smfsoln4(1617)

SMF SOLUTIONS 4. 16.3.2017

Q1. To fit a straight line $y = a + bx$ by least squares through a data set (x_1, y_1), ..., (x_n, y_n), we choose a, b so as to minimise

$$SS := \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

Taking $\partial SS/\partial a = 0$ and $\partial SS/\partial b = 0$ gives

$$\begin{aligned} \partial SS/\partial a := -2\sum_{i=1}^n e_i &= -2\sum_{i=1}^n (y_i - a - bx_i), \\ \partial SS/\partial b := -2\sum_{i=1}^n x_i e_i &= -2\sum_{i=1}^n x_i(y_i - a - bx_i). \end{aligned}$$

To find the minimum, we equate both these to zero:

$$\sum_{i=1}^n (y_i - a - bx_i) = 0, \quad \text{and} \quad \sum_{i=1}^n x_i(y_i - a - bx_i) = 0.$$

This gives the *normal equations*. Using bar notation, $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$, dividing both sides by n and rearranging, these are

$$a + b\bar{x} = \bar{y}, \quad \text{and} \quad a\bar{x} + b\overline{x^2} = \overline{xy}.$$

Multiply the first by $\bar{x}$ and subtract from the second:

$$b = (\overline{xy} - \bar{x}\bar{y})/(\overline{x^2} - (\bar{x})^2), \quad \text{and then} \quad a = \bar{y} - b\bar{x}.$$

Again using bar notation, the *sample variance* s_x^2 or s_{xx} is defined as the average, $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$, of $(x_i - \bar{x})^2$. Then by linearity of average, or 'bar',

$$s_x^2 = s_{xx} = \overline{(x - \bar{x})^2} = \overline{x^2 - 2x.\bar{x} + \bar{x}^2} = \overline{(x^2)} - 2\bar{x}.\bar{x} + (\bar{x})^2 = \overline{(x^2)} - (\bar{x})^2.$$

The *sample covariance* s_{xy} of x and y is the average of $(x - \bar{x})(y - \bar{y})$. So

$$s_{xy} = \overline{(x - \bar{x})(y - \bar{y})} = \overline{xy - x.\bar{y} - \bar{x}.y + \bar{x}.\bar{y}} = \overline{(xy)} - \bar{x}.\bar{y} - \bar{x}.\bar{y} + \bar{x}.\bar{y} = \overline{(xy)} - \bar{x}.\bar{y}.$$

Thus the slope b is given by $b = s_{xy}/s_{xx}$, the ratio of the sample covariance to the sample x-variance.

Q2. With two regressors u and v and response variable y, given a sample of size n of points $(Uu_1, v_1, y_1), \ldots, (u_n, v_n, y_n)$ we have to fit a least-squares *plane* – that is, choose parameters a, b, c to minimise the sum of squares

$$SS := \sum_{i=1}^n (y_i - c - au_i - bv_i)^2.$$

Taking $\partial SS/\partial c = 0$ gives

$$\sum_{i=1}^n (y_i - c - au_i - bv_i) = 0 : \qquad c = \bar{y} - a\bar{u} - b\bar{v}.$$

We re-write SS as

$$SS = \sum_{i=1}^{n} [(y_i - \bar{y}) - a(u_i - \bar{u}) - b(v_i - \bar{v})]^2.$$

Then $\partial SS/\partial a = 0$ and $\partial SS/\partial b = 0$ give

$$\sum_{i=1}^{n} (u_i - \bar{u})[(y_i - \bar{y}) - a(u_i - \bar{u}) - b(v_i - \bar{v})],$$

$$\sum_{i=1}^{n} (v_i - \bar{v})[(y_i - \bar{y}) - a(u_i - \bar{u}) - b(v_i - \bar{v})].$$

Multiply out, divide by n to turn the sums into averages, and re-arrange using our earlier notation: these become

$$as_{uu} + bs_{uv} = s_{yu},$$

$$as_{uv} + bs_{vv} = s_{yv}.$$

These are the *normal equations* for a and b. The determinant is

$$s_{uu}s_{vv} - s_{uv}^2 = s_{uu}s_{vv}(1 - r_{uv}^2)$$

(as $r_{uv} := s_{uv}/(s_u.s_v)$), $\neq 0$ iff $r_{uv} \neq \pm 1$, i.e., iff the (u_i, v_i) are not collinear, and this is the condition for the normal equations to have a unique solution.

Q3. Grain traders need to monitor spring weather, using the weather reports during the growing season to calculate amounts of spring rainfall and spring sunshine, u and v say. Established firms will have from past records estimates of the variances of u, v and their covariance with yield y. This can be used in the normal equations of Q2 to obtain a *prediction* in the spring of harvest yields several months later in the summer. This from past knowledge of price, supply and demand will enable a prediction of price. This will guide traders in their trading strategy: purchase of options, futures etc.

In the Great Grain Steal of 1972, the USSR foresaw that it would have a harvest failure, and would need to make massive grain purchases in the US and Canadian markets. Large trades move markets, and the risk was that this would drive prices through the roof. By careful planning, the USSR buyers were able to coordinate a large number of simultaneous purchases, of moderate size, and the deed was done before the market could react.
Note. This inspired Frederick Forsyth's novel *The devil's alternative* (1979).

Q4.

$$y_i = \sum_{j=1}^{p} a_{ij}\beta_j + \epsilon_i, \qquad \epsilon_i \quad iid \quad N(0, \sigma^2),$$

the likelihood and log-likelihood are

$$\begin{aligned} L &= \frac{1}{\sigma^n 2\pi^{\frac{1}{2}n}} . \prod_{i=1}^{n} \exp\{-\frac{1}{2}(y_i - \sum_{j=1}^{p} a_{ij}\beta_j)^2/\sigma^2\} \\ &= \frac{1}{\sigma^n 2\pi^{\frac{1}{2}n}} . \exp\{-\frac{1}{2}\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} a_{ij}\beta_j)^2/\sigma^2\}, \end{aligned}$$

$$\ell := \log L = const - n \log \sigma - \frac{1}{2}[\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} a_{ij}\beta_j)^2]/\sigma^2. \qquad (*)$$

Maximise w.r.t. β_r in $(*)$ (Fisher, MLE) – equivalently, minimise [...]: $\partial\ell/\partial\beta_r = 0$ (Least Squares):

$$\sum_{i=1}^{n} a_{ir}(y_i - \sum_{j=1}^{p} a_{ij}\beta_j) = 0 \qquad (r = 1, \ldots, p):$$

$$\sum_{j=1}^{p}(\sum_{i=1}^{n} a_{ir}a_{ij})\beta_j = \sum_{i=1}^{n} a_{ir}y_i.$$

Write $C = (c_{ij})$ for the $p \times p$ matrix $C := A^T A$, which we note is *symmetric*: $C^T = C$. Then

$$c_{ij} = \sum_{k=1}^{n}(A^T)_{ik}A_{kj} = \sum_{k=1}^{n} a_{ki}a_{kj}.$$

So this says

$$\sum_{j=1}^{p} c_{rj}\beta_j = \sum_{i=1}^{n} a_{ir}y_i = \sum_{i=1}^{n}(A^T)_{ri}y_i.$$

In matrix notation, this is

$$(C\beta)_r = (A^T y)_r \qquad (r = 1, \ldots, p): \qquad C\beta = A^T y, \qquad C := A^T A. \quad (NE)$$

These are the *normal equations*. As A ($n \times p$, with $p << n$) has full rank, A has rank p, so $C := A^T A$ has rank p, so is non-singular. So the normal equations have solution

$$\hat{\beta} = C^{-1}A^T y = (A^T A)^{-1}A^T y.$$

Multiplying both sides by A,

$$Py = A(A^T A)^{-1}A^T y = A\hat{\beta}.$$

NHB