estimated in order to calculate to fitted values $\widehat{\theta}_i$ (for example, see Agresti, 1990, page 479). Expression (2.5) is, in fact, the usual chi-squared goodness of fit statistic for count data which is often written as

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(m)$$

where o_i denotes the observed frequency and e_i denotes the corresponding expected frequency. In this case $o_i = Y_i$, $e_i = \widehat{\theta}_i$ and $\sum r_i^2 = X^2$.

For the data on chronic medical conditions, for model (2.1)

$$\sum r_i^2 = 6 \times (-1.088)^2 + 10 \times (-0.169)^2 + \ldots + 1 \times 1.669^2 = 46.759.$$

This value is consistent with $\sum r_i^2$ being an observation from the central chi-squared distribution with $m = 23 + 26 - 1 = 48$ degrees of freedom. (Recall from Section 1.4.2, that if $X^2 \sim \chi^2(m)$ then $\mathrm{E}(X^2) = m$ and notice that the calculated value $X^2 = \sum r_i^2 = 46.759$ is near the expected value of 48.)

Similarly, for model (2.2)

$$\sum r_i^2 = 6 \times (-1.193)^2 + \ldots + 1 \times 2.184^2 = 43.659$$

which is consistent with the central chi-squared distribution with $m = 49-2 = 47$ degrees of freedom. The difference between the values of $\sum r_i^2$ from models (2.1) and (2.2) is small: $46.759-43.659 = 3.10$. This suggests that model (2.2) with two parameters, may not describe the data much better than the simpler model (2.1). If this is so, then the data provide evidence supporting the null hypothesis $\mathrm{H}_0 : \theta_1 = \theta_2$. More formal testing of the hypothesis is discussed in Chapter 4.

The next example illustrates steps of the model fitting process with continuous data.

2.2.2 Birthweight and gestational age

The data in Table 2.3 are the birthweights (in grams) and estimated gestational ages (in weeks) of 12 male and female babies born in a certain hospital. The mean ages are almost the same for both sexes but the mean birthweight for boys is higher than the mean birthweight for girls. The data are shown in the scatter plot in Figure 2.2. There is a linear trend of birthweight increasing with gestational age and the girls tend to weigh less than the boys of the same gestational age. The question of interest is whether the rate of increase of birthweight with gestational age is the same for boys and girls.

Let Y_{jk} be a random variable representing the birthweight of the kth baby in group j where $j = 1$ for boys and $j = 2$ for girls and $k = 1, \ldots, 12$. Suppose that the Y_{jk}'s are all independent and are Normally distributed with means $\mu_{jk} = \mathrm{E}(Y_{jk})$, which may differ among babies, and variance σ^2 which is the same for all of them.

A fairly general model relating birthweight to gestational age is

$$\mathrm{E}(Y_{jk}) = \mu_{jk} = \alpha_j + \beta_j x_{jk}$$

Table 2.3 *Birthweight and gestational age for boys and girls.*

	Boys		Girls	
	Age	Birthweight	Age	Birthweight
	40	2968	40	3317
	38	2795	36	2729
	40	3163	40	2935
	35	2925	38	2754
	36	2625	42	3210
	37	2847	39	2817
	41	3292	40	3126
	40	3473	37	2539
	37	2628	36	2412
	38	3176	38	2991
	40	3421	39	2875
	38	2975	40	3231
Means	38.33	3024.00	38.75	2911.33

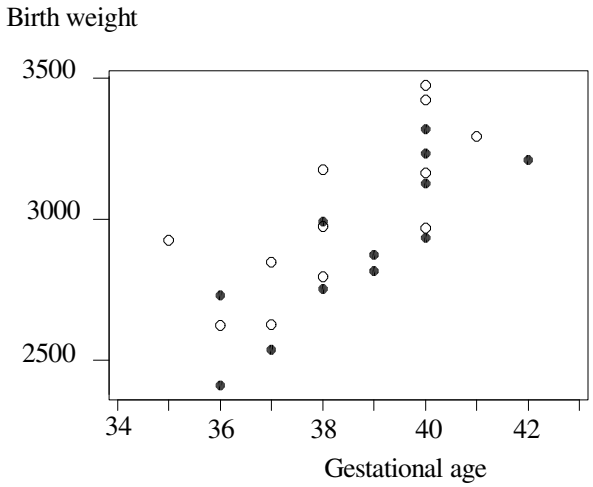

Figure 2.2 *Birthweight plotted against gestational age for boys (open circles) and girls (solid circles); data in Table 2.3.*

where x_{jk} is the gestational age of the kth baby in group j. The intercept parameters α_1 and α_2 are likely to differ because, on average, the boys were heavier than the girls. The slope parameters β_1 and β_2 represent the average increases in birthweight for each additional week of gestational age. The question of interest can be formulated in terms of testing the null hypothesis H_0 : $\beta_1 = \beta_2 = \beta$ (that is, the growth rates are equal and so the lines are parallel), against the alternative hypothesis $\mathrm{H}_1 : \beta_1 \neq \beta_2$.

We can test H_0 against H_1 by fitting two models

$$\mathrm{E}(Y_{jk}) = \mu_{jk} = \alpha_j + \beta x_{jk}; \quad Y_{jk} \sim N(\mu_{jk}, \sigma^2), \tag{2.6}$$

$$\mathrm{E}(Y_{jk}) = \mu_{jk} = \alpha_j + \beta_j x_{jk}; \quad Y_{jk} \sim N(\mu_{jk}, \sigma^2). \tag{2.7}$$

The probability density function for Y_{jk} is

$$f(y_{jk}; \mu_{jk}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(y_{jk} - \mu_{jk})^2].$$

We begin by fitting the more general model (2.7). The log-likelihood function is

$$\begin{aligned} l_1(\alpha_1, \alpha_2, \beta_1, \beta_2; \mathbf{y}) &= \sum_{j=1}^{J}\sum_{k=1}^{K}[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_{jk} - \mu_{jk})^2] \\ &= -\frac{1}{2}JK\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk} - \alpha_j - \beta_j x_{jk})^2 \end{aligned}$$

where $J = 2$ and $K = 12$ in this case. When obtaining maximum likelihood estimates of $\alpha_1, \alpha_2, \beta_1$ and β_2 we treat the parameter σ^2 as a known constant, or **nuisance parameter,** and we do not estimate it.

The maximum likelihood estimates are the solutions of the simultaneous equations

$$\begin{aligned} \frac{\partial l_1}{\partial \alpha_j} &= \frac{1}{\sigma^2}\sum_k (y_{jk} - \alpha_j - \beta_j x_{jk}) = 0, \\ \frac{\partial l_1}{\partial \beta_j} &= \frac{1}{\sigma^2}\sum_k x_{jk}(y_{jk} - \alpha_j - \beta_j x_{jk}) = 0, \end{aligned} \tag{2.8}$$

where $j = 1$ or 2.

An alternative to maximum likelihood estimation is least squares estimation. For model (2.7), this involves minimizing the expression

$$S_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk} - \mu_{jk})^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk} - \alpha_j - \beta_j x_{jk})^2. \tag{2.9}$$

The least squares estimates are the solutions of the equations

$$\begin{aligned}\frac{\partial S_1}{\partial \alpha_j} &= -2\sum_{k=1}^{K}(y_{jk} - \alpha_j - \beta_j x_{jk}) = 0,\\ \frac{\partial S_1}{\partial \beta_j} &= -2\sum_{k=1}^{K} x_{jk}(y_{jk} - \alpha_j - \beta_j x_{jk}) = 0. \end{aligned} \tag{2.10}$$

The equations to be solved in (2.8) and (2.10) are the same and so maximizing l_1 is equivalent to minimizing S_1. For the remainder of this example we will use the least squares approach.

The estimating equations (2.10) can be simplified to

$$\begin{aligned}\sum_{k=1}^{K} y_{jk} - K\alpha_j - \beta_j \sum_{k=1}^{K} x_{jk} &= 0,\\ \sum_{k=1}^{K} x_{jk}y_{jk} - K\alpha_j \sum_{k=1}^{K} x_{jk} - \beta_j \sum_{k=1}^{K} x_{jk}^2 &= 0\end{aligned}$$

for $j = 1$ or 2. These are called the **normal equations**. The solution is

$$\begin{aligned} b_j &= \frac{K\sum_k x_{jk}y_{jk} - (\sum_k x_{jk})(\sum_k y_{jk})}{K\sum_k x_{jk}^2 - (\sum_k x_{jk})^2},\\ a_j &= \overline{y}_j - b_j\overline{x}_j, \end{aligned}$$

where a_j is the estimate of α_j and b_j is the estimate of β_j, for $j = 1$ or 2. By considering the second derivatives of (2.9) it can be verified that the solution of equations (2.10) does correspond to the minimum of S_1. The numerical value for the minimum value for S_1 for a particular data set can be obtained by substituting the estimates for α_j and β_j and the data values for y_{jk} and x_{jk} into (2.9).

To test $\mathrm{H}_0 : \beta_1 = \beta_2 = \beta$ against the more general alternative hypothesis H_1, the estimation procedure described above for model (2.7) is repeated but with the expression in (2.6) used for μ_{jk}. In this case there are three parameters, α_1, α_2 and β, instead of four to be estimated. The least squares expression to be minimized is

$$S_0 = \sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk} - \alpha_j - \beta x_{jk})^2. \tag{2.11}$$

From (2.11) the least squares estimates are given by the solution of the simultaneous equations

$$\begin{aligned}\frac{\partial S_0}{\partial \alpha_j} &= -2\sum_{k=1}^{K}(y_{jk} - \alpha_j - \beta x_{jk}) = 0,\\ \frac{\partial S_0}{\partial \beta} &= -2\sum_{j=1}^{J}\sum_{k=1}^{K} x_{jk}(y_{jk} - \alpha_j - \beta x_{jk}) = 0, \end{aligned} \tag{2.12}$$

Table 2.4 *Summary of data on birthweight and gestational age in Table 2.3 (summation is over k=1,...,K where K=12).*

	Boys ($j=1$)	Girls ($j=2$)
$\sum x$	460	465
$\sum y$	36288	34936
$\sum x^2$	17672	18055
$\sum y^2$	110623496	102575468
$\sum xy$	1395370	1358497

for $j = 1$ and 2. The solution is

$$\begin{aligned} b &= \frac{K\sum_j\sum_k x_{jk}y_{jk} - \sum_j(\sum_k x_{jk}\sum_k y_{jk})}{K\sum_j\sum_k x_{jk}^2 - \sum_j(\sum_k x_{jk})^2}, \\ a_j &= \overline{y}_j - b\overline{x}_j. \end{aligned}$$

These estimates and the minimum value for S_0 can be calculated from the data.

For the example on birthweight and gestational age, the data are summarized in Table 2.4 and the least squares estimates and minimum values for S_0 and S_1 are given in Table 2.5. The fitted values $\widehat{y}_{jk}$ are shown in Table 2.6. For model (2.6), $\widehat{y}_{jk} = a_j + bx_{jk}$ is calculated from the estimates in the top part of Table 2.5. For model (2.7), $\widehat{y}_{jk} = a_j + b_j x_{jk}$ is calculated using estimates in the bottom part of Table 2.5. The residual for each observation is $y_{jk} - \widehat{y}_{jk}$. The standard deviation s of the residuals can be calculated and used to obtain approximate standardized residuals $(y_{jk} - \widehat{y}_{jk})/s$. Figures 2.3 and 2.4 show for models (2.6) and (2.7), respectively: the standardized residuals plotted against the fitted values; the standardized residuals plotted against gestational age; and Normal probability plots. These types of plots are discussed in Section 2.3.4. The Figures show that:

1. Standardized residuals show no systematic patterns in relation to either the fitted values or the explanatory variable, gestational age.
2. Standardized residuals are approximately Normally distributed (as the points are near the solid lines in the bottom graphs).
3. Very little difference exists between the two models.

The apparent lack of difference between the models can be examined by testing the null hypothesis H_0 (corresponding to model (2.6)) against the alternative hypothesis H_1 (corresponding to model (2.7)). If H_0 is correct, then the minimum values $\widehat{S}_1$ and $\widehat{S}_0$ should be nearly equal. If the data support this hypothesis, we would feel justified in using the simpler model (2.6) to describe the data. On the other hand, if the more general hypothesis H_1 is true then $\widehat{S}_0$ should be much larger than $\widehat{S}_1$ and model (2.7) would be preferable.

To assess the relative magnitude of the values $\widehat{S}_1$ and $\widehat{S}_0$ we need to use the

Table 2.5 *Analysis of data on birthweight and gestational age in Table 2.3.*

Model	Slopes	Intercepts	Minimum sum of squares
(2.6)	$b = 120.894$	$a_1 = -1610.283$	$\widehat{S}_0 = 658770.8$
		$a_2 = -1773.322$	
(2.7)	$b_1 = 111.983$	$a_1 = -1268.672$	$\widehat{S}_1 = 652424.5$
	$b_2 = 130.400$	$a_2 = -2141.667$	

Table 2.6 *Observed values and fitted values under model (2.6) and model (2.7) for data in Table 2.3.*

Sex	Gestational age	Birthweight	Fitted value under (2.6)	Fitted value under (2.7)
Boys	40	2968	3225.5	3210.6
	38	2795	2983.7	2986.7
	40	3163	3225.5	3210.6
	35	2925	2621.0	2650.7
	36	2625	2741.9	2762.7
	37	2847	2862.8	2874.7
	41	3292	3346.4	3322.6
	40	3473	3225.5	3210.6
	37	2628	2862.8	2874.7
	38	3176	2983.7	2986.7
	40	3421	3225.5	3210.6
	38	2975	2983.7	2986.7
Girls	40	3317	3062.5	3074.3
	36	2729	2578.9	2552.7
	40	2935	3062.5	3074.3
	38	2754	2820.7	2813.5
	42	3210	3304.2	3335.1
	39	2817	2941.6	2943.9
	40	3126	3062.5	3074.3
	37	2539	2699.8	2683.1
	36	2412	2578.9	2552.7
	38	2991	2820.7	2813.5
	39	2875	2941.6	2943.9
	40	3231	3062.5	3074.3

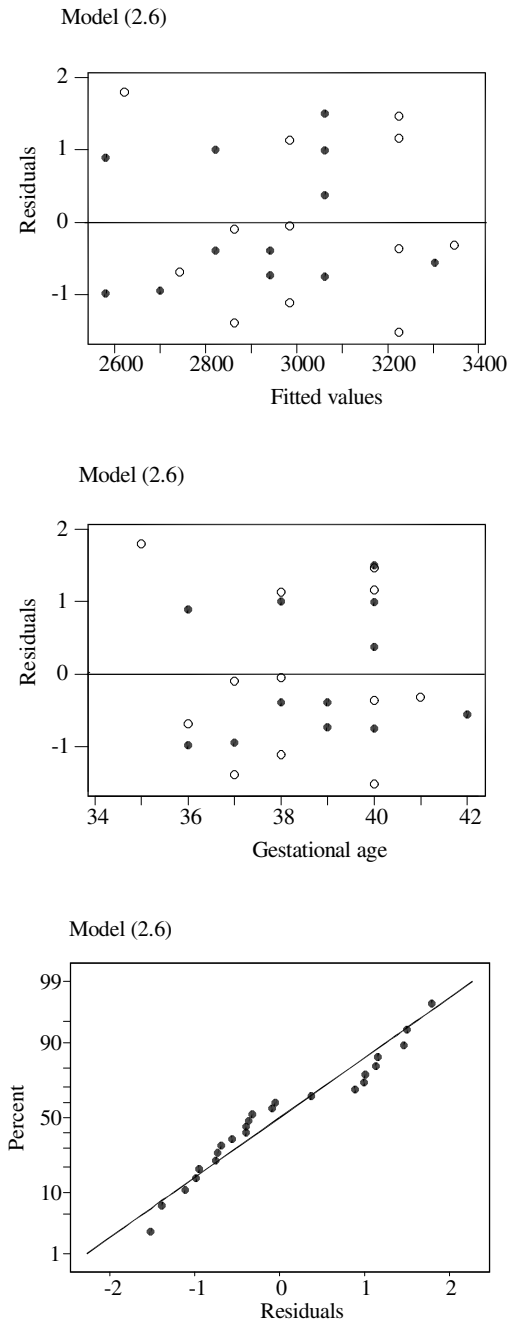

Figure 2.3 *Plots of standardized residuals for Model (2.6) for the data on birthweight and gestational age (Table 2.3); for the top and middle plots, open circles correspond to data from boys and solid circles correspond to data from girls.*

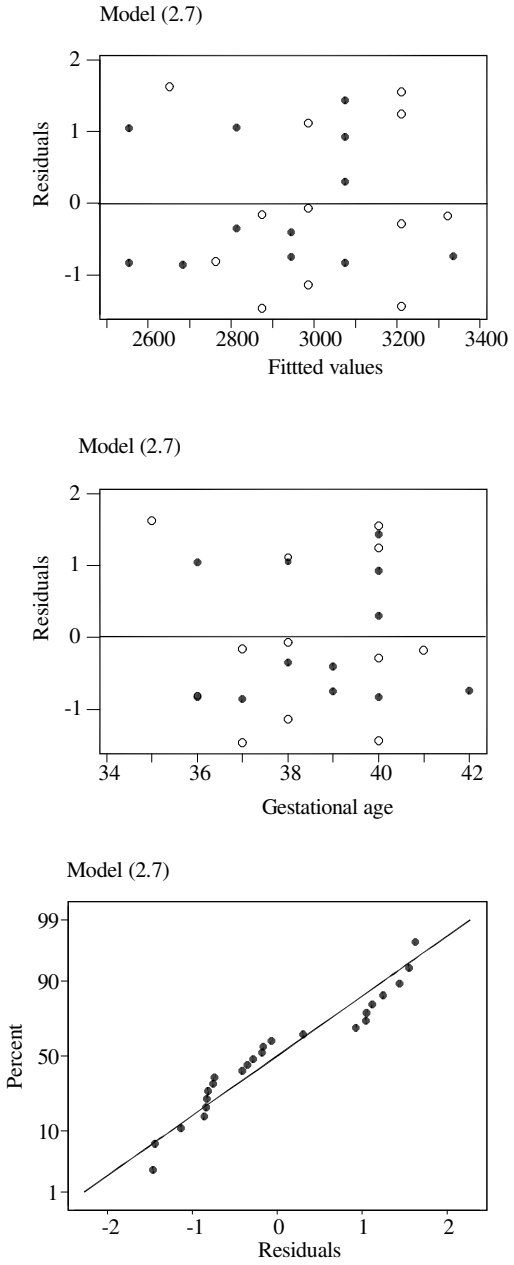

Figure 2.4 *Plots of standardized residuals for Model (2.7) for the data on birthweight and gestational age (Table 2.3); for the top and middle plots, open circles correspond to data from boys and solid circles correspond to data from girls.*

sampling distributions of the corresponding random variables

$$\widehat{S}_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{jk} - a_j - b_j x_{jk})^2$$

and

$$\widehat{S}_0 = \sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{jk} - a_j - b x_{jk})^2.$$

It can be shown (see Exercise 2.3) that

$$\begin{aligned}\widehat{S}_1 &= \sum_{j=1}^{J}\sum_{k=1}^{K}[Y_{jk} - (\alpha_j + \beta_j x_{jk})]^2 - K\sum_{j=1}^{J}(\overline{Y}_j - \alpha_j - \beta_j \overline{x}_j)^2 \\ &\quad - \sum_{j=1}^{J}(b_j - \beta_j)^2(\sum_{k=1}^{K} x_{jk}^2 - K\overline{x}_j^2)\end{aligned}$$

and that the random variables Y_{jk}, $\overline{Y}_j$ and b_j are all independent and have the following distributions:

$$\begin{aligned}Y_{jk} &\sim N(\alpha_j + \beta_j x_{jk}, \sigma^2), \\ \overline{Y}_j &\sim N(\alpha_j + \beta_j \overline{x}_j, \sigma^2/K), \\ b_j &\sim N(\beta_j, \sigma^2/(\sum_{k=1}^{K} x_{jk}^2 - K\overline{x}_j^2)).\end{aligned}$$

Therefore $\widehat{S}_1/\sigma^2$ is a linear combination of sums of squares of random variables with Normal distributions. In general, there are JK random variables $(Y_{jk} - \alpha_j - \beta_j x_{jk})^2/\sigma^2$, J random variables $(\overline{Y}_j - \alpha_j - \beta_j \overline{x}_j)^2 K/\sigma^2$ and J random variables $(b_j - \beta_j)^2(\sum_k x_{jk}^2 - K\overline{x}_j^2)/\sigma^2$. They are all independent and each has the $\chi^2(1)$ distribution. From the properties of the chi-squared distribution in Section 1.5, it follows that $\widehat{S}_1/\sigma^2 \sim \chi^2(JK - 2J)$. Similarly, if H_0 is correct then $\widehat{S}_0/\sigma^2 \sim \chi^2[JK - (J+1)]$. In this example $J = 2$ so $\widehat{S}_1/\sigma^2 \sim \chi^2(2K - 4)$ and $\widehat{S}_0/\sigma^2 \sim \chi^2(2K - 3)$. In each case the value for the degrees of freedom is the number of observations minus the number of parameters estimated.

If β_1 and β_2 are not equal (corresponding to H_1), then $\widehat{S}_0/\sigma^2$ will have a non-central chi-squared distribution with $JK - (J+1)$ degrees of freedom. On the other hand, provided that model (2.7) describes the data well, $\widehat{S}_1/\sigma^2$ will have a central chi-squared distribution with $JK - 2J$ degrees of freedom.

The statistic $\widehat{S}_0 - \widehat{S}_1$ represents the improvement in fit of (2.7) compared to (2.6). If H_0 is correct, then

$$\frac{1}{\sigma^2}(\widehat{S}_0 - \widehat{S}_1) \sim \chi^2(J-1).$$

If H_0 is not correct then $(\widehat{S}_0 - \widehat{S}_1)/\sigma^2$ has a non-central chi-squared distribu-

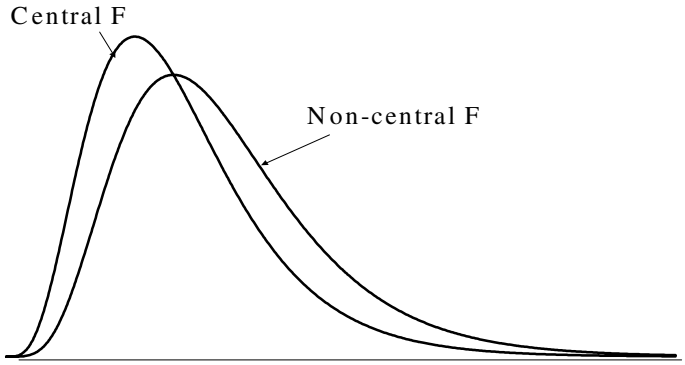

Figure 2.5 *Central and non-central F distributions.*

tion. However, as σ^2 is unknown, we cannot compare $(\widehat{S}_0 - \widehat{S}_1)/\sigma^2$ directly with the $\chi^2(J-1)$ distribution. Instead we eliminate σ^2 by using the ratio of $(\widehat{S}_0 - \widehat{S}_1)/\sigma^2$ and the random variable $\widehat{S}_1/\sigma^2$ with a central chi-squared distribution, each divided by the relevant degrees of freedom,

$$F = \frac{(\widehat{S}_0 - \widehat{S}_1)/\sigma^2}{(J-1)} \Big/ \frac{\widehat{S}_1/\sigma^2}{(JK-2J)} = \frac{(\widehat{S}_0 - \widehat{S}_1)/(J-1)}{\widehat{S}_1/(JK-2J)}.$$

If H_0 is correct, from Section 1.4.4, F has the central distribution $F(J-1, JK-2J)$. If H_0 is not correct, F has a non-central F-distribution and the calculated value of F will be larger than expected from the central F-distribution (see Figure 2.5).

For the example on birthweight and gestational age, the value of F is

$$\frac{(658770.8 - 652424.5)/1}{652424.5/20} = 0.19$$

This value is certainly not statistically significant when compared with the $F(1, 20)$ distribution. Thus the data do not provide evidence against the hypothesis $\mathrm{H}_0 : \beta_0 = \beta_1$, and on the grounds of simplicity model (2.6), which specifies the same slopes but different intercepts, is preferable.

These two examples illustrate the main ideas and methods of statistical modelling which are now discussed more generally.

2.3 Some principles of statistical modelling

2.3.1 Exploratory data analysis

Any analysis of data should begin with a consideration of each variable separately, both to check on data quality (for example, are the values plausible?) and to help with model formulation.

1. What is the scale of measurement? Is it continuous or categorical? If it