

Chapter 2

Measurement, Estimation and Prediction

Abstract Measurement is commonly taken for granted in statistical work but, in the fields where missing observations occur, it is often the main objective. This is because the quantities to be ‘measured’ turn out to be represented by the parameters or random variables of a statistical model. Measurement then becomes a matter of predicting the values of random variables or of estimating the parameters of a distribution. When the unobserved variables are latent and, possibly indeterminate in number, the key idea is to determine their conditional distribution given what has been observed. This is essentially a routine matter involving the manipulation of probability functions. However, it is necessary to make clear what has to be defined and what are the constraints imposed by the logic of probability theory. This is important because much controversy, for example in relation to factor scores, has resulted from a failure to appreciate this point. We also introduce the one-parameter exponential family of distributions. This achieves a substantial simplification without incurring a serious loss of generality. In fact, it permits a considerable degree of unification of existing models and the development of new ones.

Keywords Conditional distributions • Estimation • Exponential family • Factor scores • Measurement • Prediction • Missing values

2.1 Measurement

In psychometrics and related branches of Science there is much discussion of measurement. In psychometrics, for example, there is the classical measurement model which supposes that what we observe differs from what we seek to measure by an ‘error’. There is no comparable theory of measurement in Statistics where the term measurement is used in less specific ways. It is important, therefore, to be clear about how the general term ‘measurement’ is linked to the standard statistical procedures.

Measurement is commonly defined as the assignment of numbers to objects in such a way that the numbers are related in ways which reflect the relationship between the objects. In one of the simplest cases, the length of objects, rods say, is reflected in the numbers which measure length. So if two rods of the same length are put end to end, the measure length of the combination will be twice that of each individual rod. It is not immediately obvious how this relates to statistical theory. The objects with which we deal in a statistical model are either parameters or random variables. The former are treated as fixed and the latter as varying in a way that can be described by a probability distribution. In Statistics the process of assigning numbers to parameters is known as *estimation* and the corresponding procedure for random variables is *prediction*. In statistical language, then, measurement is achieved by estimating unknown parameters or by providing predictors for random variables.

In the last chapter we saw that the unobserved variables in our models, the y s, could be regarded either as parameters or as random variables. We shall therefore need to consider the estimation and prediction problems to which these give rise.

2.2 Estimation

With one exception, the estimation problems posed by our models for unobserved variables are standard and straightforward and therefore require no special discussion. Thus, in the notation introduced in [Sect. 1.1](#), if the y s are to be regarded as parameters they are no different from the θ s and can, in principle at least, be estimated by standard methods. The important exception occurs with latent variable models where the number of y s may be proportional to the sample size. The asymptotic theory which is used to support the method of maximum likelihood in such cases, for example, requires the sample size to go to infinity with the number of parameters remaining fixed. In particular, this difficulty arises with the Rasch model which we shall look at in more detail in [Chap. 4](#).

2.3 Prediction

All that we can know about the random variables in a statistical model is contained in their distribution conditional on all else that is known at the time the prediction has to be made. Any prediction for a random variable, based on a single number, will then be some measure of location of that distribution—often the mean. The key step, which lies behind all subsequent analysis, is then the determination of the relevant conditional distribution. In the remainder of this chapter we shall therefore set out the theory which is common to all of the models mentioned in [Chap. 1](#) and which will be worked out in more detail in the following chapters.

2.4 Some Basic Distributional Results

All of the diverse procedures we shall meet share the same basic structure. There are two classes of variable to be distinguished: the observed, denoted by \mathbf{x} and the unobserved variables, denoted by \mathbf{y} . The model, whatever the particular application, specifies the joint probability distribution of \mathbf{x} and \mathbf{y} but any inference has to be based on \mathbf{x} alone since that is all that we can observe. The relationship between the two joint distributions is

$$f(\mathbf{x}) = \int f(\mathbf{x}, \mathbf{y}) d\mathbf{y} \quad (2.1)$$

where the integral is over the range space of \mathbf{y} and which, for reasons stated in Chapter 1, we have assumed \mathbf{y} to be continuous. For the moment, any unknown parameters on which the distributions depend are to be understood, even though they are not made explicit. It is clear that further progress depends upon being able to specify the link between \mathbf{x} and \mathbf{y} and then this must be added to the specification. Equation (2.1) may place some restrictions on what models are possible. If, for example, we factorise the joint distribution as $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})f(\mathbf{y}|\mathbf{x})$, the factor $f(\mathbf{x})$ can be taken outside the integral where it cancels with the same factor on the left hand side. This produces the trivial and otherwise obvious result that the conditional distribution of \mathbf{y} given \mathbf{x} must integrate to one. A more interesting case arises if we make the alternative factorisation $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y})f(\mathbf{x}|\mathbf{y})$, for then we have

$$f(\mathbf{x}) = \int f(\mathbf{y})f(\mathbf{x}|\mathbf{y}) d\mathbf{y} \quad (2.2)$$

It is clear from this equation that, though it does place some restrictions on the choice of the two distributions within the integral, the latter are not uniquely determined by Eq. (2.2). Once one member of the pair $\{f(\mathbf{y}), f(\mathbf{x}|\mathbf{y})\}$ is specified the other is determined by Eq. (2.2). Thus, in general, there will be infinitely many such pairs satisfying Eq. (2.2). This representation, and the associated equations, will form the starting point of almost every chapter. We shall illustrate the indeterminacy by a simple example in Chap. 3.

There is one important example of the situation we have described which is of considerable generality and widespread application, especially to latent variable models. This arises when the x s are assumed to be mutually independent, given \mathbf{y} . That is, we suppose that

$$f(\mathbf{x}|\mathbf{y}) = \prod_i f(x_i|\mathbf{y}) \quad (2.3)$$

and we let

$$f(x_i|\mathbf{y}) = F(x_i)G(\alpha_i)\exp(\alpha_i x_i) \quad (2.4)$$

with

$$\alpha_i = \alpha_i(0) + \alpha_i(1)y_1 + \alpha_i(2)y_2 + \dots + \alpha_i(m)y_m \quad (2.5)$$

The probability function in Eq. (2.4) is known as the one-parameter exponential family. The family includes both continuous and discrete distributions—among which are the normal, Poisson, gamma distributions and many others. The parameter α_i is known as the canonical parameter and we have supposed in Eq. (2.5) that it is a linear function of the unobserved variables. First, under these assumptions, we start from the conditional distribution of \mathbf{y} given \mathbf{x} , given by

$$\begin{aligned} f(\mathbf{y}|\mathbf{x}) &= \frac{f(\mathbf{x}, \mathbf{y})}{f(\mathbf{x})}, \\ &= \frac{f(\mathbf{y})f(\mathbf{x}|\mathbf{y})}{\int f(\mathbf{y})f(\mathbf{x}|\mathbf{y})d\mathbf{y}}. \end{aligned} \tag{2.6}$$

Next we substitute from Eq. (2.4) into Eq. (2.3) and then use the expression given by Eq. (2.6). If we look first at the parts which depend on the \mathbf{x} s we note that the factor $\prod \psi(x_i)$ occurs in both numerator and denominator of Eq. (2.6) and thus cancels. In the remainder, \mathbf{x} s only occur in the sums $\sum \alpha_i x_i$. So if we substitute the expression for α_i from Eq. (2.5) the sum becomes $\sum_j y_j X_j$ where $X_j = \sum_i x_i \alpha_j(i)$. It is clear, therefore, that the distribution of \mathbf{y} given \mathbf{x} depends on the \mathbf{x} s only through the m linear functions $\{X_j\}$.

As we shall see later, this result has important practical implications. It supports the widespread empirical practice of choosing linear functions of the variables as indicators of an underlying latent variable. Furthermore, it delineates the circumstances under which such a practice may be justified. A fuller account of these manipulations will be found in Bartholomew et al. (2011),

Reference

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis* (3rd ed.). Chichester, UK: Wiley.

Unobserved Variables

Models and Misunderstandings

Bartholomew, D.J.

2013, VII, 86 p. 5 illus., Softcover

ISBN: 978-3-642-39911-4