

a fitted model. This reverse prediction problem is called *calibration* (see Brown, 1993).

Apart from the situation where the observer has no control over the explanatory variables, *designed experiments* are also used to measure the effects of certain explanatory variables or to test empirical beliefs statistically. The linear model can be used by the experimenter as a basis for choosing the values of the controllable variables so that the answers to the crucial questions are best answered. Chapter 6 briefly outlines the basic issues of experimental design.

If one of the explanatory variables is controllable, then one might ask: Which value of this variable will produce a desired level of response (within a certain margin of error)? This task, which is related to calibration, is called the problem of *control*. The linear model provides a framework for solving this problem (see Press, 1971, Chapter 14).

A somewhat related question in polynomial regression is the following: How should one choose the explanatory variable so that the expected response is optimized? A similar question may be asked when the response is modelled as a polynomial in *several* variables. Typically there are constraints on the range of the explanatory variables. In the context of the assumed model, the problem reduces to finding the maximum or minimum of the estimated *response surface* within a certain range of the variables. An example without range constraints is given in Exercise 1.15. Various methods for response surfaces are described in detail by Khuri and Cornell (1996) and Box and Draper (1987).

The linear model is also used as a basis for imputation of missing data. The idea is to fill in the void using information from related variables, as in the case of prediction (see Titterington and Sedransk, 1987, for details). Diagnostic tools developed in the context of the linear model are sometimes used to detect other defects in the data such as bad or incorrect data (see Belsley et al., 1980).

1.7 A tour through the rest of the book

Chapters 2 and 3 provide a brief summary of various linear-algebraic and statistical concepts that are needed for the later chapters. The

summary is meant to be a refresher as well as a reference for those who have already studied these topics in another course. We tried to make the material somewhat self-contained, so that the reader does not have to constantly refer to other sources for these basic results.

In Chapter 4 we consider estimation of linear functions of β in the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$. After addressing the question of estimability of linear functions of β , we develop the theory of optimal estimation of these functions by means of linear functions of \mathbf{y} . The theory is built around linear zero functions (LZFs) — linear functions of \mathbf{y} which have zero mean. We also show the connection of the optimal estimator with estimators obtained by the least squares method, and provide statistical interpretations of the residual sum of squares. Subsequently we introduce the ideas of reparametrization, nuisance parameters, linear restrictions and collinearity in the linear model.

While continuing with the model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ in Chapter 5, we obtain confidence intervals/regions for parametric functions and tests of linear hypotheses. To this end, we discuss the sampling distribution of the estimated model parameters, and examine which linear hypotheses are testable. The generic test statistic is motivated via an intuitive decomposition of the sum of squares in terms of LZFs. Finally, we discuss the problem of prediction on the basis of the linear model, and the effect of collinearity on various aspects of inference.

In Chapter 6 we deal with the analysis of designed experiments. Some standard designs and their analyses are put in the context of the model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, and the explained sum of squares is decomposed further in terms of interpretable LZFs. The chapter ends with a discussion of models where controllable and other factors are present simultaneously. Discussion in this chapter is confined to a few basic designs.

Chapter 7, which deals with the model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, runs somewhat parallel to the developments of Chapters 4 and 5. We explain the importance of the general linear model, and in particular, the singular linear model. We show how the ideas used in Chapters 4 and 5 can also be used in the general case, leading to analogous results. LZFs continue to play a key role in the development and interpretations.

A weakness of the analysis of Chapter 7 is that the error dispersion matrix is assumed to be known, up to a constant. In Chapter 8 we relax this assumption by allowing the dispersion matrix to be known up to a few parameters. We discuss some general estimation methods in this context. Special attention is given to the cases of block diagonal error dispersion matrix, serial and spatial correlation of errors, variance components and mixed effects model. We also identify some special cases where the lack of knowledge of the error dispersion matrix may not matter very much.

We deal with updates in the general linear model in Chapter 9, as some observations are either included in or excluded from the model. We characterize the changes in terms of LZFs, which in turn throw light on the changes that take place in the statistical quantities of interest. We also outline applications of these results in various areas such as diagnostics, design and recursive prediction. Subsequently, we consider inclusion or exclusion of one or more parameters in the general linear model.

In Chapter 10, we generalize the main results of Chapter 7 to the case of the multivariate general linear model with possibly singular dispersion matrix. LZFs once again facilitate the interpretations which are similar to those given in Chapters 4, 5 and 7.

The foundational issues related to linear inference in the linear model are taken up in Chapter 11. The theoretical development parallels the general theory of inference reviewed in Chapter 3. No distributional assumption is necessary to develop this theory. The chapter is then wrapped up with a discussion of alternative linear estimators in the linear model, geometric interpretation of optimal linear estimators in the linear model (including the singular dispersion case) and large sample properties of the estimator of Chapter 7.

Topics that we left out deliberately include nonlinear methods of inference such as Bayesian, robust and rank-based methods. For information on these topics we refer the interested reader to Broemeling (1985), Rousseeuw and Leroy (1987), Boldin, Simonova and Tyurin (1997), Hettmansperger and McKean (1998) and Rao and Toutenburg (1999). This book does not include resampling methods for the linear

model which are discussed in detail by Efron and Tibshirani (1993) and Shao and Tu (1995). We also avoid detailed treatment of the generalized linear model, which can be found in McCullagh and Nelder (1989) and Dobson (2001). In regression diagnostics, we provide the statistical theory in terms of leverages, residuals, description of several plots, deletion diagnostics and indicators of collinearity in various chapters, but do not go into detailed data-analytic examples. Other data-analytic methods such as transformation of variables and treatment of missing data are also omitted. These material can be found in Belsley et al. (1980), Dodge (1985), Atkinson (1987), Cook and Weisberg (1994) and Ryan (1997), among other books.

1.8 Exercises

- 1.1 *Signal detection.* An important problem of optical and electrical communication is to detect whether a ‘signal’ is present in a sequence of observations, which is modelled as

$$y_t = a \cos(\omega t + \phi) + \epsilon_t, \quad t = 1, 2, \dots, n,$$

where t is the time index, $a \cos(2\pi f t + \phi)$ is a sinusoidal function of t (called *signal*) having amplitude a , frequency f and phase ϕ , and ϵ_t is the error at time t (called *noise*). If a is zero, the observations consist only of noise, that is, there is no signal. Assuming that the frequency is known, show that the above *signal plus noise model* with unspecified amplitude and phase reduces to (1.3.1) after a suitable transformation of the parameters. What are the explanatory variables of this linear model? Formulate the hypothesis of ‘no signal’ in terms of the parameters of the linear model.

- 1.2 The salary (y) of an employee in an organization is modelled as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon,$$

where x_1 and x_2 are binary indicators of graduation from high-school and college, respectively, x_3 is the indicator of at least