Chapter 1

Introduction

Linear models have a dominant role in statistical theory and practice. Most standard statistical methods are special cases of the general linear model, and rely on the corresponding theory for justification.

The goal of this course is to develop the theoretical basis for analyses based on a linear model. We shall be concerned with laying the theoretical foundation for simple as well as complex data sets.

Linear model is one of the oldest topics in the statistics curriculum. The main role of linear model in statistical practice, however, has begun to undergo a fundamental change due in large measure to available computing. Balanced experiments were often required to make analysis possible. This has produced a fundamental change in the way we can think about linear models, as much less stress can be placed on the special cases where computations are easy and more can be placed on general ideas. Topics that might have been standard, such as the recovery of interblock information in an incomplete block experiment, is of much less interest when computers can be used to appropriately maximize functions.

However, standard results are so elegant, and so interesting, that they deserve study in their own right, and for that reason we will study the traditional body of material that makes up linear models, including many standard simple models as well as a general approach.

The goal of these notes is to develop a *coordinate-free approach* to linear models. Coordinates can often sever to make problems unnecessarily complex, and understanding the features of a problems that are not dependent on coordinates is extremely valuable. The problems introduced by parameters are more easily understood given the coordinate-free background.

1.1 Some simple examples

1.1.1 One sample problem

The simplest linear model has data $y_i, i = 1, \ldots, n$ such that each y_i has the same distribution with mean μ and variance $\sigma^2 > 0$. Normality, or other distributional assumptions, are sometimes needed, but will not be used in the first few weeks of the course. In the one-sample problem, the goals are to learn about μ and possibly σ^2.

A *model* for this problem can we obtained by writing:

$$\begin{aligned} y_i &= \mu + (y_i - \mu) \\ &= \mu + \varepsilon_i \end{aligned}$$

where $\varepsilon_i = y_i - \mu$. Each observation is then taken to be the sum of a fixed part, in this case the parameter μ, and a random part ε_i, a random variable with zero mean and variance σ^2. In the spirit of this course, we will collect the responses into a vector $y = (y_1, \ldots, y_n)^T$, and the ε_i into $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Writing J_n to be a vector of length n of all ones, the one-sample model can be written as

$$y = J_n\mu + \varepsilon$$

We will soon be learning the linear algebraic background to interpret this equation. The vector on the left is any arbitrary vector in n-dimensional space. On the right we have two vectors. ε is also an arbitrary vector in n-dimensional space, while $J_n\mu$ is a vector that is *constrained* to live in a part of n-dimensional space. This will be a characteristic form of (fixed-effect) linear models.

1.1.2 One way layout

Suppose we let y_{ij} be the jth observation in the ith population, $i = 1, \ldots, p; j = 1, \ldots, n_i$ be $n = \sum n_i$ independent observations. We then specify a mean structure:

$$\begin{aligned} \mathrm{E}(y_{ij}|\text{Group} = i) &= \mu_i \\ \mathrm{Var}(y_{ij}|\text{Group} = i) &= \sigma^2 \end{aligned} \tag{1.1}$$

so each group has its own mean but a common variance. This model is (somewhat) more complex because the mean now depends on the index and is therefore conditional. In matrix terms, suppose that $p = 3$, and write $y = (y_{11}, y_{12}, \ldots, y_{3n_3})^T$

to be the vector of responses. Then we can write

$$y = \begin{pmatrix} J_{n_1} & 0 & 0 \\ 0 & J_{n_2} & 0 \\ 0 & 0 & J_{n_3} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \varepsilon \tag{1.2}$$

This is a model which is just like the one-sample model except that the description of the fixed part is more complicated. The fixed part is now in a more complex space of dimension p rather than 1.

In such a model we may wish to address several goals:

1. Estimate the cell means μ_i, and obtain estimates of uncertainty.

2. Test hypotheses such as $\mu_1 = \mu_2 = \ldots = \mu_p$, or $\mu_i = \mu_j$ or more generally $\sum \alpha_i \mu_i$ = constant, where the α_i are known numbers.

3. Estimate the index of the largest of the μ_i. A *comparative experiment* is one in which several treatments indexed here from $1, \ldots, p$, are to be compared, and the goal is to decide which is the best one or the best few. This leads to many interesting questions, in particular many questions concerning how to make inferences when faced with multiple objectives (comparing many treatments).

and so on. This model is linear because it is linear in the unknown location parameters μ_i.

Parameterizaton. A general form of the one-way model given by (1.2), is

$$y = X\beta + \varepsilon$$

where

$$X = \begin{pmatrix} J_{n_1} & 0 & 0 \\ 0 & J_{n_2} & 0 \\ 0 & 0 & J_{n_3} \end{pmatrix}; \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

This is in fact a *parametric* or *coordinate* version of a linear model because of the fixed choice of X. In fact, (1.2) is just one of many possible ways of writing the linear model for the one-way classification. If A is any $p \times p$ nonsingular matrix, meaning that there is a matrix A^{-1} such that $AA^{-1} = I$, we can write

$$\begin{aligned} y &= XAA^{-1}\beta + \varepsilon \\ &= (XA)(A^{-1}\beta) + \varepsilon \\ &= X^*\gamma + \varepsilon \end{aligned}$$

which is a completely equivalent form of this linear model, but with parameters γ rather than β. There are several different choices that are commonly used for A (we set $p = 3$ for illustration): set $\gamma = A^{-1}\beta \equiv (\alpha_0, \alpha_1, \alpha_2)^T$.

1. The three parameters are the overall mean $\alpha_0 = \mu \equiv (\mu_1 + \mu_2 + \mu_3)/3$, $\alpha_1 = \mu_1 - \mu$, and $\alpha_2 = \mu_2 - \mu$. This is the "effects" parameterization seen most often.
$$A_1 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{pmatrix} \qquad \text{What is } A_1^{-1}?$$

2. This sets the parameters to be $\alpha_0 = \mu_1$, $\alpha_1 = \mu_2 - \mu_1$ and $\alpha_2 = \mu_3 - \mu_1$. This parameterization is the default used by R.
$$A_2 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad \text{What is } A_2^{-1}?$$

3. This is called the Helmert parameterization, and is the default used by S-Plus. It is convenient for computing, but usually not convenient for interpretation.
$$A_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & -2 \end{pmatrix} \qquad \text{What is } A_3^{-1}?$$

The approach to linear models we use will try to avoid specific parameterization, since it is not relevant to many important topics.

1.1.3 One-way random effects

The one-way model we have just discussed was a *conditional* model for fixed groups. Suppose that the groups were in fact a random sample from a population of groups. Since all that changes from group to group is the mean, one way to view this problem is to assume that the μ_i are random draws from a population, with mean μ and variance τ^2. The rules for iterated mean and variance can then be applied to get the unconditional model,

$$\begin{aligned} \mathrm{E}(y_{ij}) &= \mathrm{E}\left[\mathrm{E}(y_{ij}|\text{Group} = i)\right] \\ &= \mathrm{E}\left[\mu_i\right] \\ &= \mu \end{aligned}$$

and

$$\begin{aligned}
\text{Var}(y_{ij}) &= \text{E}\left[\text{Var}(y_{ij}|\text{Group} = i)\right] + \text{Var}\left[\text{E}(y_{ij}|\text{Group} = i)\right] \\
&= \text{E}\left[\sigma^2\right] + \text{Var}\left[\mu_i\right] \\
&= \sigma^2 + \tau^2
\end{aligned}$$

Thus, the unconditional model is that $\text{E}(y_{ij}) = \mu$ but $\text{var}(y_{ij}) = \sigma^2 + \tau^2$. In addition, although the y_{ij} are conditionally independent given group, they are unconditionally correlated, since $\text{cov}(y_{ij}, y_{ik}) = \tau^2$. The simpler mean structure for the random effects model is offset by a more complex variance structure.

1.1.4 Simple linear regression

The simple linear regression model is a special case of (1.1), if we take

$$\mu_i = \beta_0 + \beta_1 x_i \tag{1.3}$$

and further assume that the x_i are known, fixed constants. The model can be written as

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}, i = 1, \ldots, p; j = 1, \ldots, n_i \tag{1.4}$$

One usually sees this model written as

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k, \text{ where } k = 1, \ldots, n = \sum n_i \tag{1.5}$$

losing the identification of observations with a population. For this model we may wish to:

1. Estimate the βs and σ^2.

2. Make tests concerning the βs, in particular of $\beta_1 = 0$.

3. Obtain interval estimates for $\beta_0 + \beta_1 x_i$. This is the *prediction problem.*

4. Examine the assumption that the cell means μ_i are linear in the xs.

and so on. You should have all seen the simple regression model in great detail, and we shall look at regression only as a special case of the general linear model.