and *dependent* variables for $X$ and $Y$ as these are easily confused with the broader meanings of terms. *Regression analysis* is another term used for linear modeling although regressions can also be nonlinear.

When $p = 1$, it is called *simple* regression but when $p > 1$ it is called *multiple* regression or sometimes *multivariate* regression. When there is more than one response, then it is called *multivariate multiple* regression or sometimes (confusingly) multivariate regression. We will not cover this in this book, although you can just do separate regressions on each $Y$.

The response must be a continuous variable, but the explanatory variables can be continuous, discrete or categorical, although we leave the handling of categorical explanatory variables to later in the book. Taking the example presented above, a regression with `diastolic` and `bmi` as $X$s and `diabetes` as $Y$ would be a multiple regression involving only quantitative variables which we tackle first. A regression with `diastolic` and `test` as $X$s and `bmi` as $Y$ would have one predictor that is quantitative and one that is qualitative, which we will consider later in Chapter 14 on *analysis of covariance*. A regression with `test` as $X$ and `diastolic` as $Y$ involves just qualitative predictors — a topic called *analysis of variance* (*ANOVA*), although this would just be a simple two sample situation. A regression of `test` as $Y$ on `diastolic` and `bmi` as predictors would involve a qualitative response. A *logistic regression* could be used, but this will not be covered in this book.

Regression analyses have two main objectives:

1. Prediction of future or unseen responses given specified values of the predictors.
2. Assessment of the effect of, or relationship between, explanatory variables and the response. We would like to infer causal relationships if possible.

You should be clear on the objective for the given data because some aspects of the resulting analysis may differ. Regression modeling can also be used in a descriptive manner to summarize the relationships between the variables. However, most end users of data have more specific questions in mind and want to direct the analysis toward a particular set of goals.

It is rare, except in a few cases in the precise physical sciences, to know (or even suspect) the true model. In most applications, the model is an empirical construct designed to answer questions about prediction or causation. It is usually not helpful to think of regression analysis as the search for some true model. The model is a means to an end, not an end in itself.

## 1.4 History

In the $18^{th}$ century, accurate navigation was a difficult problem of commercial and military interest. Although, it is relatively easy to determine latitude from Polaris, also known as the North Star, finding longitude then was difficult. Various attempts were made to devise a method using astronomy. Contrary to popular supposition, the Moon does not always show the same face and moves such that about 60% of its surface is visible at some time.

Tobias Mayer collected data on the locations of various landmarks on Moon,

including the Manilius crater, as they moved relative to the earth. He derived an equation describing the motion of the moon (called *libration*) taking the form:

$$\texttt{arc} = \beta + \alpha\texttt{sinang} + \gamma\texttt{cosang}$$

He wished to obtain values for the three unknowns $\alpha, \beta$ and $\gamma$. The variables `arc`, `sinang` and `cosang` can be observed using a telescope. A full explanation of the story behind the data and the derivation of the equation can be found in Stigler (1986).

Since there are three unknowns, we need only three distinct observations of the set of three variables to find a unique solution for $\alpha, \beta$ and $\gamma$. Embarassingly for Mayer, there were 27 sets of observations available. Astronomical measurements were naturally subject to some variation and so there was no solution that fit all 27 observations. Let's take a look at the first few lines of the data:

```
> data(manilius, package="faraway")
> head(manilius)
    arc sinang  cosang group
1 13.167 0.8836 -0.4682     1
2 13.133 0.9996 -0.0282     1
3 13.200 0.9899  0.1421     1
4 14.250 0.2221  0.9750     3
5 14.700 0.0006  1.0000     3
6 13.017 0.9308 -0.3654     1
```

Mayer's solution was to divide the data into three groups so that observations within each group were similar in some respect. He then computed the sum of the variables within each group. We can also do this:

```
> (moon3 <- aggregate(manilius[,1:3],list(manilius$group), sum))
  Group.1    arc  sinang  cosang
1       1 118.13  8.4987 -0.7932
2       2 140.28 -6.1404  1.7443
3       3 127.53  2.9777  7.9649
```

Now there are just three equations in three unknowns to be solved. The solution is:

```
> solve(cbind(9,moon3$sinang,moon3$cosang), moon3$arc)
[1] 14.54459 -1.48982  0.13413
```

Hence the computed values of $\alpha, \beta$ and $\gamma$ are -1.49, 14.5 and 0.134 respectively. One might question how Mayer selected his three groups, but this solution does not seem unreasonable.

Similar problems with more linear equations than unknowns continued to arise until 1805, when Adrien Marie Legendre published the method of least squares. Suppose we recognize that the equation is not exact and introduce an error term, $\varepsilon$:

$$\texttt{arc}_i = \beta + \alpha\texttt{sinang}_i + \gamma\texttt{cosang}_i + \varepsilon_i$$

where $i = 1, \ldots, 27$. Now we find $\alpha, \beta$ and $\gamma$ that minimize the sum of the squared errors: $\sum \varepsilon^2$. We will investigate this in much greater detail in the chapter to follow but for now we simply present the solution using R:

```
> lmod <- lm(arc ~ sinang + cosang, manilius)
> coef(lmod)
(Intercept)      sinang      cosang
  14.561624   -1.504581    0.091365
```

We observe that this solution is quite similar to Mayer's. The least squares solution is more satisfactory in that it requires no arbitrary division into groups. Carl Friedrich Gauss claimed to have devised the method of least squares earlier but without publishing it. At any rate, he did publish in 1809 showing that the method of least squares was, in some sense, optimal.

For many years, the method of least squares was confined to the physical sciences where it was used to resolve problems of overdetermined linear equations. The equations were derived from theory and least squares was used as a method to fit data to these equations to estimate coefficients like $\alpha, \beta$ and $\gamma$ above. It was not until later in the 19th century that linear equations (or models) were suggested empirically from the data rather than from theories of physical science. This opened up the field to the social and life sciences.

Francis Galton, a nephew of Charles Darwin, was important in this extension of statistics into social science. He coined the term *regression to mediocrity* in 1875 from which the rather peculiar term *regression* derives. Let's see how this terminology arose by looking at one of the datasets he collected at the time on the heights of parents and children in Galton (1886). We load the `HistData` package of historical statistical datasets and plot some of the data as seen in Figure 1.5. You will need to install this package using `install.packages("HistData")` if you have not already done so.

```
> data(GaltonFamilies, package="HistData")
> plot(childHeight ~ midparentHeight, GaltonFamilies)
```

We see that `midparentHeight`, defined as the father's height plus 1.08 times the mother's height divided by two, is correlated with the `childHeight`, both in inches. Now we might propose a linear relationship between the two of the form:

$$\texttt{childHeight} = \alpha + \beta \texttt{midparentHeight} + \varepsilon$$

We can estimate $\alpha$ and $\beta$ using R and plot the resulting fit as follows:

```
> lmod <- lm(childHeight ~ midparentHeight, GaltonFamilies)
> coef(lmod)
    (Intercept) midparentHeight
       22.63624         0.63736
> abline(lmod)
```

For the simple case of a response $y$ and a single predictor $x$, we can write the equation in the form:

$$\frac{y-\bar{y}}{SD_y} = r\frac{(x-\bar{x})}{SD_x}$$

where $r$ is the correlation between $x$ and $y$. The equation can be expressed in words as: the response in standard units is the correlation times the predictor in standard units. We can verify that this produces the same results as above by rearranging the equation in the form $y = \alpha + \beta x$ and computing the estimates:

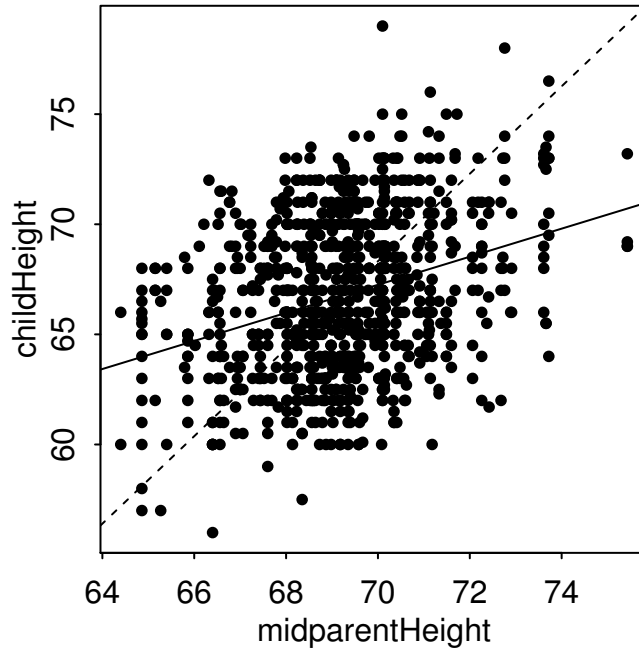


Figure 1.5 The height of child is plotted against a combined parental height defined as (father's height + 1.08 × mother's height)/2.

```
> (beta <- with(GaltonFamilies, cor(midparentHeight, childHeight) * sd
    (childHeight) / sd(midparentHeight)))
[1] 0.63736
> (alpha <- with(GaltonFamilies, mean(childHeight) - beta * mean(
    midparentHeight)))
[1] 22.636
```

Now one might naively expect that a child with parents who are, for example, one standard deviation above average in height, to also be one standard deviation above average in height, give or take. The supposition would set $r = 1$ in the equation and leads to a line which we compute and plot below:

```
> (beta1 <- with(GaltonFamilies, sd(childHeight) / sd(midparentHeight)
    ))
[1] 1.9859
> (alpha1 <- with(GaltonFamilies, mean(childHeight) - beta1 * mean(
    midparentHeight)))
[1] -70.689
> abline(alpha1, beta1, lty=2)
```

The resulting dashed line is added to Figure 1.5. The lines cross at the point of the averages. We can see that a child of tall parents is predicted by the least squares line to have a height which is above average but not quite as tall as the parents as the dashed line would have you believe. Similarly children of below average height

parents are predicted to have a height which is still below average but not quite as short as the parents. This is why Galton used the phrase "regression to mediocrity" and the phenomenom is sometimes called the regression effect.

This applies to any $(x,y)$ situation like this. For example, in sports, an athlete may have a spectacular first season only to do not quite as well in the second season. Sports writers come up with all kinds of explanations for this but the regression effect is likely to be the unexciting cause. In the parents and children example, although it does predict that successive descendants in the family will come closer to the mean, it does not imply the same of the population in general since random fluctuations will maintain the variation, so no need to get too pessimistic about mediocrity! In many other applications of linear modeling, the regression effect is not of interest because different types of variables are measured. Unfortunately, we are now stuck with the rather gloomy word of regression thanks to Galton.

Regression methodology developed rapidly with the advent of high-speed computing. Just fitting a regression model used to require extensive hand calculation. As computing hardware has improved, the scope for analysis has widened. This has led to an extensive development in the methodology and the scale of problems that can be tackled.

## Exercises

1. The dataset `teengamb` concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.
2. The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988. Make a numerical and graphical summary of the data as in the previous question.
3. The dataset `prostate` is from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data as in the first question.
4. The dataset `sat` comes from a study entitled "Getting What You Pay For: The Debate Over Equity in Public School Expenditures." Make a numerical and graphical summary of the data as in the first question.
5. The dataset `divusa` contains data on divorces in the United States from 1920 to 1996. Make a numerical and graphical summary of the data as in the first question.