the response on the predictor. In the `snowfall` example, there might be no dependence at all. The `turkey` growth example is different from the others because the average value of the response seems to change nonlinearly with the value of the predictor on the horizontal axis.

The scatterplots for these examples are all typical of graphs one might see in problems with one response and one predictor. Examination of the summary graph is a first step in exploring the relationships these graphs portray.

Anscombe (1973) provided the artificial data given in the file `anscombe` that consists of 11 pairs of points (x_i, y_i) , to which the simple linear regression mean function $E(y|x) = \beta_0 + \beta_1 x$ is fit. Each data set leads to an identical summary analysis with the same estimated slope, intercept, and other summary statistics, but the visual impression of each of the graphs is very different. The first example in Figure 1.9a is as one might expect to observe if the simple linear regression model were appropriate. The graph of the second data set given in Figure 1.9b suggests that the analysis based on simple linear regres-

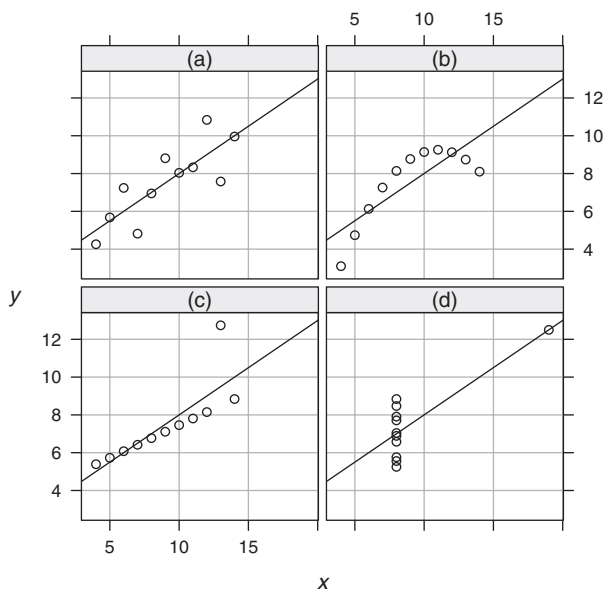


Figure 1.9 Four hypothetical data sets (Anscombe, 1973).

sion is incorrect and that a smooth curve, perhaps a quadratic polynomial, could be fit to the data with little remaining variability. Figure 1.9c suggests that the prescription of simple regression may be correct for most of the data, but one of the cases is too far away from the fitted regression line. This is called the *outlier problem*. Possibly the case that does not match the others should be deleted from the data set, and the regression should be refit from the remaining cases. This will lead to a different fitted line. Without a context for the data, we cannot judge one line “correct” and the other “incorrect.” The final set graphed in Figure 1.9d is different from the others in that there is not enough information to make a judgment concerning the mean function. If the separated point were deleted, we could not even estimate a slope. We must distrust an analysis that is so heavily dependent upon a single case.

1.5 TOOLS FOR LOOKING AT SCATTERPLOTS

Because looking at scatterplots is so important to fitting regression models, we establish some common vocabulary for describing the information in them and some tools to help us extract the information they contain.

The summary graph is of the response Y versus the predictor X . The mean function for the graph is defined by (1.1), and it characterizes how Y changes on the average as the value of X is varied. We may have a parametric model for the mean function and will use data to estimate the parameters. The

variance function also characterizes the graph, and in many problems we will assume at least at first that the variance function is constant. The scatterplot also will highlight separated points that may be of special interest because they do not fit the trend determined by the majority of the points.

A *null plot* has a horizontal straight line as its mean function, constant variance function, and no separated points. The scatterplot for the snowfall data appears to be a null plot.

1.5.1 Size

We may need to interact with a plot to extract all the available information, by changing scales, by resizing, or by removing linear trends. An example of this is given in Problem 1.3.

1.5.2 Transformations

In some problems, either or both of Y and X can be replaced by transformations so the summary graph has desirable properties. Most of the time, we will use power transformations, replacing, for example, X by X^λ for some number λ . Because logarithmic transformations are so frequently used, we will interpret $\lambda = 0$ as corresponding to a log transform.

1.5.3 Smoothers for the Mean Function

In the smallmouth bass data in Figure 1.5, we computed an estimate of $E(\text{Length}|\text{Age})$ using a simple nonparametric smoother obtained by averaging the repeated observations at each value of Age . Smoothers can also be defined when we do not have repeated observations at values of the predictor by averaging the observed data for all values of X *close to*, but not necessarily equal to, x . The literature on using smoothers to estimate mean functions has exploded in recent years, with fairly elementary treatments given by Bowman and Azzalini (1997), Green and Silverman (1994), Härdle (1990), and Simonoff (1996). Although these authors discuss nonparametric regression as an end in itself, we will generally use smoothers as *plot enhancements* to help us understand the information available in a scatterplot and to help calibrate the fit of a parametric mean function to a scatterplot.

For example, Figure 1.10 repeats Figure 1.1a, this time adding the estimated straight-line mean function and smoother called a `loess` smooth (Cleveland, 1979). Roughly speaking, the `loess` smooth estimates $E(Y|X=x)$ at the point x by fitting a straight line to a fraction of the points closest to x ; we used the fraction of 0.20 in this figure because the sample size is so large, but it is more usual to set the fraction to about $2/3$. The smoother is obtained by joining the estimated values of $E(Y|X=x)$ for many values of x . The `loess` smoother and the straight line agree almost perfectly for `mheight` close to average, but they

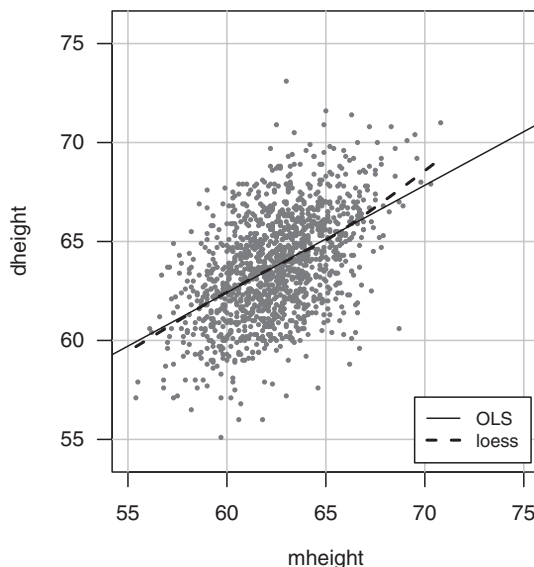


Figure 1.10 Heights data with the OLS line and a loess smooth with $\text{span} = 0.20$.

agree less well for larger values of `mheight` where there is much less data. Smoothers tend to be less reliable at the edges of the plot. We briefly discuss the `loess` smoother in Appendix A.5, but this material is dependent on the results in Chapters 2 and 3.

1.6 SCATTERPLOT MATRICES

With one predictor, a scatterplot provides a summary of the regression relationship between the response and the predictor. With many predictors, we need to look at many scatterplots. A *scatterplot matrix* is a convenient way to organize these plots.

Fuel Consumption

The goal of this example is to understand how fuel consumption varies over the 50 United States and the District of Columbia (Federal Highway Administration, 2001). Table 1.1 describes the variables to be used in this example; the data are given in the file `fuel2001`. The data were collected by the U.S. Federal Highway Administration.

Both `Drivers` and `FuelC` are state totals, so these will be larger in states with more people and smaller in less populous states. `Income` is computed per person. To make all these comparable and to attempt to eliminate the effect of size of the state, we compute rates $\text{Dlic} = \text{Drivers}/\text{Pop}$ and

Table 1.1 Variables in the Fuel Consumption Data^a

Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use, thousands of gallons
Income	Per person personal income for the year 2000, in thousands of dollars
Miles	Miles of Federal-aid highway miles in the state
Pop	2001 population age 16 and over
Tax	Gasoline state tax rate, cents per gallon
Fuel	$1000 \times \text{FuelC}/\text{Pop}$
Dlic	$1000 \times \text{Drivers}/\text{Pop}$
log(Miles)	Natural logarithm of Miles

^aAll data are for 2001, unless otherwise noted. The last three variables do not appear in the data file, but are computed from the previous variables, as described in the text.

$\text{Fuel} = \text{FuelC}/\text{Pop}$. Additionally, we replace Miles by its logarithm before doing any further analysis. Justification for replacing Miles with $\log(\text{Miles})$ is deferred to Problem 8.7.

Many problems will require replacing the observable predictors like `Drivers` and `Pop` with a function of them like `Dlic`. We will use the term *predictor* to correspond to the original variables, and the new term *regressor*, described more fully in Section 3.3, to refer to variables that are computed from the predictors. In some instances this distinction is artificial, but in others the distinction can clarify issues.

The scatterplot matrix for the fuel data is shown in Figure 1.11. Except for the diagonal, a scatterplot matrix is a 2D array of scatterplots. The variable names on the diagonal label the axes. In Figure 1.11, the variable `log(Miles)` appears on the horizontal axis of all the plots in the rightmost column and on the vertical axis of all the plots in the bottom row.³

Each plot in a scatterplot matrix is relevant to a particular one predictor regression of the variable on the vertical axis, given the variable on the horizontal axis. For example, the plot of `Fuel` versus `Tax` in the top row and second column of the scatterplot matrix in Figure 1.11 is relevant for the regression of `Fuel` on `Tax`. We can interpret this plot as we would a scatterplot for simple regression. We get the overall impression that `Fuel` decreases on the average as `Tax` increases, but there is lot of variation. We can make similar qualitative judgments about the each of the regressions of `Fuel` on the other variables. The overall impression is that `Fuel` is at best weakly related to each of the variables in the scatterplot matrix.

Does this help us understand how `Fuel` is related to all four predictors simultaneously? The marginal relationships between the response and each of

³The scatterplot matrix program used to draw Figure 1.11, which is the `pairs` function in R, has the diagonal running from the top left to the lower right. Other programs, such as the `splo` function in R, has the diagonal from lower left to upper right. There seems to be no compelling reason to prefer one over the other.

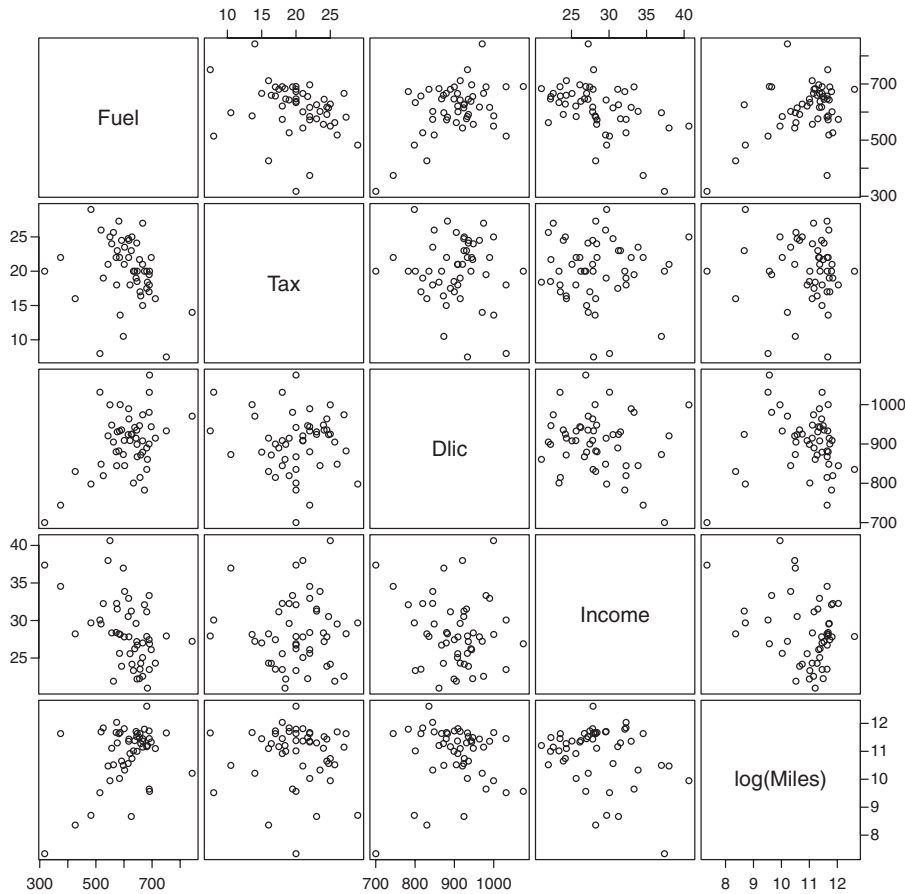


Figure 1.11 Scatterplot matrix for the fuel data.

the variables are *not* sufficient to understand the *joint* relationship between the response and the more than one predictor at a time. The interrelationships among the predictors are also important. The pairwise relationships between the predictors can be viewed in the remaining cells of the scatterplot matrix. In Figure 1.11, the relationships between all pairs of predictors appear to be very weak, suggesting that for this problem, the marginal plots including `Fuel` are quite informative about the multiple regression problem. General considerations for other scatterplot matrices will be developed in later chapters.

1.7 PROBLEMS

1.1 United Nations (Data file: UN11) The data in the file UN11 contains several variables, including `ppgdp`, the gross national product per person

in U.S. dollars, and `fertility`, the birth rate per 1000 females, both from the year 2009. The data are for 199 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries. The data were collected from United Nations (2011). We will study the dependence of `fertility` on `ppgdp`.⁴

1.1.1 Identify the predictor and the response.

1.1.2 Draw the scatterplot of `fertility` on the vertical axis versus `ppgdp` on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be plausible for a summary of this graph?

1.1.3 Draw the scatterplot of `log(fertility)` versus `log(ppgdp)` using natural logarithms. Does the simple linear regression model seem plausible for a summary of this graph? If you use a different base of logarithms, the *shape* of the graph won't change, but the *values on the axes* will change.

1.2 Smallmouth bass data (Data file: `wblake`) Compute the means and the variances for each of the eight subpopulations in the smallmouth bass data. Draw a graph of average length versus `Age` and compare with Figure 1.5. Draw a graph of the standard deviations versus `age`. If the variance function is constant, then the plot of standard deviation versus `Age` should be a null plot. Summarize the information.

1.3 (Data file: `Mitchell`) The data shown in Figure 1.12 give average soil temperature in degrees C at 20 cm depth in Mitchell, Nebraska for 17 years beginning January 1976, plotted versus the month number. The data were collected by K. Hubbard (Burnside et al., 1996).

1.3.1 Summarize the information in the graph about the dependence of soil temperature on month number.

1.3.2 The data used to draw Figure 1.12 are in the file `Mitchell`. Redraw the graph, but this time make the length of the horizontal axis at least 4 times the length of the vertical axis. Repeat Problem 1.3.1.

1.4 Old Faithful (Data file: `oldfaith`) The data file gives information about eruptions of Old Faithful Geyser during October 1980. Variables are the `Duration` in seconds of the current eruption, and the `Interval`, the time in minutes to the next eruption. The data were collected by volunteers and were provided by the late Roderick Hutchinson. Apart from missing data for the period from midnight to 6 a.m., this is a complete record of eruptions for that month.

⁴In the third edition of this book, similar data from 2000 were used in this problem. Those data are still available in the R package that accompanies this book and is called `UN1`.

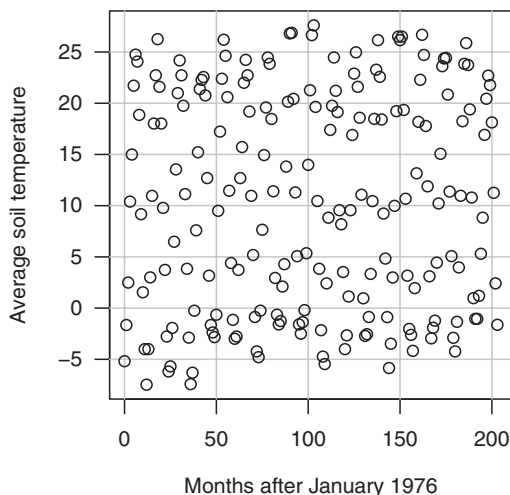


Figure 1.12 Monthly soil temperature data.

Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption.

Draw the relevant summary graph for predicting interval from duration and summarize your results.

1.5 Water runoff in the Sierras (Data file: `water`) Can Southern California's water supply in future years be predicted from past data? One factor affecting water availability is stream runoff. If runoff could be predicted, engineers, planners, and policy makers could do their jobs more efficiently. The data file contains 43 years' worth of precipitation measurements taken at six sites in the Sierra Nevada mountains (labeled `APMAM`, `APSAB`, `APSLAKE`, `OPBPC`, `OPRC`, and `OPSLAKE`) and stream runoff volume at a site near Bishop, California, labeled `BSAAM`.

Draw the scatterplot matrix for these data and summarize the information available from these plots.

1.6 Professor ratings (Data file: `Rateprof`) In the website and online forum `RateMyProfessors.com`, students rate and comment on their instructors. Launched in 1999, the site includes millions of ratings on thousands of instructors. The data file includes the summaries of the ratings of 364 instructors at a large campus in the Midwest (Bleske-Rechek and Fritsch, 2011). Each instructor included in the data had at least 10 ratings over a several year period. Students provided ratings of 1–5 on quality,

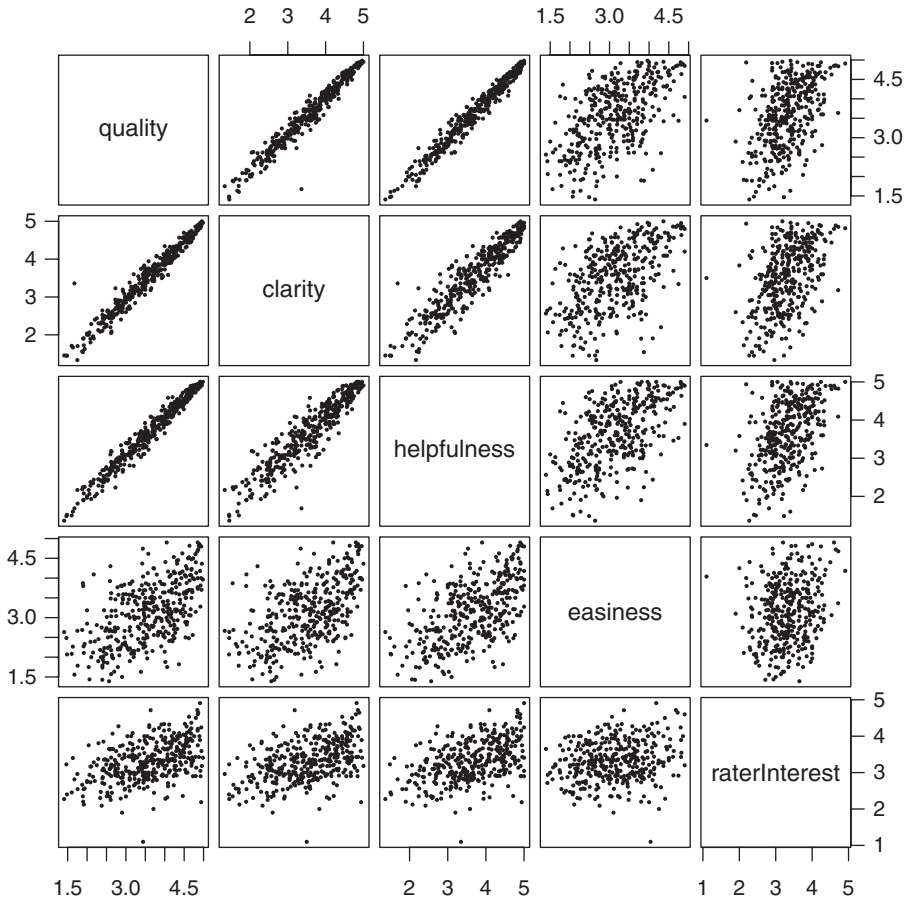


Figure 1.13 Average professor ratings from the file `Rateprof`.

helpfulness, clarity, easiness of instructor's courses, and raterInterest in the subject matter covered in the instructor's courses. The data file provides the averages of these five ratings, and these are shown in the scatterplot matrix in Figure 1.13.

Provide a brief description of the relationships between the five ratings.