Warren Ewens Gregory Grant

Statisical Methods in Bioinformatics:

An Introduction

Second Edition

With 30 Figures

Springer

If possible, these equations are then solved explicitly for $\hat{\xi}_{\text{MM}}$ and $\hat{\phi}_{\text{MM}}$.

Example 2. If in the gamma distribution (1.75) both λ and k are unknown and are to be estimated, equations (1.76), (1.31) and (8.35) show that the method of moments estimators $\hat{\alpha}$ and $\hat{k}$ are found from the equations

$$\hat{\lambda}_{\text{MM}} = \frac{n\hat{k}_{\text{MM}}}{\sum X_i}, \quad n^{-1}\sum X_i^2 = \left(\frac{\hat{k}_{\text{MM}}}{\hat{\lambda}_{\text{MM}}}\right)^2 + \frac{\hat{k}_{\text{MM}}}{\hat{\lambda}^2_{\text{MM}}}. \tag{8.36}$$

The values of $\hat{\lambda}_{\text{MM}}$ and $\hat{k}_{\text{MM}}$ are readily found from these equations (see Problem 8.9).

We now compare these estimators with the corresponding maximum likelihood estimators. The maximum likelihood estimator of k, namely $\hat{k}_{\text{MLE}}$, is independent of λ and is thus given by (8.34). The maximum likelihood estimator of λ, namely $\hat{\lambda}_{\text{MLE}}$, is $n\hat{k}_{\text{MLE}}/\sum X_i$. This equation is of the same form as the first equation in (8.36), implying that when $\hat{k}_{\text{MLE}}$ and $\hat{k}_{\text{MM}}$ are close, then $\hat{\lambda}_{\text{MLE}}$ and $\hat{\lambda}_{\text{MM}}$ are also close.

8.4.3 Least Squares and Multiple Regression

In programma.

Another estimation procedure, which in some cases is equivalent to the maximum likelihood method, is that of least squares. We illustrate it in the context of the general linear model, with which it is most closely associated.

In describing the least squares approach it is convenient to depart from our standard convention and to use the notation Y for a random variable, whether it be discrete or continuous. Suppose first that $Y_1, Y_2, \ldots, Y_n$ are independently but not identically distributed random variables, Y_j having a probability distribution with mean of the form $\mu_j = \alpha + \beta x_j$ and variance σ^2. This model is most conveniently written in the form

$$Y_j = \alpha + \beta x_j + E_j, \tag{8.37}$$

where $E_1, E_2, \ldots, E_n$ are iid random variables with mean 0 and variance σ^2. The model is most frequently used when one wishes to estimate the way in which some random variable Y_j depends on some fixed quantity x_j. This is the *simple regression model*, and is used very widely in applied statistics.

The form of equation (8.37) explains the choice of the notation Y for the random variable involved in least squares calculations, since if the term E_j is ignored, this is the equation of a straight line in the standard cartesian form $y = mx + b$.

In the model (8.37), α and β are unknown parameters that we might wish to estimate. The least squares estimators of α and β are found as the values that minimize the sum of squares

$$\sum_{j=1}^{n} E_j^2 = \sum_{j=1}^{n} \left(Y_j - \alpha - \beta x_j\right)^2. \tag{8.38}$$

The resulting least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ are given explicitly by

$$\hat{\beta} = \frac{\sum_{j=1}^{n} Y_j(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \tag{8.39}$$

where $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$, $\bar{Y} = (Y_1 + Y_2 + \cdots + Y_n)/n$. The estimators $\hat{\beta}$ and $\hat{\alpha}$ are unbiased (see Problem 8.10).

The fact that an explicit expression is available for both $\hat{\alpha}$ and $\hat{\beta}$ should not pass without comment. If the mean of Y_j were not a linear function of α and β it might not be possible to find explicit expressions for $\hat{\alpha}$ and $\hat{\beta}$, and the best that can be done might be to find $\hat{\alpha}$ and $\hat{\beta}$ by a purely numerical procedure. We take up this comment again below.

Given observed values $y_1, y_2, \ldots, y_n$ of $Y_1, Y_2, \ldots, Y_n$, the estimates of β and α are, respectively,

$$\hat{\beta} = \frac{\sum_{j=1}^{n} y_j(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{8.40}$$

Here we have abused notation and, for purposes of typographical clarity, have used the same symbol for the estimators and the estimates of α and β.

If $Y_1, Y_2, \ldots, Y_n$ are independent normal random variables, each having variance σ^2 and with Y_j having mean $\alpha + \beta x_j$, the maximum likelihood estimators of α and β are the least squares estimators (8.39) of these parameters (see Problem 8.11).

The model described above assumes that the various Y_j random variables all have the same variance. If some of the random variables have variances greatly exceeding that of the remaining random variables, the estimates of the parameters might be unduly influenced by those random variables with a large variance. In this case it might be thought desirable to minimize the weighted sum of squares $\sum_{j=1}^{n} w_j(Y_j - \alpha - \beta x_j)^2$ rather than the unweighted sum in (8.38), where w_j is a weighting factor associated with Y_j and is small for those random variables with a large variance. If we use the suffix "w" for the weighted least squares estimates of α and β, these estimates are given by

$$\begin{aligned} \hat{\beta}_w &= \frac{(\sum w_j)(\sum w_j x_j y_j) - (\sum w_j y_j)(\sum w_j x_j)}{(\sum w_j)(\sum w_j x_j^2) - (\sum w_j x_j)^2}, \\ \hat{\alpha}_w &= \frac{\sum w_j y_j - \hat{\beta}_{\mathrm{w}} \sum w_j x_j}{\sum w_j}, \end{aligned} \tag{8.41}$$

all sums being over $j = 1, 2, \ldots, n$.

No

A further application of weighted least squares estimation is in the construction of *loess* curves, discussed in detail by Cleveland and Devlin (1988), following the earlier work of Cleveland (1979). (The word "loess" is an acronym (LOcally weighted regrESSion), and was chosen by Cleveland and Devlin (1988), because of its use in describing geological strata. The spelling "lowess" occurs often in the literature. Since here we describe the work of Cleveland and Devlin (1988), we adopt their spelling convention.)

Suppose that the relation between Y and x is nonlinear. Clearly any linear estimation procedure, weighted or unweighted, is inappropriate. On the other hand, a collection of linear estimation procedures, each one carried out over a short range of x values, might be reasonable. Further, it might be desirable that in any such local regression centered around the value x, higher weights are given to values of x_i close to x than to values further from x. With these aims in mind, Cleveland and Devlin (1988), suggest the following procedure.

We first consider some particular value of x, say x_j. We choose some number d and weighting factors $w_{j-d}, w_{j-d+1}, \ldots, w_{j+d}$ and carry out a weighted regression of $Y_{j-d}, Y_{j-d+1}, \ldots, Y_{j+d}$ on $x_{j-d}, x_{j-d+1}, \ldots, x_{j+d}$, using these weights. Cleveland and Devlin (1988) suggest values of d and forms of the weights that lead to suitable loess curves.

This procedure will lead to regression estimates $\hat{\beta}_{w,j}$ and $\hat{\alpha}_{w,j}$, for the weighted regression centered on x_j. The observed value y_j is then replaced by $y_j^* = \hat{\alpha}_{w,j} + \hat{\beta}_{w,j} x_j$, the value of Y corresponding to x_j predicted by this (short) weighted least-squares line. This entire procedure is then carried out for each value of j, with special calculations at boundary values where $j < d$ and $j > n - d$. The various values of the y_j^* so found are now joined to form a loess curve, which will generally be far smoother than the curve joining the original y_j values and provide a better fit to the data than an ordinary linear regression.

We return to loess curves in Section 13.1.3, where their use in connection with microarray analysis is discussed.

A second generalization of (8.37) arises when the mean μ_j of Y_j is of the form $\alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk}$, for some collection of known constants $x_{j1}, x_{j2}, \ldots, x_{jk}$, so that we write

$$Y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + E_j, \quad j = 1, 2, \ldots, n. \tag{8.42}$$

Here α, β_1, β_2 , ..., β_k are unknown parameters that we wish to estimate, and in the unweighted case the $E_j, j = 1, 2, \ldots, n$ are assumed to be iid random variables with mean 0 and variance σ^2. This is the *multiple regression*, or *general linear*, model, and is important in many statistical procedures. A particular case of this model is the polynomial regression model, for which x_{ji} is of the form $(x_j)^i$.

No

The least squares estimators of $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are those which minimize the (unweighted) sum of squares $\sum_{j=1}^{n} E_j^2$. To find these estimators it is convenient to write the multiple regression model (8.42) in the matrix and vector form

$$\mathbf{Y} = C\boldsymbol{\beta} + \mathbf{E}, \tag{8.43}$$

Here $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$, $\mathbf{E} = (E_1, E_2, \ldots, E_n)'$, $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2, \ldots, \beta_k)'$ and C is an $n \times (k + 1)$ matrix whose first column consists of 1's and