smfl1.tex
Lecture 1. 17.1.2011

I. REGRESSION

1. LEAST SQUARES

The idea of regression is to take some sample of size n from some unknown population (typically n is large – the larger the better), and seek how best to represent it in terms of a smaller number of variables, typically involving p parameters (p to be kept as small as possible, to give a parsimonious representation of the data – so p is much smaller than n, $p << n$). Usually we will have p explanatory variables, and represent the data as a linear combination of them (the coefficients being the parameters) plus some random error, as best we can. To do this, we use the *method of least squares*, and choose the coefficients so as to minimise the sum of squares (SS) of the differences between the observed data points and the linear combination. This gives us a fitted value; what is left over is called a residual; thus

$$\textit{data} = \textit{true value} + \textit{error} = \textit{fitted value} + \textit{residual}.$$

If the data forms an n-vector y and the parameters form a p-vector β, the model equation is

$$y = A\beta + \epsilon,$$

where A is an $n \times p$ matrix of constants (the *design matrix*), and ϵ is an n-vector of errors. In the full-rank case (where A has rank p), it can be shown ([BF], 3.1) that the *least-squares estimates* (LSEs) of β are

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

and (Gauss-Markov Theorem) that this gives the minimum-variance unbiased (= 'best') linear estimator (or BLUE): in this sense *least-squares is best*.

Geometrically, the Method of Least Squares projects n-dimensional reality onto the best approximating p-dimensional subspace. Indeed, the key role is played by the *projection matrix* $P = A(A^T A)^{-1} A^T$ (or $P = AC^{-1}A^T$ with $C := A^T A$ the *information matrix*; P is $n \times n$, C is $p \times p$). P is also called the *hat matrix*, H, as it projects the data y onto the fitted values $\hat{y} = A\hat{\beta}$.

To make good statistical sense of this, we need a statistical model for the error structure. We will use the *multivariate normal* distribution (Section 3),

whose estimation theory follows in Section 4.

The most basic case is $p = 2$, where one fits a line (two parameters, slope and intercept) through n data points in the plane. One can show (see e.g. [BF], 1.2) that the least-squares (best) line is

$$y = a + bx, \quad b = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} = s_{xy}/s_{xx} = r_{xy}s_y/s_x, \quad a = \overline{y} - b\overline{x}.$$

(here s_{xy} is the sample covariance between x and y, $s_{xx} = s_x^2$ is the sample variance of x, $r_{xy} = s_{xy}/(s_x s_y)$ the sample correlation coefficient). This is the *sample regression line*. By LLN, its large-sample limit is the *(population) regression line*,

$$y = \alpha + \beta x, \quad \beta = \rho\sigma_2/\sigma_1, \quad \alpha = Ey - \beta Ex: \quad y - Ey = (\rho\sigma_2/\sigma_1)(x - Ex).$$

The multivariate normal reduces in this case to the *bivariate normal* in Section 2; we treat this in full because of its fundamental importance and of how well it illustrates the general case.

Motivating examples:

1. *CAPM*. The capital asset pricing model looks at individual risky assets and compares them with 'the market', or some proxy for it such as an index. One seeks to 'pick winners' by maximising 'beta', or the slope of the linear trend of asset price versus market price.

2. *Examination scores* (BF, 1.4). Here x is the 'incoming score' of an entrant to an elite academic programme, y is the 'graduating score'; the question is how well does the institution pick its intake (i.e., how well does x predict y).

3. *Galton's height data* (BF, 1.3). Here y = offspring's height (adult sons, say), x = average of parents' heights.

2. THE BIVARIATE NORMAL DISTRIBUTION

Recall two of the key ingredients of statistics:

a. *The normal distribution*, $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(x - \mu)^2/\sigma^2\},$$

which has mean $EX = \mu$ and variance $varX = \sigma^2$.

b. *Linear regression by the method of least squares.* This is for *two-dimensional* (or bivariate) data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Two questions arise: (i) Why linear? (ii) What (if any) is the two-dimensional analogue of the normal law?

Mathematical preliminaries. Writing

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}$$

for the standard normal density, $\int$ for $\int_{-\infty}^{\infty}$, we shall need
(i) *recognising normal integrals:* (a) $\int \phi(x)dx = 1$ (*'normal density'*, (b) $\int x\phi(x)dx = 0$ (*'normal mean'* - or, 'symmetry'), (c) $\int x^2\phi(x)dx = 1$ (*'normal variance'*),
(ii) *completing the square*: as for solving quadratic equations!

In view of the work above, we need an analogue in *two* dimensions of the normal distribution $N(\mu, \sigma^2)$ in one dimension. Just as in one dimension we need *two* parameters, μ and σ, in two dimensions we must expect to need *five*, by above.

Consider the following bivariate density:

$$f(x, y) = c \exp\{-\frac{1}{2}Q(x, y)\},$$

where c is a constant, Q a positive definite quadratic form in x and y:

$$c = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, \quad Q = \frac{1}{1-\rho^2}\Big[\Big(\frac{x-\mu_1}{\sigma_1}\Big)^2 - 2\rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big(\frac{y-\mu_2}{\sigma_2}\Big) + \Big(\frac{y-\mu_2}{\sigma_2}\Big)^2\Big].$$

Here $\sigma_i > 0$, μ_i are real, $-1 < \rho < 1$. Since f is clearly non-negative, to show that f is a (probability) density (function) (in two dimensions), it suffices to show that f integrates to 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1, \qquad \text{or} \qquad \int\int f = 1.$$

Write

$$f_1(x) := \int_{-\infty}^{\infty} f(x, y)dy, \qquad f_2(y) := \int_{-\infty}^{\infty} f(x, y)dx.$$

Then to show $\int\int f = 1$, we need to show $\int_{-\infty}^{\infty} f_1(x)dx = 1$ (or $\int_{-\infty}^{\infty} f_2(y)dy = 1$). Then f_1, f_2 are densities, in *one* dimension. If $f(x, y) = f_{X,Y}(x, y)$ is the *joint* density of *two* random variables X, Y, then $f_1(x)$ is the density $f_X(x)$ of X, $f_2(y)$ the density $f_Y(y)$ of Y (f_1, f_2, or f_X, f_Y, are called the *marginal* densities of the *joint* density f, or $f_{X,Y}$).

To perform the integrations, we have to *complete the square*. We have

$$(1-\rho^2)Q \equiv \Big[\Big(\frac{y-\mu_2}{\sigma_2}\Big) - \rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big]^2 + (1-\rho^2)\Big(\frac{x-\mu_1}{\sigma_1}\Big)^2$$

(reducing the number of occurrences of y to 1, as we intend to integrate out y first). Then (taking the terms free of y out through the y-integral)

$$f_1(x) = \frac{\exp(-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2)}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\Big(\frac{-\frac{1}{2}(y-c_x)^2}{\sigma_2^2(1-\rho^2)}\Big)dy, \tag{$*$}$$

where

$$c_x := \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

The integral is 1 ('normal density'). So

$$f_1(x) = \frac{\exp(-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2)}{\sigma_1\sqrt{2\pi}},$$

which integrates to 1 ('normal density'), proving

Fact 1. $f(x,y)$ is a joint density function (two-dimensional), with marginal density functions $f_1(x), f_2(y)$ (one-dimensional). So we can write

$$f(x,y) = f_{X,Y}(x,y), \qquad f_1(x) = f_X(x), \qquad f_2(y) = f_Y(y).$$

Fact 2. X, Y are normal: X is $N(\mu_1, \sigma_1^2)$, Y is $N(\mu_2, \sigma_2^2)$. For, we showed $f_1 = f_X$ to be the $N(\mu_1, \sigma_1^2)$ density above, and similarly for Y by symmetry.

Fact 3. $EX = \mu_1, EY = \mu_2, varX = \sigma_1^2, varY = \sigma_2^2$.

This identifies four of the five parameters: two means μ_i, two variances σ_i^2.

Next, recall the definition of conditional probability:

$$P(A|B) := P(A \cap B)/P(B).$$

In the *discrete* case, if X, Y take possible values x_i, y_j with probabilities $f_X(x_i), f_Y(y_j)$, (X,Y) takes possible values (x_i, y_j) with probabilities $f_{X,Y}(x_i, y_j)$:

$$f_X(x_i) = P(X = x_i) = \Sigma_j P(X = x_i, Y = y_j) = \Sigma_j f_{X,Y}(x_i, y_j).$$

Then the *conditional* distribution of Y *given* $X = x_i$ is

$$f_{Y|X}(y_j|x_i) = P(Y = y_j \;\&\; X = x_i)/P(X = x_i) = f_{X,Y}(x_i, y_j)/\Sigma_j f_{X,Y}(x_i, y_j).$$

In the *density* case, we have to replace *sums* by *integrals*. Thus the conditional *density* of Y *given* $X = x$ is

$$f_{Y|X}(y|x) := f_{X,Y}(x,y)/f_X(x) = f_{X,Y}(x,y)/\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy.$$

Returning to the bivariate normal:

Fact 4. The conditional distribution of y given $X = x$ is $N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),\ \sigma_2^2(1-\rho^2))$.