

MthStat 465, Spring 2005, Lecture Number 25
Linear Models

1. INTRODUCTION

We will now proceed to connect statistical concepts to the problem of least squares. Earlier, given a sequence of ordered pairs $(x_1, y_1), \dots, (x_N, y_N)$ we explored various means of choosing a graph $y = f(x)$ that fit these ordered pairs best. We realized we had to define “best fit”, and criterion of least squares was described as both reasonable and computable. We now want to revisit this issue and give some precise meaning to the measurement errors we discussed.

2. LINEAR MODELS

We presume that some sequence of inputs are to be used, which we denote by x_1, x_2, \dots, x_N . Each x may in fact be a single number, an ordered pair, or, more generally, an element of an arbitrary domain. We suppose that k functions of these x_j are given, and that y_k satisfies

$$y_j = A_1 f_1(x_j) + \dots + A_k f_k(x_j).$$

The term linear is used because the equations are linear in the unknown values of A_1, A_2, \dots, A_k . The nature of the function f_1, f_2, \dots, f_k is irrelevant so long as these functions are known.

2.1. Motion in a vacuum: The accepted model for motion of a small mass under the influence of gravity is $p = a + bt + ct^2$ where p denotes position and t denotes time. Here $f_1(x) = 1$, $f_2(x) = x$ and $f_3(x) = x^2$, $A_1 = a$, $A_2 = b$ and $A_3 = c$. It is clear that at least three different times need to be used if we are to have any chance to solve for a , b and c .

2.2. Motion of a periodic wave. One model for the wave might be

$$W = A_0 + A_1 \cos(t) + A_2 \cos(2t) + A_3 \cos(3t) + B_1 \sin(t) + B_2 \sin(2t) + B_3 \sin(3t)$$

The idea would be to observe the wave at several times and find A 's and B 's.

2.3. Exponential growth and decay. The accepted model here is $z = Ae^{kt}$. This does not appear linear. However, the model is equivalent to $\ln(z) = \ln(A) + kt$, which is a linear model with $y = \ln(z)$, $A_1 = \ln(A)$ and $A_2 = k$.

2.4. Logistic Model. Suppose that $z = At/(Bt + C) = t/((B/A) + (C/A))$. Taking reciprocals we get $(1/z) = (B/A)(1/t) + (C/A)$ so we have a linear model with $y = 1/z$, $x = 1/t$, $A_1 = (C/A)$ and $A_2 = B/A$.

3. RANDOM ERRORS

Of course, we don't presume that we can measure the output y in our model completely accurately. To account for this, we presume that $\epsilon_1, \dots, \epsilon_N$ are independent random variables with the same distribution and with expected value equal to 0 and standard deviation equal to some positive constant σ .

The complete model is:

$$y_j = A_1 f_1(x_j) + \dots + A_k f_k(x_j) + \epsilon_j$$

where

- N and k are fixed positive integers with $k \leq N$;

- The functions f_1, \dots, f_k are given;
- The sequence x_1, x_2, \dots, x_N is given;
- The sequence of random errors $\epsilon_1, \dots, \epsilon_N$ has the properties described above.

The model can be given in matrix form as follows:

- Let

$$\vec{y} = \begin{bmatrix} y_1 \\ \dots \\ y_N \end{bmatrix} \quad \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \dots \\ \epsilon_N \end{bmatrix} \quad \vec{A} = \begin{bmatrix} A_1 \\ \dots \\ A_k \end{bmatrix}$$

- Let B be the matrix with N rows and k columns with $f_j(x_i)$ in the j^{th} row and i^{th} column.

Then the linear model can be written as $\vec{y} = B\vec{A} + \vec{\epsilon}$.