1.6 Writing statistical reports

In Section 1.7, we present a data example, and illustrate some of the procedures outlined above. Since a linear models course is often the first opportunity for the student to do substantial analyses of data and write statistical reports, we take the opportunity to offer some remarks about writing statistical reports.

In general, a report on a statistical data analysis consists of the following three parts:

(i) Presentation of the problem and the data.
(ii) Statistical analysis.
(iii) Conclusions.

The three parts may be divided into subsections, if necessary.

The application of the statistical procedures and methods, as outlined in the previous sections of this chapter, all belong to part (ii). In part (i) one presents the problem and the data, explaining the circumstances under which the data were obtained, and the principal questions that the statistical analysis is intended to answer. In (iii), the results of the statistical analysis are discussed, with reference to the circumstances under which the data were obtained, responding, as far as possible, to the questions posed in part (i). Note that (i) and (iii) must be phrased, as far as possible, in a non-statistical language, directed essentially at the researcher who obtained the data. Statistical jargon, such as "test", "estimator", "likelihood" etc. should, as far as possible, be confined to part (ii).

A statistical report should be a clear and fluently written text, such that it can be understood by readers who know the basics of statistical data analysis, but not necessarily the conventions of the statistics course you are currently following. Only the most relevant tables and graphical displays should be included in part (ii), whereas computer programs and their output should be put in an appendix. A statistical report should *not* consist of annotated computer output.

Finally, a word about parsimony. A statistical report that manages to communicate its message in a brief form is preferable to a lengthy and obscure report. The length of the report should be in reasonable proportion to the size of the data being analyzed. I sometimes suggest the following rule of thumb to my students

$$\#\ \text{pages} \sim k\sqrt{n},$$

where n is the number of observations and k is the number of independent variables in the data. Thus, a simple linear regression with $n = 25$ and $k = 1$ rarely deserves more than a five-page report. Of course, not every

statistical report should have this size, but if your report is longer than the rule suggests, making it shorter will probably also make it clearer. Actually, the formula is probably too generous, especially for large data sets.

1.7 Analysis of the spinach data

(i) Presentation of the problem.

As mentioned in Example 1.1, the data in Table 1.1 represent the relationship between the percentage dry matter of fresh spinach (x) and the percentage preserved ascorbic acid after drying at 90°C (y). The data are from an investigation concerning the preservation of ascorbic acid in vegetables during drying and storing, so consequently percentage preserved ascorbic acid after drying is chosen as response variable (y). The questions relevant to this investigation are if the relationship between x and y can be said to be linear in the x-interval under study, which ranges from 6 to 15 percent dry matter, and what the magnitude of the deviation from the linear relationship is. Furthermore, we may ask how precisely the parameters of the linear relationship have been estimated.

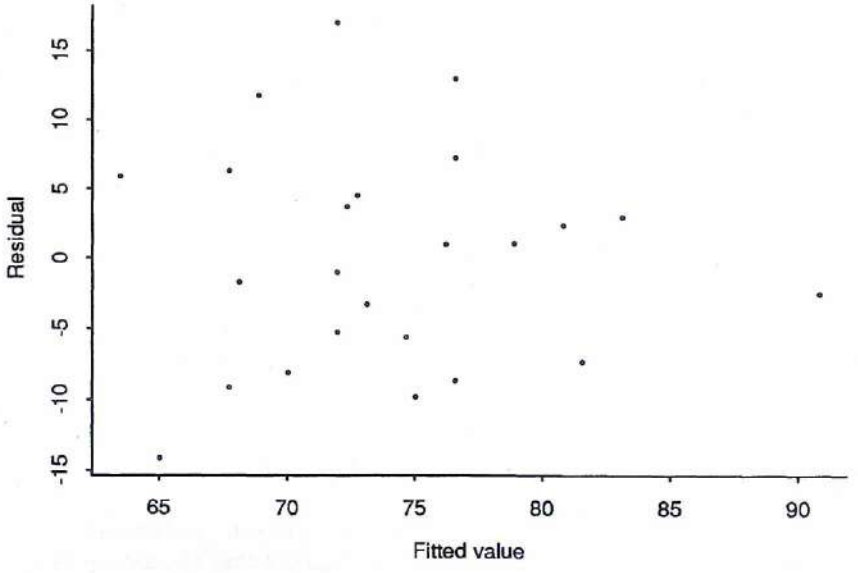

Figure 7.1 *Plot of residuals, spinach data*

i) *Statistical analysis*

Figure 1.2 shows the scatterplot of y versus x, and as noted ear-
:r, the plot does not suggest any substantial departures from the linear
gression model for these data. The statistical model that we use is hence

$$Y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2), \quad i = 1, \ldots, 24,$$

here $Y_1, \ldots, Y_{24}$ are independent, the data y_i representing a realization
the random variable Y_i. A further check of the model is provided by
ıe plot of residuals of the normal plot of residuals, shown in Figures 7.1
ıd 7.2. The first plot shows that the variance is constant, and the second
ıows a nice linear relationship, confirming the normality of the residuals.
ı any case, it is difficult to reject normality based on a sample of only 24
)servations.

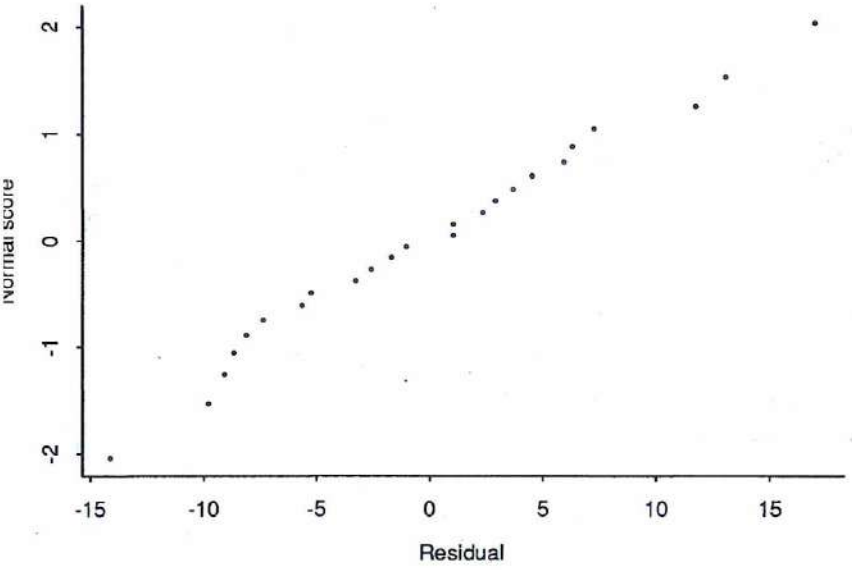

Figure 7.2 *Normal plot of residuals, spinach data.*

To complete the verification of the model, we note that the assump-
on of independence of the 24 observations requires that the 24 experiments
ere in some sense performed separately, both in space and time, although
e have no specific information about this point here.

Table 7.1. Parameter estimates for linear regression model, spinach data

Parameter	Estimate	s.e.
β_1	33.48	11.10
β_2	3.85	1.04
	$\tilde{\sigma}^2 = 64.84$	d.f. $= 22$

The parameter estimates for the model and their standard errors are given in Table 7.1. Based on these values, the estimated linear relationship between $\mathrm{E}(Y)$ and x is given by

$$\mathrm{E}(Y) = 33.48 + 3.85x, \tag{7.1}$$

with a standard deviation estimated by $\tilde{\sigma} = 8.05$. A 95% confidence interval for β_2 is $[1.70, 6.00]$. The t-test for the hypothesis $\beta_2 = 0$ is

$$t(y) = \frac{3.85}{1.04} = 3.70,$$

with 22 degrees of freedom, which gives a p-value of less than 0.01. There is hence a strong indication that β_2 is not zero.

(iii) *Conclusions*

The statistical analysis shows that the data may reasonably be described by a linear regression model, the estimated relationship being given by (7.1). The estimates and their standard errors in Table 7.1 show that the parameters are not very precisely estimated, particularly so for the intercept β_1. The statistical test for the hypothesis that the slope is zero rejects the hypothesis. The percentage preserved ascorbic acid after drying hence depends on the percentage dry matter of fresh spinach, with a slope between about 1.7 and 6.0 (95% confidence interval). Hence, equation (7.1) may be useful for predicting y from x, but a vertical deviation ("prediction error") of about $1.96 \times 8.05 = 15.77$ should be expected. For example, for $x = 10$, a value of y between 56.21 and 87.75 is expected with probability 95%, with a median value of 71.98.

We have here used the normal distribution as a basis for the prediction interval. A more detailed discussion of prediction, which is given in Section 4.5, shows that the correct prediction intervals should be based on the t-distribution, although the above approach is approximately correct.