

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  $\phi$ , and  $\epsilon_t$  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

Running distance (meters)	Men's record (seconds)	Women's record (seconds)
100	9.79	10.49
200	19.32	21.34
400	43.18	47.60
800	101.11	113.28
1000	131.96	148.98
1500	206.00	230.46
2000	284.79	325.36
3000	440.67	486.11
5000	759.36	868.09
10000	1582.75	1771.78

Table 1.1 World record running times data (source: International Association of Athletics Federations, <http://www.iaaf.org/Results/Records/index.html>)

one post-graduate degree,  $x_4$  is the number of years in service and  $\epsilon$  is the error term of the model.

- What are the possible sources of the model error?
- Interpret the parameters  $\beta_0, \dots, \beta_4$ .
- Which constraints on the parameters correspond to the hypothesis: 'salary does not depend on the educational background'?

1.3 Table 1.1 gives the men's and women's world record times for various running distances, recognized by the International Association of Athletics Federations (IAAF) as of 16 August, 2002. It may be assumed that the log of the record time is approximately a linear function of the log of the running distance. Identify the matrix and vectors of (1.3.2) if a linear model is used for the men's log-record times and another one for the women's log-record times. Construct a single 'grand model' with four parameters which can be used as a substitute for these two models, and identify the corresponding matrix and vectors.

1.4 (a) If a single linear model is used for all the log-record times

Year	Population (billion)	Year	Population (billion)
1981	4.533	1991	5.367
1982	4.613	1992	5.450
1983	4.694	1993	5.531
1984	4.774	1994	5.611
1985	4.855	1995	5.691
1986	4.938	1996	5.769
1987	5.024	1997	5.847
1988	5.110	1998	5.925
1989	5.196	1999	6.003
1990	5.284	2000	6.080

Table 1.2 World population data (Source: U.S. Census Bureau, International Data Base, <http://www.census.gov/ipc/www/idbnew.html>)

for the data of Table 1.1, and the gender effect is represented by an additional (binary) explanatory variable, then identify the matrix and vectors of (1.3.2).

- (b) Identify a constraint on the parameters of the ‘grand model’ of Exercise 1.3 which would make it equivalent to the model of part (a).
- 1.5 Table 1.2 gives the midyear population of the world for the years 1981-2000. Suppose that a linear model is used to express the world population approximately in terms of the year. Identify the matrix and vectors of (1.3.2), and interpret the parameters  $\beta_0$  and  $\beta_1$ .
- 1.6 Show that if the explanatory variables are random and (1.3.2)–(1.3.3) represent a model of  $\mathbf{y}$  conditional on  $\mathbf{X}$ , then the model error  $\epsilon$  must be uncorrelated with  $\mathbf{X}$ . [See Exercise 3.7 for a stronger version of this result.]
- 1.7 Consider the *piecewise linear model*

$$y = \begin{cases} \alpha_0 + \alpha_1 x + \epsilon & \text{if } x \leq x_0, \\ \beta_0 + \beta_1 x + \epsilon & \text{if } x > x_0. \end{cases}$$

Show that if  $x_0$  is known, this model can be rewritten as a linear model — with a suitable choice of explanatory variables.

[Note: Usually the *change point*  $x_0$  is unknown, and therefore the piecewise linear model is in fact a nonlinear model.]

- 1.8 If the linear model of Exercise 1.7 is used for the world population data of Table 1.2 with  $x_0$  chosen as the year 1990, and the model is expressed as (1.3.2), identify the matrix and vectors.
- 1.9 According to the piecewise linear model of Exercise 1.7,  $E(y|x)$  may be discontinuous at  $x_0$ . Observe that the discontinuity disappears if the restriction  $\beta_0 - \alpha_0 = (\alpha_1 - \beta_1)x_0$  is imposed. Rewrite this continuous, piecewise linear model as a linear model — with a suitable choice of explanatory variables.
- 1.10 If the linear model of Exercise 1.9 is used for the world population data of Table 1.2 with  $x_0$  chosen as the year 1990, and the model is expressed as (1.3.2), identify the matrix and vectors.
- 1.11 Consider the models (1.4.2) and (1.4.3) of Example 1.4.4. Show that the errors  $\epsilon$  and  $\delta$  both have zero mean only if  $v$  and  $1/v$  are uncorrelated. Is this condition likely to hold?
- 1.12 *Cobb-Douglas model.* This model for production function postulates that the production ( $q$ ) is related to labour ( $l$ ) and capital ( $c$ ) via the equation

$$q = al^\alpha c^\beta \cdot u,$$

where  $a$ ,  $\alpha$  and  $\beta$  are unspecified constants and  $u$  is the (multiplicative) model error. The model is transformed to a linear model via a log-transformation of both sides of the equation. Assume that the additive error of the transformed model has zero mean. If  $\delta$  is defined as  $q - al^\alpha c^\beta$ , the additive error of the original model, show that this error has larger variance when the mean response of the transformed model is larger.

- 1.13 *Logistic regression model.* Suppose that the response is a binary variable whose conditional mean ( $\pi$ ) given the explanatory variables  $x_1, \dots, x_p$  is given by the equation

$$\log \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

Is this model a special case of any of the models discussed in this chapter? Can it be linearized by a suitable transformation of the response?

- 1.14 The manufacturer of a medicine for common cold claims that this medicine provides 30% longer relief than that provided by a competing brand. In order to test this claim, an experiment is conducted with a number of adult volunteers who were given a standard dose of one medicine or the other. The duration of relief was measured, and other possibly influencing factors such as gender were recorded too. Is it possible to formulate the problem in such a way that the claim amounts to a simple condition on the parameters of a linear model which may be fitted to the above data?
- 1.15 *Response surface.* Consider the quadratic regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \quad \text{Var}(\epsilon) = \sigma^2,$$

with independent observations. If  $\beta_2 > 0$ , determine the value of  $x$  which will minimize the expected response.

[See Exercise 5.8 for inference of this value from data.]

- 1.16 *Errors in variables.* Suppose that for a given value of the random explanatory variable  $x$ , the response ( $y$ ) is given by the linear model

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Suppose that  $x$  is observed with some random error, and the observation  $x_o$  is represented by the model

$$x_o = x + \delta,$$

where  $\delta$  has zero mean and is independent of  $\epsilon$  and  $x$ .

- (a) A model involving  $y$  and  $x_o$  may be obtained by eliminating  $x$  from the two equations. Show that this model is *not* a special case of (1.1.1), by calculating the correlation between the model error and  $x_o$ .
- (b) Is the model represented by the original pair of equations a special case of any of the models considered in this chapter?
- 1.17 Suppose that the effectiveness of a new drug (for which there is no competitor) is studied in the following way. A random

sample of 10 clinics is selected without replacement from all the clinics in the country, and a random sample of 10 patients is selected without replacement from these clinics. The selected patients are administered the drug and the ‘improvement in status’ is recorded. The model for this response is

$$y_{ij} = \mu + \delta_i + \epsilon_{ij}, \quad i, j = 1, \dots, 10,$$

where  $\mu$  is a constant,  $\delta_i$  is the effect of the  $i$ th clinic, and  $\epsilon_{ij}$  is a random term corresponding to the  $j$ th patient of the  $i$ th clinic. The objective of the study is to measure the average ‘improvement in status’, irrespective of the clinic. Should the  $\delta_i$ s be modelled as fixed parameters or random quantities? Which parameter of this model should be the focus of inference? Identify the model from among all those considered in this chapter for which the above model is a special case.