Chapter 20
Estimation lines (the so-called “regression” lines)

Section 20.1: introduction

In the preceding chapter, the correlation coefficient has been treated as a tool for detecting the existence of a straight-line statistical relationship between two variates and for measuring the intensity of such a relationship. In this chapter, the presence of a straight-line relationship will be taken advantage of for estimating the value of one of the two variates, called the predicted variate, from that of the other, called the predictor variate. For instance, the relationship between the latitude $X$ and the mean length $Y$ in mm reached at the age of four years by the small-mouth black bass *Micropterus dolomieui* is illustrated in figure 20.1.1. The analysis was done originally by Gérard Pageau (1967), an ichthyologist who kindly made these interesting data available. Since the lower left and upper right corners of this scatter diagram are empty, the mean length reached at the age of four years by the various local populations appears to decrease when the latitude increases. This phenomenon seems easy enough to interpret since the increase of latitude entails a decrease of the mean water temperature and a shortening of the annual growth season for this poikilothermal fish (of which the body temperature follows that of the environment).

![Scatter diagram](image)

**Figure 20.1.1**
Scatter diagram of the mean length $Y$ in mm reached at four years of age by the small-mouth black bass *Micropterus dolomieui* according to latitude $X$ (data compiled and analyzed by Pageau, 1967)
The simplest way of confirming whether the mean length changes when the latitude increases is perhaps to classify the data into several subgroups according to latitude and to calculate the mean length for each subgroup (figure 20.1.2). This yields a sequence of conditional means of the length of the small-mouth black bass, the condition being that the latitude is restricted to a specified interval within each subgroup. Thus, in figure 20.1.2, the data are divided into three subgroups of approximately equal sizes (comprising 12, 11 and 12 pairs of observations) and the mean lengths of these three successive subgroups are 313.1, 279.7 and 250.6 mm. Therefore, the mean length reached at the age of four years by the small-mouth black bass seems to decrease when latitude increases.

However, the preceding approach has the drawback of suggesting that the change of mean length with latitude is discontinuous, something which the original data certainly do not indicate (figure 20.1.1). The decrease of the mean length of the small-mouth black bass could be described in a more gradual manner by dividing the data into a greater number of narrower subgroups but, since the whole sample contains only 35 pairs of observations, the sizes of narrower subgroups would be very small and their means would be strongly affected by sampling fluctuations (section 9.5). Subgroup means would then vary in an irregular manner and would poorly reflect the relationship between mean length and latitude. Fortunately, it may be shown theoretically that, when two variates follow a bivariate normal distribution (section 19.2), the range of the predictor variate $X$ can be divided into a greater and greater number of narrower and narrower subgroups, and the parametric means of those subgroups all lie on a straight line of which the equation is

$$
\hat{Y} = [\mu_Y - (\sigma_{xy}/\sigma_X^2)\mu_X] + (\sigma_{xy}/\sigma_X^2)X.
$$

**Figure 20.1.2**

Mean length $Y$ in mm of the small-mouth black bass at four years of age in three subgroups of local populations at increasing latitudes.