smfd7.tex
Day 7. 7.11.2014.

IV. 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 a known $n \times p$ 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* (I.5). 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?

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)

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

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

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

Fact 4. The conditional distribution of y given $X = x$ is

$$N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),\ \sigma_2^2(1 - \rho^2)).$$

Proof. Go back to completing the square (or, return to (*) with $\int$ and dy deleted):

$$f(x,y) = \frac{\exp(-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2)}{\sigma_1\sqrt{2\pi}} . \frac{\exp(-\frac{1}{2}(y-c_x)^2/(\sigma_2^2(1-\rho^2)}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}}.$$

The first factor is $f_1(x)$, by Fact 2. So, $f_{Y|X}(y|x) = f(x,y)/f_1(x)$ is the second factor:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\{-\frac{1}{2}(y-c_x)^2/(\sigma_2^2(1-\rho^2))\},$$

where c_x is the linear function of x given below (*). This gives Fact 4, and
Fact 5. The conditional mean $E(Y|X=x)$ is *linear* in x:

$$E(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

Note. This simplifies when X and Y are equally variable, $\sigma_1 = \sigma_2$:

$$E(Y|X=x) = \mu_2 + \rho(x-\mu_1)$$

(recall $EX = \mu_1, EY = \mu_2$). Recall that in Galton's height example, this says: for every inch of mid-parental height above/below the average, $x - \mu_1$, the parents pass on to their child, *on average*, ρ inches, and continuing in this way: *on average*, after n generations, each inch above/below average becomes *on average* ρ^n inches, and $\rho^n \to 0$ as $n \to \infty$, giving *regression towards the mean.*
(A regression function is a *conditional mean* – see Section 5.)
Fact 6. The conditional variance of Y given $X = x$ is

$$var(Y|X=x) = \sigma_2^2(1-\rho^2).$$

Recall (Fact 3) that the variability (= variance) of Y is $varY = \sigma_2^2$. By Fact 5, the variability remaining in Y when X is given (i.e., not accounted for by knowledge of X) is $\sigma_2^2(1-\rho^2)$. Subtracting: the variability of Y which *is* accounted for by knowledge of X is $\sigma_2^2\rho^2$. That is: ρ^2 is the *proportion of the variability* of Y accounted for by knowledge of X. So ρ is a measure of the *strength of association* between Y and X.

Recall that the *covariance* is defined by

$$cov(X,Y) := E[(X-EX)(Y-EY)] = E[(X-\mu_1)(Y-\mu_2)] = E(XY)-(EX)(EY),$$

and the *correlation coefficient* ρ, or $\rho(X,Y)$, defined by

$$\rho = \rho(X,Y) := cov(X,Y)/(\sqrt{varX}\sqrt{varY}) = E[(X-\mu_1)(Y-\mu_2)]/\sigma_1\sigma_2$$

is the usual measure of the strength of association between X and Y ($-1 \le \rho \le 1$; $\rho = \pm 1$ iff one of X, Y is a function of the other).
Fact 7. The correlation coefficient of X, Y is ρ.
Proof.

$$\rho(X,Y) := E\Big[\Big(\frac{X-\mu_1}{\sigma_1}\Big)\Big(\frac{Y-\mu_2}{\sigma_2}\Big)\Big] = \int\int \Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big(\frac{y-\mu_2}{\sigma_2}\Big) f(x,y)dxdy.$$

Substitute for $f(x,y) = c\exp(-\frac{1}{2}Q)$, and make the change of variables $u := (x-\mu_1)/\sigma_1$, $v := (y-\mu_2)/\sigma_2$:

$$\rho(X,Y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int\int uv \exp\{-\frac{1}{2}[u^2 - 2\rho uv + v^2]/(1-\rho^2)\} du dv.$$

Completing the square, $[u^2 - 2\rho uv + v^2] = (v-\rho u)^2 + (1-\rho^2)u^2$. So

$$\rho(X,Y) = \frac{1}{\sqrt{2\pi}} \int u \exp(-\frac{1}{2}u^2) du . \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \int v \exp\{-\frac{1}{2}(v-\rho u)^2/(1-\rho^2)\} dv.$$

Replace v in the inner integral by $(v-\rho u)+\rho u$, and calculate the two resulting integrals separately. The first is zero ('normal mean', or symmetry), the second is ρu ('normal density'). So

$$\rho(X,Y) = \frac{1}{\sqrt{2\pi}} . \rho \int u^2 \exp(-\frac{1}{2}u^2) du = \rho$$

('normal variance'), as required.

This completes the identification of all five parameters in the bivariate normal distribution: two means μ_i, two variances σ_i^2, one correlation ρ.
Note 1. The above holds for $-1 < \rho < 1$; always, $-1 \leq \rho \leq 1$. When $\rho = \pm 1$, one of X, Y is a linear function of the other, as with temperature (Fahrenheit and Centigrade). This is not really two-dimensional: we can (and should) use only *one* of X and Y, and reduce to one dimension.
Note 2. The slope of the regression line $y = c_x$ is $\rho\sigma_2/\sigma_1 = (\rho\sigma_1\sigma_2)/(\sigma_1^2)$, which can be written as $cov(X,Y)/varX = \sigma_{12}/\sigma_{11}$, or σ_{12}/σ_1^2: the line is

$$y - EY = \frac{\sigma_{12}}{\sigma_{11}}(x - EX).$$

This is the *population* version (what else?!) of the *sample regression line*

$$y - \bar{Y} = \frac{S_{XY}}{S_{XX}}(x - \bar{X}),$$

from linear regression (Section 1).

The case $\rho = \pm 1$ – apparently two-dimensional, but really one-dimensional – is *singular*; the case $-1 < \rho < 1$ - genuinely two-dimensional - is *non-singular*, or (III) *full rank*. We note in passing
Fact 8. The bivariate normal law has *elliptical contours*. For, the contours

are $Q(x,y) = const$, which are ellipses (as Galton found).
Characteristic Function (CF) and Moment Generating Function (MGF).

Recall the CF $\phi(t) := E[e^{itX}]$ and MGF $M(t) := E[e^{tX}]$. For X $N(\mu, \sigma^2)$, $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$ [SP, Problems 5]. So $M_{X-\mu}(t) = \exp(\frac{1}{2}\sigma^2 t^2)$, and the CF is $\phi_{X-\mu}(t) = \exp(-\frac{1}{2}\sigma^2 t^2)$. Then (check) $\mu = EX = M'_X(0)$, $varX = E[(X-\mu)^2] = M''_{X-\mu}(0)$. Similarly in the bivariate case:

$$\phi_{X,Y}(t_1,t_2) := E[\exp\{i(t_1X+t_2Y)\}], \quad M_{X,Y}(t_1,t_2) := E[\exp\{i(t_1X+t_2Y)\}].$$

For the bivariate normal,

$$\phi(t_1,t_2) = E[\exp\{i(t_1X+t_2Y)\}] = \int\int \exp\{i(t_1x+t_2y)\}f(x,y)dxdy$$

$$= \int \exp\{it_1x\}f_1(x)dx \int \exp\{it_2y\}f(y|x)dy.$$

The inner integral is the CF of $Y|X=x$, which is $N(c_x, \sigma_2^2, (1-\rho^2))$, so is $\exp(ic_xt_2 - \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2)$. By Fact 4, $c_xt_2 = [\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)]t_2$, so

$$\phi(t_1,t_2) = \exp\{i(t_2\mu_2 - t_2\frac{\sigma_2}{\sigma_1}\mu_1 - \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2)\}\int \exp\{i([t_1+t_2\rho\frac{\sigma_2}{\sigma_1}]x)\}f_1(x)dx.$$

Since $f_1(x)$ is $N(\mu_1, \sigma_1^2)$, the inner integral is a normal CF, which is thus $\exp\{i(\mu_1[t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}] - \frac{1}{2}\sigma_1^2[\ldots]^2)\}$. Combining the two terms and simplifying:
Fact 9. The joint MGF and joint CF of X, Y are

$$M_{X,Y}(t_1,t_2) = M(t_1,t_2) = \exp(\mu_1t_1 + \mu_2t_2 + \frac{1}{2}[\sigma_1^2t_1^2 + 2\rho\sigma_1\sigma_2t_1t_2 + \sigma_2^2t_2^2]),$$

$$\phi_{X,Y}(t_1,t_2) = \phi(t_1,t_2) = \exp(i\mu_1t_1 + i\mu_2t_2 - \frac{1}{2}[\sigma_1^2t_1^2 + 2\rho\sigma_1\sigma_2t_1t_2 + \sigma_2t_2^2]).$$

Fact 10. X, Y are independent if and only if $\rho = 0$.
Proof. For densities: X, Y are independent iff the joint density $f_{X,Y}(x,y)$ *factorises* as the *product* of the marginal densities $f_X(x).f_Y(y)$. For MGFs, CFs: X, Y are independent iff the joint MFG $M_{X,Y}(t_1,t_2)$, or CF, *factorises* as the *product* of the marginals. From Fact 9, this occurs iff $\rho = 0$..
Note. X, Y independent implies X, Y uncorrelated ($\rho = 0$) in general (when the correlation exists). The converse if *false* in general, but *true*, by Fact 10, in the bivariate normal case.

3. The Multivariate Normal Distribution.

With one regressor, we used the bivariate normal distribution as above. Similarly for two regressors, we use the trivariate normal. With any number of regressors, as here, we need a general *multivariate normal* - or *'multinormal'* - distribution in n dimensions. We must expect that in n dimensions, to handle a random n-vector $\mathbf{X} = (X_1, \cdots, X_n)^T$, we will need
(i) a *mean vector* $\mu = (\mu_1, \cdots, \mu_n)^T$ with $\mu_i = EX_i$, $\mu = E\mathbf{X}$,
(ii) a *covariance matrix* $\mathbf{\Sigma} = (\sigma_{ij})$, with $\sigma_{ij} = cov(X_i, X_j)$: $\mathbf{\Sigma} = cov\mathbf{X}$.

First, note the effect of a linear transformation:

Proposition 1. If $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$, with $\mathbf{Y}, \mathbf{b}$ m-vectors, $\mathbf{A}$ an $m \times n$ matrix and $\mathbf{X}$ an n-vector,
(i) the mean vectors are related by $E\mathbf{Y} = \mathbf{A}E\mathbf{X} + \mathbf{b} = \mathbf{A}\mu + \mathbf{b}$,
(ii) the covariance matrices are related by $\mathbf{\Sigma_Y} = \mathbf{A\Sigma A}^T$.

Proof. (i) This is just linearity of E: $Y_i = \sum_j a_{ij}X_j + b_i$, so

$$EY_i = \sum\nolimits_j a_{ij}EX_j + b_i = \sum\nolimits_j a_{ij}\mu_j + b_i,$$

for each i. In vector notation, this is $\mu_{\mathbf{Y}} = \mathbf{A}\mu + \mathbf{b}$.
(ii) $Y_i - EY_i = \sum_k a_{ik}(X_k - EX_k) = \sum_k a_{ik}(X_k - \mu_k)$, so

$$cov(Y_i, Y_j) = E[\sum\nolimits_r a_{ir}(X_r - \mu_r)\sum\nolimits_s a_{js}(X_s - \mu_s)] = \sum\nolimits_{rs} a_{ir}a_{js}E[(X_r - \mu_r)(X_s - \mu_s)]$$

$$= \sum\nolimits_{rs} a_{ir}a_{js}\sigma_{rs} = \sum\nolimits_{rs} \mathbf{A}_{ir}\mathbf{\Sigma}_{rs}(\mathbf{A}^T)_{sj} = (\mathbf{A\Sigma A}^T)_{ij},$$

identifying the elements of the matrix product $\mathbf{A\Sigma A}^T$. //

Corollary. Covariance matrices $\mathbf{\Sigma}$ are non-negative definite.

Proof. Let $\mathbf{a}$ be any $n \times 1$ matrix (row-vector of length n); then $Y := \mathbf{aX}$ is a scalar. So $Y = Y^T = \mathbf{Xa}^T$. Taking $\mathbf{a} = \mathbf{A}^T, \mathbf{b} = \mathbf{0}$ above, Y has variance [$= 1 \times 1$ covariance matrix] $\mathbf{a}^T\mathbf{\Sigma a}$. But variances are non-negative. So $\mathbf{a}^T\mathbf{\Sigma a} \geq \mathbf{0}$ for all n-vectors $\mathbf{a}$. This says that $\mathbf{\Sigma}$ is non-negative definite. //