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

data = true value + error = fitted value + 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 because of its fundamental importance and of how well it illustrates the general case, also as it illustrates the ‘concrete’ way to do conditioning, which seems at first sight ‘abstract’ when done the Kolmogorov way (F22, SP) via σ-fields.

Motivating examples:
1. *CAPM* (I.6, W2). 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*, $N(\mu, \sigma^2)$, with mean μ, variance σ^2 and density

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

b. *Linear regression by the method of least squares* (IV.1). 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:

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

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

$$c = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, \quad Q = \frac{1}{1-\rho^2}\Big[\Big(\frac{x-\mu_1}{\sigma \quad {}_1}\Big)^2 - 2\rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big(\frac{y-\mu_2}{\sigma_2}\Big) + \Big(\frac{y-\mu_2}{\sigma_2}\Big)^2\Big].$$

Here $\sigma_i > 0$, μ_i are real, $-1 < \rho < 1$.

A full treatment of this basic and vitally important case is given in M5F22 Problems/Solutions 4. Recall that the crux is *completing the square*:

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

where

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

This leads as before to the ten key facts:

Fact 1. $f(x,y)$ is a joint density function (two-dimensional). Its marginal density functions $f_1(x), f_2(y)$ (one-dimensional) are given by

Fact 2. X, Y are normal: $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$. So:

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

Fact 4. Fact 5. The conditional mean $E(Y|X = x)$ is *linear* in x:

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

Fact 6. The conditional variance of Y given $X = x$ is

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

Fact 7. The correlation coefficient of X, Y is ρ.

Fact 8. The bivariate normal law has *elliptical contours.* For, the contours are $Q(x,y) = const$, which are ellipses (as Galton found).

Fact 9. The joint MGF and joint CF of X, Y are

$$M_{X,Y}(t_1,t_2) = M(t_1,t_2) = \exp(\mu_1 t_1 + \mu_2 t_2 + \frac{1}{2}[\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2]),$$

$$\phi_{X,Y}(t_1,t_2) = \phi(t_1,t_2) = \exp(i\mu_1 t_1 + i\mu_2 t_2 - \frac{1}{2}[\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2 t_2^2]).$$

Fact 10. X, Y are independent if and only if $\rho = 0$.

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. //

We turn now to a technical result, which is important in reducing n-dimensional problems to one-dimensional ones.

Theorem (Cramér-Wold device). The distribution of a random n-vector $\mathbf{X}$ is completely determined by the set of all one-dimensional distributions of linear combinations $\mathbf{t}^T\mathbf{X} = \sum_i t_i X_i$, where $\mathbf{t}$ ranges over all fixed n-vectors.

Proof. $Y := \mathbf{t}^T\mathbf{X}$ has CF

$$\phi_Y(t) := E\exp\{itY\} = E\exp\{it\mathbf{t}^T\mathbf{X}\}.$$

If we know the distribution of each Y, we know its CF $\phi_Y(t)$. In particular, taking $t = 1$, we know $E\exp\{i\mathbf{t}^T\mathbf{X}\}$. But this is the CF of $\mathbf{X} = (X_1, \cdots, X_n)^T$ evaluated at $\mathbf{t} = (t_1, \cdots, t_n)^T$. But this determines the distribution of $\mathbf{X}$. //

Thus by the Cramér-Wold device, to define an n-dimensional distribution it suffices to define the distributions of *all linear combinations*.

The Cramér-Wold device suggests a way to *define* the multivariate normal distribution. The definition below seems indirect, but it has the advantage of handling the full-rank and singular cases together ($\rho = \pm 1$ as well as $-1 < \rho < 1$ for the bivariate case).

Definition. An n-vector $\mathbf{X}$ has an *n-variate normal* distribution iff $\mathbf{a}^T\mathbf{X}$ has a univariate normal distribution for all constant n-vectors $\mathbf{a}$.

Proposition. (i) Any linear transformation of a multinormal n-vector is multinormal,
(ii) Any vector of elements from a multinormal n-vector is multinormal. In particular, the components are univariate normal.

Proof. (i) If $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{c}$ ($\mathbf{A}$ an $m \times n$ matrix, $\mathbf{c}$ an m-vector) is an m-vector, and $\mathbf{b}$ is any m-vector,

$$\mathbf{b}^T\mathbf{Y} = \mathbf{b}^T(\mathbf{A}\mathbf{X} + \mathbf{c}) = (\mathbf{b}^T\mathbf{A})\mathbf{X} + \mathbf{b}^T\mathbf{c}.$$

If $\mathbf{a} = \mathbf{A}^T\mathbf{b}$ (an m-vector), $\mathbf{a}^T\mathbf{X} = \mathbf{b}^T\mathbf{A}\mathbf{X}$ is univariate normal as $\mathbf{X}$ is multinormal. Adding the constant $\mathbf{b}^T\mathbf{c}$, $\mathbf{b}^T\mathbf{Y}$ is univariate normal. This holds for all $\mathbf{b}$, so $\mathbf{Y}$ is m-variate normal.

(ii) Take a suitable matrix $\mathbf{A}$ of 1s and 0s to pick out the required sub-vector.

Theorem 1. If $\mathbf{X}$ is n-variate normal with mean μ and covariance matrix $\mathbf{\Sigma}$, its CF is

$$\phi(\mathbf{t}) := E\exp\{i\mathbf{t}^T\mathbf{X}\} = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}.$$

Proof. By Proposition 1, $Y := \mathbf{t}^T\mathbf{X}$ has mean $\mathbf{t}^T\mu$ and variance $\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}$. By definition of multinormality, $Y = \mathbf{t}^T\mathbf{X}$ is univariate normal. So Y is $N(\mathbf{t}^T\mu, \mathbf{t}^T\mathbf{\Sigma}\mathbf{t})$, so Y has CF

$$\phi_Y(t) := E\exp\{itY\} = E\exp\{it\mathbf{t}^T\mathbf{X}\} = \exp\{it\mathbf{t}^T\mu - \frac{1}{2}t^2\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}.$$

Taking $t = 1$ (as in the proof of the Cramér-Wold device),

$$E\exp\{i\mathbf{t}^T\mathbf{X}\} = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}. \qquad //$$

Corollary. The components of $\mathbf{X}$ are independent iff $\mathbf{\Sigma}$ is diagonal.

Proof. The components are independent iff the joint CF factors into the product of the marginal CFs. This factorization takes place, into $\Pi_j \exp\{i\mu_j t_j - \frac{1}{2}\sigma_{jj}t_j^2\}$, in the diagonal case only. //

Corollary. Two Gaussian random variables X_i, X_j are independent iff they are uncorrelated, i.e. their correlation coefficient $\sigma_{ij} = 0$ is zero.

Proof. Taking all the t_k for $k \neq i, j$ reduces the joint CF above to a bivariate CF, which factorises as above (or as in Fact 10 of the bivariate normal distribution, IV.2 above) iff the cross-terms in $t_i t_j$ are absent, i.e. iff $\sigma_{ij} = 0$. //

So for Gaussians, *uncorrelated implies independent* (when the means exist). *Independent always implies uncorrelated*, by the Multiplication Theorem: for independence,

$$E[(X_i - EX_i)(X_j - EX_j)] = E[X_i - EX_i].E[X_j - EX_j] = 0.0 = 0.$$

So for Gaussians, *uncorrelated is equivalent to independent.* This useful property is wildly false in general! E.g.,

$$X := \cos 2\pi U, \quad Y := \sin 2\pi U, \qquad U \sim U(0,1)$$

are both uncorrelated (check), but heavily dependent (each determines U to within two values, so determined the other to within two values).

Recall that a covariance matrix $\mathbf{\Sigma}$ is always
(a) symmetric ($\sigma_{ij} = \sigma_{ji}$, as $\sigma_{ij} = cov(X_i, X_j)$),
(b) non-negative definite, written $\mathbf{\Sigma} \geq 0$: $\mathbf{a}^T \mathbf{\Sigma} \mathbf{a} \geq 0$ for all n-vectors $\mathbf{a}$.
Suppose that $\mathbf{\Sigma}$ is, further, *positive definite*, written $\mathbf{\Sigma} > 0$:

$$\mathbf{a}^T \mathbf{\Sigma} \mathbf{a} > 0 \qquad \text{unless} \qquad \mathbf{a} = \mathbf{0}.$$

The Multinormal Density.

If $\mathbf{X}$ is n-variate normal, $N(\mu, \mathbf{\Sigma})$, its density (in n dimensions) need not exist (e.g. the singular case $\rho = \pm 1$ with $n = 2$). But if $\mathbf{\Sigma} > \mathbf{0}$ (so $\mathbf{\Sigma}^{-1}$ exists), $\mathbf{X}$ has a density. The link between the multinormal density below and the multinormal MGF above is due to the English statistician F. Y. Edgeworth (1845-1926) in 1893.

Theorem (Edgeworth). If μ is an n-vector, $\mathbf{\Sigma} > \mathbf{0}$ a symmetric positive definite $n \times n$ matrix, then
(i)

$$f(\mathbf{x}) := \frac{1}{(2\pi)^{\frac{1}{2}n} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)\}$$

is an n-dimensional probability density function (of a random n-vector $\mathbf{X}$, say),
(ii) $\mathbf{X}$ has CF $\phi(\mathbf{t}) = \exp\{i\mathbf{t}^T \mu - \frac{1}{2}\mathbf{t}^T \mathbf{\Sigma} \mathbf{t}\}$,
(iii) $\mathbf{X}$ is multinormal $N(\mu, \mathbf{\Sigma})$.

Proof. Write $\mathbf{Y} := \mathbf{\Sigma}^{-\frac{1}{2}} \mathbf{X}$ ($\mathbf{\Sigma}^{-\frac{1}{2}}$ exists as $\mathbf{\Sigma} > \mathbf{0}$, by above). Then $\mathbf{Y}$ has covariance matrix $\mathbf{\Sigma}^{-\frac{1}{2}} \mathbf{\Sigma} (\mathbf{\Sigma}^{-\frac{1}{2}})^T$. Since $\mathbf{\Sigma} = \mathbf{\Sigma}^T$ and $\mathbf{\Sigma} = \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{\Sigma}^{\frac{1}{2}}$, $\mathbf{Y}$ has covariance matrix $\mathbf{I}$ (the components Y_i of $\mathbf{Y}$ are uncorrelated).

Change variables as above, with $\mathbf{y} = \mathbf{\Sigma}^{-\frac{1}{2}} \mathbf{x}$, $\mathbf{x} = \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{y}$. The Jacobian is (taking $\mathbf{A} = \mathbf{\Sigma}^{-\frac{1}{2}}$) $J = \partial \mathbf{x} / \partial \mathbf{y} = det(\mathbf{\Sigma}^{\frac{1}{2}}), = (det \mathbf{\Sigma})^{\frac{1}{2}}$ by the product theorem for determinants. Substituting, the integrand is

$$\exp\{-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)\} = \exp\{-\frac{1}{2}(\mathbf{\Sigma}^{\frac{1}{2}} \mathbf{y} - \mathbf{\Sigma}^{\frac{1}{2}} (\mathbf{\Sigma}^{-\frac{1}{2}} \mu))^T \mathbf{\Sigma}^{-1} (\mathbf{\Sigma}^{\frac{1}{2}} \mathbf{y} - \mathbf{\Sigma}^{\frac{1}{2}} (\mathbf{\Sigma}^{-\frac{1}{2}} \mu))\}.$$

Writing $\nu := \mathbf{\Sigma}^{-\frac{1}{2}} \mu$, this is

$$\exp\{-\frac{1}{2}(\mathbf{y} - \nu)^T \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{\Sigma}^{-1} \mathbf{\Sigma}^{\frac{1}{2}} (\mathbf{y} - \nu)\} = \exp\{-\frac{1}{2}(\mathbf{y} - \nu)^T (\mathbf{y} - \nu)\}.$$

So by the change-of-density formula, $\mathbf{Y}$ has density

$$g(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\mathbf{\Sigma}|^{\frac{1}{2}}}.|\mathbf{\Sigma}|^{\frac{1}{2}}.\exp\{-\frac{1}{2}(\mathbf{y}-\nu)^T(\mathbf{y}-\nu)\}.$$

This factorises as

$$\Pi_{i=1}^n \frac{1}{(2\pi)^{\frac{1}{2}}}\exp\{-\frac{1}{2}(y_i-\nu_i)^2\}.$$

So the components Y_i of $\mathbf{Y}$ are independent $N(\nu_i, 1)$. So $\mathbf{Y}$ is multinormal, $N(\nu, I)$.
(i) Taking $A = B = \mathbb{R}^n$, $\int_{\mathbb{R}^n} f(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^n} g(\mathbf{y})d\mathbf{y}, = 1$ as g is a probability density, as above. So f is also a probability density (non-negative and integrates to 1).
(ii) $\mathbf{X} = \Sigma^{\frac{1}{2}}\mathbf{Y}$ is a linear transformation of $\mathbf{Y}$, and $\mathbf{Y}$ is multivariate normal, $N(\nu, I)$. So $\mathbf{X}$ is multivariate normal.
(iii) $E\mathbf{X} = \mathbf{\Sigma}^{\frac{1}{2}}E\mathbf{Y} = \mathbf{\Sigma}^{\frac{1}{2}}\nu = \mathbf{\Sigma}^{\frac{1}{2}}.\mathbf{\Sigma}^{-\frac{1}{2}}\mu = \mu$, $cov\mathbf{X} = \mathbf{\Sigma}^{\frac{1}{2}}cov\mathbf{Y}(\mathbf{\Sigma}^{\frac{1}{2}})^T = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{I}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Sigma}$. So $\mathbf{X}$ is multinormal $N(\mu, \mathbf{\Sigma})$. So its CF is

$$\phi(\mathbf{t}) = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}. \qquad //$$

Note. The inverse $\mathbf{\Sigma}^{-1}$ of the covariance matrix $\mathbf{\Sigma}$ is called the *concentration matrix*, K.

Conditional independence of two components X_i, X_j of a multinormal vector given the others can be identified by vanishing of the (off-diagonal) (i, j) entry k_{ij} in the concentration matrix K. The proof needs the results on conditioning and regression in IV.6 W4 below, and the formula for the inverse of a partitioned matrix; see SMF1415 Problems 6.

Independence of Linear Forms

Given a normally distributed random vector $\mathbf{x} \sim N(\mu, \Sigma)$ and a matrix A, one may form the *linear form* $A\mathbf{x}$. One often encounters several of these together, and needs their joint distribution – in particular, to know when these are independent.

Theorem 3. Linear forms $A\mathbf{x}$ and $B\mathbf{x}$ with $\mathbf{x} \sim N(\mu, \Sigma)$ are independent iff

$$A\Sigma B^T = 0.$$

In particular, if A, B are symmetric and $\Sigma = \sigma^2 I$, they are independent iff

$$AB = 0.$$

Proof. The joint CF is

$$\phi(\mathbf{u}, \mathbf{v}) := E \exp\{i\mathbf{u}^T A\mathbf{x} + i\mathbf{v}^T B\mathbf{x}\} = E \exp\{i(A^T\mathbf{u} + B^T\mathbf{v})^T\mathbf{x}\}.$$

This is the CF of $\mathbf{x}$ at argument $\mathbf{t} = A^T\mathbf{u} + B^T\mathbf{v}$, so

$$\phi(\mathbf{u}, \mathbf{v}) = \exp\{i(\mathbf{u}^T A + \mathbf{v}^T B)\mu - \frac{1}{2}(A^T\mathbf{u} + B^T\mathbf{v})^T\Sigma(A^T\mathbf{u} + B^T\mathbf{v})\}$$

$$= \exp\{i(\mathbf{u}^T A + \mathbf{v}^T B)\mu - \frac{1}{2}[\mathbf{u}^T A\Sigma A^T\mathbf{u} + \mathbf{u}^T A\Sigma B^T\mathbf{v} + \mathbf{v}^T B\Sigma A^T\mathbf{u} + \mathbf{v}^T B\Sigma B^T\mathbf{v}]\}.$$

This factorises into a product of a function of $\mathbf{u}$ and a function of $\mathbf{v}$ iff the two cross-terms in $\mathbf{u}$ and $\mathbf{v}$ vanish, that is, iff $A\Sigma B^T = 0$ and $B\Sigma A^T = 0$; by symmetry of Σ, the two are equivalent.

4. Quadratic forms in normal variates

We give a brief treatment of this important material; for full detail see e.g. [BF], 3.4 – 3.6. Recall (IV.3)
(i) with $x \sim N(\mu, \Sigma)$, linear forms Ax, BX are independent iff $A\Sigma B^T = 0$;
(ii) for a projection, $P^2 = P$ (P is *idempotent*); for a symmetric projection, $P^T P = P$.
We restrict attention, for simplicity, to $\mu = 0$, $\Sigma = \sigma^2 I$, $x \sim N(0, \sigma^2 I)$.

It turns out that the distribution theory relevant to regression depends on *quadratic forms in normal variates*, $x^T Ax$ for a normally distributed random vector x, and that we can confine attention to projection matrices. For P a symmetric projection,

$$x^T Px = x^T P^T Px = (Px)^T(Px),$$

which reduces from *quadratic forms* to *linear* forms – which are much easier! So: if xP_1x, xP_2x are quadratic forms in normal vectors x, with P_1, P_2 projections, $x^T P_1 x$ and $x^T P_2 x$ are independent iff

$$P_1 P_2 = 0 :$$

P_1, P_2 are *orthogonal projections.* Recall that projections P_1, P_2 are *orthogonal* if their ranges are orthogonal subspaces, i.e.

$$(P_1x).(P_2x) = 0 \quad \forall\ x : \quad x^T P_1^T P_2 x = 0 \quad \forall x; \quad P_1^T P_2 = 0 \quad \forall x; \quad P_1 P_2 = 0$$

for P_i symmetric. Note that for P a projection, $I - P$ is a projection orthogonal to it:

$$(I-P)^2 = I-2P+P^2 = 1-2P+P = I-P; \quad P(I-P) = P-P^2 = P-P = 0.$$

If λ is an eigenvalue of A, λ^2 is an eigenvalue of A^2 (check). So if a projection P has eigenvalue λ, $\lambda^2 = \lambda$: $\lambda = 0$ or 1. Also, the trace is the sum of the eigenvalues; for a projection, this is the number of non-zero eigenvalues; this is the rank. So:

For a projection, the eigenvalues are 0 or 1, and the trace is the rank.

By Spectral Decomposition (III.1), a symmetric projection matrix P can be diagonalised by an orthogonal transformation O to a diagonal matrix D:

$$O^T P O = D, \qquad P = ODO^T;$$

as above, the diagonal entries d_{ii} are 0 or 1, and we may re-order so that the 1s come first. So with $y := O^T x$,

$$x^T P x = x^T ODO^T x = y^T D y = y_1^2 + \ldots + y_r^2.$$

Normality is preserved under orthogonal transformations (check!), so also $y \sim N(0, \sigma^2 I)$. So $y_1^2 + \ldots + y_r^2$ is σ^2 times the sum of r independent squares of standard normal variates, and this sum is $\chi^2(r)$ (by definition of chi-square):

$$x^T P x \sim \sigma^2 \chi^2(r).$$

If P has rank r, $I - P$ has rank $n - r$ (where n is the sample size – the dimension of the vector space we are working in):

$$x^T (I - P) x \sim \sigma^2 \chi^2(n - r),$$

and the two quadratic forms are independent.

It turns out that all this can be generalised, to the sum of several projections, not just two. This result – the key to all the distribution theory in Regression – is *Cochran's theorem* (William G. COCHRAN (1909-1980) in 1934); [BF] Th. 3.27):

Theorem (Cochran's Theorem). If

$$I = P_1 + \ldots + P_k$$

with each P_i a symmetric projection with rank n_i, then
(i) the ranks sum: $n = n_1 + \ldots + n_k$;
(ii) each quadratic form $Q_i := x^T P_i x \sim \sigma^2 \chi^2(n_i)$;
(iii) $Q_1, \ldots, Q_k$ are mutually independent;
(iv) $P_1, \ldots, P_k$ are mutually orthogonal: $P_i P_j = 0$ for $i \neq j$.

The quadratic forms that we encounter in Statistics are called *sums of squares (SS)* – for *regression* (SSR), for *error* (SSE), for the *hypothesis* (SSH), etc.

Recall the definition of the *Fisher F-distribution* with degrees of freedom (df) m and n (note the order): $F(m,n)$ is the distribution of the ratio

$$F := \frac{U/m}{V/n},$$

with U, V independent chi-square random variables with df m, n (see e.g. [BF] 2.3 for the explicit formula for the density, but we shall not need this).

Recall also (or, if you have not met these, take a look at a textbook):
(i) *Analysis of variance (ANOVA)* (see e.g. [BF] Ch. 2). Here one tests for differences between the *means* of different (normal) populations by analysing *variances*. Specifically, one looks at *within-groups* variability and *between-groups* variability, and *rejects* the null hypothesis of no difference between the group means if the second is *too big* compared to the first. As above, one forms the relevant F-statistic, and rejects if F is too big. Here one has *qualitative* factors (which group?).
(ii) *Analysis of Covariance (ANCOVA)* (see e.g. [BF] Ch. 5. Similarly for ANCOVA, where one has both qualitative factors (as with ANOVA) and quantitative ones (covariates), as with Regression.
(iii) Tests of linear hypotheses in Regression (II.4; see e.g. [BF] Ch. 6). Here we reject if SSH is too big compared to SSE.

5. Estimation theory for the multivariate normal.

Given a sample $x_1, \ldots, x_n$ from the multivariate normal $N_p(\mu, \Sigma)$, form the *sample mean* (vector) and the *sample covariance matrix* as in the one-dimensional case:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad S := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}).$$

The likelihood for a sample of size 1 is

$$L(x|\mu,\Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\},$$

so the likelihood for a sample of size n is

$$L = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\{-\frac{1}{2}\sum_1^n(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\}.$$

Writing

$$x_i - \mu = (x_i - \bar{x}) - (\mu - \bar{x}),$$

$$\sum_1^n(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = \sum_1^n(x_i-\bar{x})^T\Sigma^{-1}(x_i-\bar{x}) + n(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)$$

(the cross-terms cancel as $\sum_1^n(x_i-\bar{x}) = 0$). The summand in the first term on the right is a scalar, so is its own trace. Since $trace(AB) = trace(BA)$ and $trace(A+B) = trace(B+A)$,

$$trace(\sum_1^n(x_i-\bar{x})^T\Sigma^{-1}(x_i-\bar{x})) = trace(\Sigma^{-1}\sum_1^n(x_i-\bar{x})^T(x_i-\bar{x}))$$

$$= trace(\Sigma^{-1}.nS) = n\ trace(\Sigma^{-1}S).$$

Combining,

$$L = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\{-\frac{1}{2}n\ trace(\Sigma^{-1}S) - \frac{1}{2}n(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\}.$$

Write ('K for Konzentration' – or, 'V for variance')

$$K := \Sigma^{-1}:$$

$$\ell = const - \frac{1}{2}n\ trace(KS) - (\bar{x}-\mu)^T K(\bar{x}-\mu).$$

So by the Fisher-Neyman Theorem, $(\bar{x}, S)$ is sufficient for (μ, Σ) (equivalently, for (μ, K)). It is in fact minimal sufficient (SMF1415 Problems 2 Q2).

These natural estimators are in fact the MLEs:

Theorem. For the multivariate normal $N_p(\mu,\Sigma)$, $\bar{x}$ and S are the maximum likelihood estimators for μ, Σ.

Proof. As above, with $K = (k_{ij}) := \Sigma^{-1}$, the likelihood is

$$L = const.|K|^{n/2} \exp\{-\frac{1}{2}n \; trace(KS) - \frac{1}{2}n(\bar{x}-\mu)^T K(\bar{x}-\mu)\},$$

so the log-likelihood is

$$\ell = c + \frac{1}{2}n \log|K| - \frac{1}{2}n \; trace(KS) - \frac{1}{2}n(\bar{x}-\mu)^T K(\bar{x}-\mu).$$

The MLE $\hat{\mu}$ for μ is $\bar{x}$, as this reduces the last term (the only one involving μ) to its minimum value, 0. For $A = (a_{ij})$, its determinant is

$$|A| = \sum_j a_{ij} A_{ij}$$

for each i, or

$$|A| = \sum_i a_{ij} A_{ij}$$

for each j, expanding by the ith row or jth column, where A_{ij} is the *cofactor* (signed minor) of a_{ij}. From either,

$$\partial|A|/\partial a_{ij} = A_{ij},$$

so

$$\partial \log|A|/\partial a_{ij} = A_{ij}/|A| = (A^{-1})_{ji},$$

the (j, i) element of A^{-1}, recalling the formula for the matrix inverse (or $(A^{-1})_{ij}$ if A is symmetric). Also, if B is symmetric,

$$trace(AB) = \sum_i \sum_j a_{ij} b_{ji} = \sum_{i,j} a_{ij} b_{ij} :$$

$$\partial \; trace(AB)/\partial a_{ij} = b_{ij}.$$

Using these, and writing $S = (s_{ij})$,

$$\partial \log|K|/\partial k_{ij} = (K^{-1})_{ij} = (\Sigma)_{ij} = \sigma_{ij} \qquad (K := \Sigma^{-1}),$$

$$\partial \; trace(KS)/\partial k_{ij} = s_{ij}.$$

So

$$\partial\ell/\partial v_{ij} = \frac{1}{2}n(\sigma_{ij} - s_{ij}),$$

which is 0 for all i and j iff $\Sigma = S$. This says that S is the MLE for Σ. //

6. Conditioning and regression

In general, we should always *use what we know.* In Probability and Statistics, this goes by the technical term of *conditioning.* This rests ultimately on the formula $P(A|B) := P(A \cap B)/P(B)$ of elementary probability (applicable only when $P(B) > 0$!), and its analogue with sums replaced by integrals when densities exist (which they do not in general!). Both these elementary cases are handled above in our treatment of the bivariate normal distribution (IV.2). The general approach to conditioning is due to Kolmogorov in 1933, and uses Measure Theory and σ-fields; see e.g. [SP]. We pause to make the link between conditioning and regression.

Recall that the *conditional* density of Y *given* $X = x$ is

$$f_{Y|X}(y|x) := f_{X,Y}(x,y) / \int f_{X,Y}(x,y)dy.$$

Conditional means.

The conditional mean of Y given $X = x$ is

$$E(Y|X = x),$$

a function of x called the *regression* function (of Y on x). So, if we do not specify the value x, we get $E(Y|X)$. This is *random*, because X is random (until we observe its value, x; then we get the regression function of x as above). As $E(Y|X)$ is random, we can look at its mean and variance.

Recall (SP, Ch. II)

Theorem (Conditional Mean Formula). $E[E(Y|X)] = EY$.

Interpretation. EY takes the random variable Y, and averages out all the randomness to give a number, EY.
$E(Y|X)$ takes the random variable Y, and averages out all the randomness in Y NOT accounted for by knowledge of X.
$E[E(Y|X)]$ then averages out the remaining randomness, which IS accounted for by knowledge of X, to give EY as above.
Example: Bivariate normal distribution, $N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$, or $N(\mu, \sigma)$,

$$\mu = (\mu_1, \mu_2)^T, \qquad \sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Then

$$E(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \qquad \text{so} \qquad E(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1):$$

$$E[E(Y|X)] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(EX - \mu_1) = \mu_2 = EY, \qquad \text{as} \qquad EX = \mu_1.$$

As with the bivariate normal, we should keep some concrete instance in mind as a motivating example, e.g.:
X = incoming score of student [in medical school or university, say], Y = graduating score;
X = child's height at 2 years (say), Y = child's eventual adult height,
or X = mid-parent height, Y = child's adult height, as in Galton's study.

Recall also (SP, Ch. II)

Theorem (Conditional Variance Formula).

$$varY = E[var(Y|X)] + var(E[Y|X]).$$

Interpretation.

$$varY = \text{total variability in } Y,$$

$$E_X var(Y|X) = \text{variability in } Y \text{ not accounted for by knowledge of } X,$$

$$var_X E(Y|X) = \text{variability in } Y \text{ accounted for by knowledge of } X.$$

Example: the bivariate normal.

$$Y|X = x \text{ is } N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1-\rho^2)), \qquad varY = \sigma_2^2,$$

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

which has variance $(\rho\sigma_2/\sigma_1)^2 varX = (\rho\sigma_2/\sigma_1)^2\sigma_1^2 = \rho^2\sigma_2^2$;

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

Corollary. $E(Y|X)$ has the same mean as Y and smaller variance (if anything) than Y.

Proof. From the Conditional Mean Formula, $E[E(Y|X)] = EY$. Since $var(Y|X) \geq 0$, $E_X var(Y|X) \geq 0$, so

$$varE[Y|X] \leq varY$$

from the Conditional Variance Formula. //

This result has important applications in estimation theory. Suppose we are to estimate a parameter θ, and are considering a statistic X as a possible estimator (or basis for an estimator) of θ. We would naturally want X to contain all the information on θ contained within the entire sample. What (if anything) does this mean in precise terms? The answer lies in the concept of *sufficiency* ('data reduction' – I.4 W2) – one of the most important contributions to statistics of the great English statistician R. A. (Sir Ronald) Fisher (1880-1962) in 1920. In the language of sufficiency, the Conditional Variance Formula is seen as (essentially) the *Rao-Blackwell Theorem*, a key result in the area (see the index in your favourite Statistics book for more).

Regression.

In the bivariate normal, with X = mid-parent height, Y = child's height, $E(Y|X = x)$ is linear in x (*regression line*). In a more detailed analysis, with U = father's height, V = mother's height, Y = child's height, one would expect $E(Y|U = u, V = v)$ to be linear in u and v (*regression plane*), etc.

In an n-variate normal distribution $N_n(\mu, \Sigma)$ (we restrict attention to Σ non-singlular for simplicity), suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is partitioned into $\mathbf{X}_1 := (X_1, \cdots, X_r)^T$ and $\mathbf{X}_2 := (X_{r+1}, \cdots, X_n)^T$. Let the corresponding partition of the mean vector and the covariance matrix be

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $E\mathbf{X}_i = \mu_i$, Σ_{11} is the covariance matrix of $\mathbf{X}_1$, Σ_{22} that of $\mathbf{X}_2$, $\Sigma_{12} = \Sigma_{21}^T$ the covariance matrix of $\mathbf{X}_1$ with $\mathbf{X}_2$.

The concentration matrix (= precision matrix), $K := \Sigma^{-1}$.

By Edgeworth's Theorem, the (multinormal) distribution $N(\mu, \Sigma)$ of a Gaussian n-vector is determined by the mean μ and the covariance matrix Σ. Now the matrix entries $\sigma_{ij} := cov(X_i, X_j)$ in Σ are determined by the behaviour of the coordinates of $\mathbf{X}$ *two at a time*, while those of the concentration matrix $K := \Sigma^{-1}$ depend on the whole distribution (the X_i *all together*). This makes K a better choice than Σ for some purposes, as it captures structural information better (see below). Write as above

$$K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix},$$

To proceed, we need the formula for the *inverse of a partitioned matrix.* You can check that, when all inverses exist, this is given by

$$\left(\begin{array}{cc} A & B \\ C & D \end{array} \right)^{-1} = \left(\begin{array}{cc} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{array} \right), \quad M := (A - BD^{-1}C)^{-1},$$

Lemma. If Σ is positive definite, so is Σ_{11}.

Proof. $\mathbf{x}^T \Sigma \mathbf{x} > \mathbf{0}$ for all $\mathbf{x} \neq 0$ as Σ is positive definite. Take $\mathbf{x} = (\mathbf{x}_1, \mathbf{0})^T$, where $\mathbf{x}_1$ has the same number of components as the order of Σ_{11} [i.e., in matrix language, so that the partition of $\mathbf{x}$ is conformable with those of μ and σ above]. Then $\mathbf{x}_1 \Sigma_{11} \mathbf{x}_1 > 0$ for all $\mathbf{x}_1 \neq 0$. This says that Σ_{11} is positive definite. //

One of the most important things about regression is the link between *linearity* and *Gaussianity*: the conditional mean is *linear* in what one is conditioning on:

Theorem. The regression of $\mathbf{X}_2$ on $\mathbf{X}_1$ is linear:

$$E(\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1) = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{x}_1 - \mu_1) = \mu_2 - K_{11}^{-1} K_{12} (\mathbf{x}_1 - \mu_1).$$

This is a special case of the next result (Exam1516, Q6). We make two small changes: (a) interchange 1 and 2 (to reflect the shift of interest from what is conditioned on, as here, to what is left, as there): (b) drop bold-face in the notation (use the lightest notation that will do the job, and let context speak for itself).

Theorem (Gaussian Regression Formula). If a multinormal vector x is partitioned into x_1 and x_2, with μ, Σ, K partitioned accordingly, the conditional distribution of x_1 given x_2 in terms of μ, K is

$$x_1 | x_2 \sim N(\mu_1 - K_{11}^{-1} K_{12} (x_2 - \mu_2), K_{11}^{-1}),$$

or in terms of μ and Σ,

$$x_1 | x_2 \sim N(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}).$$

Corollary. The regression of x_1 on x_2 is linear:

$$E[x_1 | x_2] = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) = \mu_1 - K_{11}^{-1} K_{12} (x_2 - \mu_2).$$

First Proof (densities). By Edgeworth's theorem, if $x \sim N(\mu, \Sigma)$, $K := \Sigma^{-1}$,

$$f(x) \propto \exp\{-\frac{1}{2}(x-\mu)^T K (x-\mu)\}.$$

$$f(x_1, x_2) \propto \exp\{-\frac{1}{2}(x_1^T - \mu_1^T, x_2^T - \mu_2^T) \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\},$$

giving (as a scalar is its own transpose, so the two cross-terms are the same)

$$\exp\{-\frac{1}{2}[(x_1^T - \mu_1^T)K_{11}(x_1 - \mu_1) + 2(x_1^T - \mu_1^T)K_{12}(x_2 - \mu_2) + (x_2^T - \mu_2^T)K_{22}(x_2 - \mu_2)]\}.$$

So

$$f_{1|2}(x_1|x_2) = f(x_1, x_2)/f_2(x_2)$$

$$\propto \exp\{-\frac{1}{2}[(x_1^T - \mu_1^T)K_{11}(x_1 - \mu_1) + 2(x_1^T - \mu_1^T)K_{12}(x_2 - \mu_2)]\}, \qquad (*)$$

treating x_2 here as a constant and x_1 as the argument of $f_{1|2}$. By Edgeworth's theorem again, if the conditional mean of $x_1|x_2$ is ν_1,

$$f_{1|2}(x_1|x_2) \propto \exp\{-\frac{1}{2}(x_1^T - \nu_1^T)V_{11}(x_1 - \nu_1)\}, \qquad (**)$$

for some matrix V_{11}. So $x_1|x_2$ is multinormal. Equating coefficients of the quadratic term, the conditional concentration matrix of $x_1|x_2$ is $V_{11} = K_{11}$:

$$conc(x_1|x_2) = K_{11}.$$

So the conditional covariance matrix is K_{11}^{-1}. Then equating linear terms in $(*)$ and $(**)$ gives the conditional mean:

$$x_1^T K_{11} \nu_1 = x_1^T K_{11} \mu_1 - x_1^T K_{12}(x_2 - \mu_2) : \quad \nu_1 := E[x_1|x_2] = \mu_1 - K_{11}^{-1} K_{12}(x_2 - \mu_2) :$$

$$x_1|x_2 \sim N(\mu_1 - K_{11}^{-1}K_{12}(x_2 - \mu_2), K_{11}^{-1}).$$

Using the result above for the inverse of a partitioned matrix gives

$$M = K_{11}, \qquad M^{-1} = K_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

$$K_{11}^{-1}K_{12} = M^{-1}(-MBD^{-1}) = -BD^{-1} = -\Sigma_{12}\Sigma_{22}^{-1}.$$

Combining,

$$x_1|x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \qquad //$$

Second Proof (CFs). Recall that $\mathbf{AX}, \mathbf{BX}$ are independent iff $\mathbf{A\Sigma B}^T = \mathbf{0}$, or as Σ is symmetric, $\mathbf{B\Sigma A}^T = \mathbf{0}$. Now

$$\mathbf{X}_1 = \mathbf{AX} \text{ where } \mathbf{A} = (\mathbf{I}, \mathbf{0}),$$

$$\mathbf{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}_1 = \begin{pmatrix} -\Sigma_{21}\Sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \mathbf{BX}, \text{ where } \mathbf{B} = \begin{pmatrix} -\Sigma_{21}\Sigma_{11}^{-1} & \mathbf{I} \end{pmatrix}.$$

Now

$$\mathbf{B\Sigma A}^T = \begin{pmatrix} -\Sigma_{21}\Sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} -\Sigma_{21}\Sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \Sigma_{11} \\ \Sigma_{21} \end{pmatrix}$$

$$= -\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11} + \Sigma_{21} = \mathbf{0},$$

so $\mathbf{X}_1$ and $\mathbf{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}_1$ are *independent.* Since both are linear transformations of $\mathbf{X}$, which is multinormal, both are *multinormal.* Also,

$$E(\mathbf{BX}) = \mathbf{B}E\mathbf{X} = \begin{pmatrix} -\Sigma_{21}\Sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1.$$

To calculate the covariance matrix, introduce $\mathbf{C} := -\Sigma_{21}\Sigma_{11}^{-1}$, so $\mathbf{B} = (\mathbf{C}\ \mathbf{I})$, and recall $\Sigma_{12}^T = \Sigma_{21}$, so $\mathbf{C}^T = -\Sigma_{11}^{-1}\Sigma_{12}$:

$$var(\mathbf{BX}) = \mathbf{B\Sigma B}^T = \begin{pmatrix} \mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{C}^T \\ \mathbf{I} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \Sigma_{11}\mathbf{C}^T + \Sigma_{12} \\ \Sigma_{21}\mathbf{C}^T + \Sigma_{22} \end{pmatrix} = \mathbf{C}\Sigma_{11}\mathbf{C}^T + \mathbf{C}\Sigma_{12} + \Sigma_{21}\mathbf{C}^T + \Sigma_{22}$$

$$= \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11}\Sigma_{11}^{-1}\Sigma_{12} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22}$$

$$= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

By independence, the conditional distribution of $\mathbf{BX}$ given $\mathbf{X}_1 = \mathbf{AX}$ is the same as its marginal distribution, which by above is $N(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$. So given $\mathbf{X}_1$, $\mathbf{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}_1$ is $N(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$.

To pass from the conditional distribution of $\mathbf{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}_1$ given $\mathbf{X}_1$ to that of $\mathbf{X}_2$ given $\mathbf{X}_1$: just add $\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}_1$. Then

$$\mathbf{X}_2|\mathbf{X}_1 \sim N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}). \quad //$$

Here $\Sigma_{2|1} := \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is called the *partial covariance matrix* of $\mathbf{X}_2$ given $\mathbf{X}_1$.

Note. Here $\Sigma_{2|1} := \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is called the *partial covariance matrix* of $\mathbf{X}_2$ given $\mathbf{X}_1$.

The concentration matrix $K := \Sigma^{-1}$ proves its value in terms of *conditional independence*. For important result below, see e.g. Prop. 5.2 (§5.1.3) in
Steffen L. LAURITZEN, *Graphical models*, OUP, 1996.
Note the contrast with our earlier result: two components are *independent* iff the corresponding element σ_{ij} of Σ is zero.

Theorem. In a Gaussian n-vector, two components (the ith and jth, say) are conditionally independent given all the others iff the correspondng element k_{ij} in the concentration matrix K is zero.

Proof. By relabelling the indices, we can assume $i = 1$ and $j = 2$. Consider the 2-vector x_1 of the first two coordinates of x. Then, from the first proof above with $V_{11} = K_{11}$ in $(**)$, the conditional density of the first two coordinates given all the others is

$$f_{1|2}(x_1|x_2) \propto \exp\{-\frac{1}{2}(x_1^T - \nu_1^T)K_{11}(x_1 - \nu_1)\}.$$

The first two coordinates are conditionally independent given all the others iff this (bivariate) conditional density factorises into the product of the two marginals. This happens (as in the bivariate normal) iff the 2×2 matrix K_{11} is diagonal, i.e. iff its off-diagonal terms $k_{12} = k_{21}$ are zero. //

Recall (see e.g. Probability for Statistics (the PfS link on my homepage), Ch. V, Markov chains) the *Markov property*: for predicting the future, the present gives the same as the present plus the past. Equivalently: a process is Markov iff *past and future are conditionally independent given the present.* So when the process is *both Gaussian and Markov*, the last result shows that the concentration matrix will have many zeros, i.e. will be *sparse*, and the pattern of zeros will be informative: it will give the order of the components in time, to within time-reversal (as past and future are interchangeable here). For the resulting Gaussian Msrkov theory, see e.g.

H. RUE and L. HELD, *Gaussian Markov random fields: Theory and applications*, Chapman & Hall, 2005.

Note. 1. It often happens (e.g. in the Gaussian Markov case) that the concentration matrix K is *sparse* (most elements are zero), while its inverse the covariance matrix Σ is *dense* (most elements are non-zero). While inverting a matrix is straightforward theoretically, and not problematic for small matrices (unless they are *ill-conditioned* — close to being singular, and so numerically unstable), inverting *large matrices* numerically is a formidable task. So K is preferred to Σ in such cases.
2. The need to handle matrix operations numerically in an efficient way is the subject of Numerical Linear Algebra. This has grown greatly in importance recently, partly as computers have grown more powerful, partly in view of the prevalence of Big Data in applications.
3. We will meet conditional independence later in VII.7, Hierarchical models: Markov chain Monte Carlo.
4. The argument above (match the quadratic terms first, then match the linear terms) is the crux of a similar proof, involving *mixed models* in regression. Here, one has both *fixed effects* and *random effects*. The classic result is *Henderson's mixed-model equations* (widely used nowadays; it originated in studies of selective breeding in the US dairy farming industry, where it greatly improved efficiency). See e.g. [BF], §9.1.
5. Both the hierarchical models and the mixed models above occur in *Bayesian statistics*, for which see Ch. VII.

Elliptical models

The multinormal, or Gaussian, model is wonderfully convenient mathematically. In particular, the property of having linear regression is highly convenient. However, we note two properties of normal or Gaussian distributions, in any dimension:
(i) they are *symmetrical*, and so cannot model *skewness*;
(ii) they have *extremely thin tails* (so deviations of, say, 3 standard deviations from the mean are very rare).
But these contradict common observation in finance!
Skew.

Profit and loss are profoundly asymmetrical! Large unexpected profits are nice; large unexpected losses are lethal. Consequently, a given amount of profit gives less pleasure than a given amount of loss gives pain. One can

see the same effect in prices falling below a peak once the market has turned *far faster* than they increase when the market is rising (so one can detect the *arrow of time* from time series of price data).
Tails.

Inspection (EDA) of any financial data set will reveal *much fatter* tails than Gaussian. Typically, one sees *heavy tails* – tails that decay like a power (as with the Student t-distributions).

There is a third problem, that arises in portfolio management, where we have a range of assets (balanced, by Markowitzian diversification). The tails of two different components of a multinormal vector are (asymptotically) independent. By contrast, the negative tails (downside risk) of assets are usually highly dependent: in a falling market, everything falls, and the tails are heavily dependent.

For all these reasons, it is important to seek other models, which retain as many as possible of the desirable properties of the normal but not the disadvantages above. Such models exist – the *elliptical*, or *elliptically contoured*, models. These may be characterised in several ways. An elliptically contoured distribution in n dimensions with mean vector μ and covariance matrix Σ of rank k (with Cholesky decomposition $\Sigma = A^T A$) has a *stochastic representation*

$$X = \mu + RA^T u \qquad (R: \text{ risk-driver});$$

here u is a random vector uniformly distributed over the unit sphere in k dimensions and $R \geq 0$ is a scalar random variable independent of u. Alternatively, X has CF

$$\psi(t) = e^{it^T \mu} \phi(t^T \Sigma t)$$

for some scalar function ϕ. Thus $\phi(x) = e^{-\frac{1}{2}x}$ gives the Gaussian case, and choosing ϕ to decrease more slowly gives heavier tails, as required. For background, we refer to e.g. the book [MFE] and the paper [BFK].
Copulas.

Given a random n-vector $X = (X_1, \ldots, X_n)$, write $F(x) = F(x_1, \ldots, x_n) := P(X_{\leq} x_1, \ldots, X_n \leq x_n)$ for the *joint* distribution function, $F_i(x_i) := P(X_i \leq x_i)$ for the *marginal* distribution functions. Then by *Sklar's theorem* (Abe SKLAR (1915-) in 1958),

$$F(x) = C(F_1(x_1), \ldots, F_n(x_n))$$

for some distribution function $C(u) = C(u_1, \ldots, u_n)$ on the unit n-cube. This C is called the *copula*, as it *couples* the marginals together to give the joint distribution. It contains all the information on the *dependence* structure (vital for financial applications, as above!). For more, see e.g. [MFE] Ch. 5.

7. Generalised linear models (GLMs).

In Regression above, we took as our basic model

$$y = A\beta + \epsilon : \qquad Ey = A\beta; \qquad Ey_i = \sum_j a_{ij}\beta_j$$

– our data y (an n-vector) is modelled as a linear transformation (by a known matrix A, the *design matrix*, $n \times p$) of a p-vector β of parameters, plus an *error*. That is, we work with *linear combinations of predictors plus error*; in particular, the mean μ is given by a *linear predictor*, η. This simple procedure is surprisingly general and effective, but there are situations where it does not apply. We turn to these, seeking to use as much as possible of the approach above.

First, we generalise this by allowing the linear predictor η to be some (smooth and monotone, so invertible) function g of the mean μ:

$$\eta = g(\mu),$$

where g is called the *link function*, or *link*. Next, we need to specify the *error* structure. This is done by means of *exponential families* (see e.g. SMF1415 I.6.4, D2): the y_i are independent, with densities

$$f(y_i) = \exp\{\frac{\omega_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y, \phi)\};$$

here b, c are known functions, ω_i are known weights, ϕ is a scale parameter (known or unknown), and the parameter θ_i depends on η.

The case where this dependence is given by the identity,

$$\theta = \eta,$$

is particularly important; here the link is called *canonical.*

GLMs were introduced by Nelder and Wedderburn in 1972; our treatment here follows [BF] Ch. 8. The standard work is
[McN] P. McCULLAGH and J. A. NELDER, *Generalised linear models*, 2nd ed., 1989, Chapman and Hall (1st ed. 1983).
They have been extended to *hierarchical GLMs* (see Ch. VII):

[NLP] J. A. NELDER, Y. LEE and Y. PAWITAN, *Generalised linear models with random effects: unified analysis via H-likelihood*. Chapman and Hall, 2006.

Examples.

1. *Normal.* Here $g(\mu) = \mu$, the errors are normal, and the GLM reduces to the ordinary Linear Model above – as was to be expected!

2. *Poisson.* For the Poisson distribution $P(\lambda)$, writing y for the usual $k = 0, 1, 2, \ldots$,

$$f(y, k) = e^{-\lambda}\lambda^y/y! = \exp\{y \log\lambda - \lambda - \log y!\}.$$

So $\theta = \eta = \log\lambda$: the canonical link is the *logarithm*:

$$\eta = \log\lambda.$$

The Poisson law is the default option for *count data*. The log here explains the use of logs in *log-linear models* for count data – contingency tables, etc. (Pearson's chi-square goodness-of-fit test, 1900). See e.g. [BF] 8.3 – 8.5.

3. *Gamma.* The Gamma density $\Gamma(\lambda, \alpha)$ $(\lambda, \alpha > 0)$ has density $f(x) = \lambda^\alpha e^{-\lambda x} x^{\alpha-1}/\Gamma(\lambda)$ on $(0, \infty)$. The mean is $\mu = \alpha/\lambda$, and the canonical link is $\eta = 1/\mu$. The Gamma is the default option for error structure on $(0, \infty)$; here it is often used with the log-link $\eta = \log\mu$. See e.g. [BF] 8.2.3 for an application (to athletics times).