(iii) First, $f(.)$ is a density, as it is non-negative, and integrates to 1:

$$\begin{aligned} \int f(x)\,dx &= \frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)}\int_0^\infty x^{\frac{1}{2}n-1}\exp\left(-\frac{1}{2}x\right)\,dx \\ &= \frac{1}{\Gamma\left(\frac{1}{2}n\right)}\int_0^\infty u^{\frac{1}{2}n-1}\exp(-u)\,du \qquad (u := \frac{1}{2}x) \\ &= 1, \end{aligned}$$

by definition of the Gamma function. Its MGF is

$$\begin{aligned} M(t) &= \frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)}\int_0^\infty e^{tx}x^{\frac{1}{2}n-1}\exp\left(-\frac{1}{2}x\right)\,dx \\ &= \frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)}\int_0^\infty x^{\frac{1}{2}n-1}\exp\left(-\frac{1}{2}x(1-2t)\right)\,dx. \end{aligned}$$

Substitute $u := x(1-2t)$ in the integral. One obtains

$$M(t) = (1-2t)^{-\frac{1}{2}n}\frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)}\int_0^\infty u^{\frac{1}{2}n-1}e^{-u}\,du = (1-2t)^{-\frac{1}{2}n},$$

by definition of the Gamma function. □

Chi-square Addition Property. If X_1, X_2 are independent, $\chi^2(n_1)$ and $\chi^2(n_2)$, $X_1 + X_2$ is $\chi^2(n_1 + n_2)$.

Proof

$X_1 = U_1^2 + \ldots + U_{n_1}^2$, $X_2 = U_{n_1+1}^2 + \ldots + U_{n_1+n_2}^2$, with U_i iid $N(0,1)$.
So $X_1 + X_2 = U_1^2 + \cdots + U_{n_1+n_2}^2$, so $X_1 + X_2$ is $\chi^2(n_1 + n_2)$. □

Chi-Square Subtraction Property. If $X = X_1 + X_2$, with X_1 and X_2 independent, and $X \sim \chi^2(n_1 + n_2)$, $X_1 \sim \chi^2(n_1)$, then $X_2 \sim \chi^2(n_2)$.

Proof

As X is the independent sum of X_1 and X_2, its MGF is the product of their MGFs. But X, X_1 have MGFs $(1-2t)^{-\frac{1}{2}(n_1+n_2)}$, $(1-2t)^{-\frac{1}{2}n_1}$. Dividing, X_2 has MGF $(1-2t)^{-\frac{1}{2}n_2}$. So $X_2 \sim \chi^2(n_2)$. □

2.2 Change of variable formula and Jacobians

Recall from calculus of several variables the change of variable formula for multiple integrals. If in

$$I := \int \ldots \int_A f(x_1, \ldots, x_n)\ dx_1 \ldots dx_n = \int_A f(\mathbf{x})\ d\mathbf{x}$$

we make a one-to-one change of variables from $\mathbf{x}$ to $\mathbf{y}$ — $\mathbf{x} = \mathbf{x}(\mathbf{y})$ or $x_i = x_i(y_1, \ldots, y_n) \quad (i = 1, \ldots, n)$ — let B be the region in $\mathbf{y}$-space corresponding to the region A in $\mathbf{x}$-space. Then

$$I = \int_A f(\mathbf{x})\ d\mathbf{x} = \int_B f(\mathbf{x}(\mathbf{y})) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|\ d\mathbf{y} = \int_B f(\mathbf{x}(\mathbf{y}))|J|\ d\mathbf{y},$$

where J, the determinant of partial derivatives

$$J := \frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \frac{\partial(x_1, \cdots, x_n)}{\partial(y_1, \cdots, y_n)} := \det\left(\frac{\partial x_i}{\partial y_j}\right)$$

is the *Jacobian* of the transformation (after the great German mathematician C. G. J. Jacobi (1804–1851) in 1841 – see e.g. Dineen (2001), Ch. 14). Note that in one dimension, this just reduces to the usual rule for change of variables: $dx = (dx/dy).dy$. Also, if J is the Jacobian of the change of variables $\mathbf{x} \to \mathbf{y}$ above, the Jacobian $\partial\mathbf{y}/\partial\mathbf{x}$ of the inverse transformation $\mathbf{y} \to \mathbf{x}$ is J^{-1} (from the product theorem for determinants: $det(AB) = detA.detB$ – see e.g. Blyth and Robertson (2002a), Th. 8.7).

Suppose now that $\mathbf{X}$ is a random n-vector with density $f(\mathbf{x})$, and we wish to change from $\mathbf{X}$ to $\mathbf{Y}$, where $\mathbf{Y}$ corresponds to $\mathbf{X}$ as $\mathbf{y}$ above corresponds to $\mathbf{x}$: $\mathbf{y} = \mathbf{y}(\mathbf{x})$ iff $\mathbf{x} = \mathbf{x}(\mathbf{y})$. If $\mathbf{Y}$ has density $g(\mathbf{y})$, then by above,

$$P(\mathbf{X} \in A) = \int_A f(\mathbf{x})\ d\mathbf{x} = \int_B f(\mathbf{x}(\mathbf{y})) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|\ d\mathbf{y},$$

and also

$$P(\mathbf{X} \in A) = P(\mathbf{Y} \in B) = \int_B g(\mathbf{y}) d\mathbf{y}.$$

Since these hold for all B, the integrands must be equal, giving

$$g(\mathbf{y}) = f(\mathbf{x}(\mathbf{y}))|\partial\mathbf{x}/\partial\mathbf{y}|$$

as the density g of $\mathbf{Y}$.

In particular, if the change of variables is linear:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}, \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{y} - \mathbf{A}^{-1}\mathbf{b}, \quad \partial\mathbf{y}/\partial\mathbf{x} = |\mathbf{A}|, \quad \partial\mathbf{x}/\partial\mathbf{y} = |\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}.$$

2.3 The Fisher F-distribution

Suppose we have two independent random variables U and V, chi–square distributed with degrees of freedom (df) m and n respectively. We divide each by its df, obtaining U/m and V/n. The distribution of the *ratio*

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

will be important below. It is called the *F-distribution* with *degrees of freedom* (m, n), $F(m, n)$. It is also known as the (Fisher) *variance-ratio distribution.*

Before introducing its density, we define the *Beta function*,

$$B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx,$$

wherever the integral converges ($\alpha > 0$ for convergence at 0, $\beta > 0$ for convergence at 1). By *Euler's integral for the Beta function*,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

(see e.g. Copson (1935), §9.3). One may then show that the density of $F(m, n)$ is

$$f(x) = \frac{m^{\frac{1}{2}m}n^{\frac{1}{2}n}}{B(\frac{1}{2}m, \frac{1}{2}m)} . \frac{x^{\frac{1}{2}(m-2)}}{(mx + n)^{\frac{1}{2}(m+n)}} \qquad (m, n > 0, \quad x > 0)$$

(see e.g. Kendall and Stuart (1977), §16.15, §11.10; the original form given by Fisher is slightly different).

There are two important features of this density. The first is that (to within a normalisation constant, which, like many of those in Statistics, involves ratios of Gamma functions) it behaves near zero like the power $x^{\frac{1}{2}(m-2)}$ and near infinity like the power $x^{-\frac{1}{2}n}$, and is smooth and unimodal (has one peak). The second is that, like all the common and useful distributions in Statistics, its percentage points are *tabulated.* Of course, using tables of the F-distribution involves the complicating feature that one has *two* degrees of freedom (rather than one as with the chi-square or Student t-distributions), and that these must be taken in the correct *order.* It is sensible at this point for the reader to take some time to gain familiarity with use of tables of the F-distribution, using whichever standard set of statistical tables are to hand. Alternatively, all standard statistical packages will provide percentage points of F, t, χ^2, etc. on demand. Again, it is sensible to take the time to gain familiarity with the statistical package of your choice, including use of the online Help facility.

One can derive the density of the F distribution from those of the χ^2 distributions above. One needs the formula for the density of a quotient of random variables. The derivation is left as an exercise; see Exercise 2.1. For an introduction to calculations involving the F distribution see Exercise 2.2.

2.4 Orthogonality

Recall that a square, non-singular $(n \times n)$ matrix A is *orthogonal* if its inverse is its transpose:

$$A^{-1} = A^T.$$

We now show that the property of being independent $N(0, \sigma^2)$ is preserved under an orthogonal transformation.

Theorem 2.2 (Orthogonality Theorem)

If $X = (X_1, \ldots, X_n)^T$ is an n-vector whose components are independent random variables, normally distributed with mean 0 and variance σ^2, and we change variables from X to Y by

$$Y := AX$$

where the matrix A is orthogonal, then the components Y_i of Y are again independent, normally distributed with mean 0 and variance σ^2.

Proof

We use the Jacobian formula. If $A = (a_{ij})$, since $\partial Y_i / \partial X_j = a_{ij}$, the Jacobian $\partial Y / \partial X = |A|$. Since A is orthogonal, $AA^T = AA^{-1} = I$. Taking determinants, $|A|.|A^T| = |A|.|A| = 1$: $|A| = 1$, and similarly $|A^T| = 1$. Since length is preserved under an orthogonal transformation,

$$\sum\nolimits_1^n Y_i^2 = \sum\nolimits_1^n X_i^2.$$

The joint density of $(X_1, \ldots, X_n)$ is, by independence, the product of the marginal densities, namely

$$f(x_1, \ldots, x_n) = \prod\nolimits_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\} = \frac{1}{(2\pi)^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2}\sum\nolimits_1^n x_i^2\right\}.$$

From this and the Jacobian formula, we obtain the joint density of $(Y_1, \ldots, Y_n)$ as

$$f(y_1, \ldots, y_n) = \frac{1}{(2\pi)^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2}\sum\nolimits_1^n y_i^2\right\} = \prod\nolimits_1^n \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y_i^2\right\}.$$

But this is the joint density of n independent standard normals – and so $(Y_1, \ldots, Y_n)$ are independent standard normal, as claimed. □

Helmert's Transformation.

There exists an orthogonal $n \times n$ matrix P with first row

$$\frac{1}{\sqrt{n}}(1, \ldots, 1)$$

(there are many such! Robert Helmert (1843–1917) made use of one when he introduced the χ^2 distribution in 1876 – see Kendall and Stuart (1977), Example 11.1 – and it is convenient to use his name here for any of them.) For, take this vector, which spans a one-dimensional subspace; take $n-1$ unit vectors not in this subspace and use the Gram–Schmidt orthogonalisation process (see e.g. Blyth and Robertson (2002b), Th. 1.4) to obtain a set of n orthonormal vectors.

2.5 Normal sample mean and sample variance

For $X_1, \ldots, X_n$ independent and identically distributed (iid) random variables, with mean μ and variance σ^2, write

$$\overline{X} := \frac{1}{n}\sum\nolimits_1^n X_i$$

for the *sample mean* and

$$S^2 := \frac{1}{n}\sum\nolimits_1^n (X_i - \overline{X})^2$$

for the *sample variance.*

Note 2.3

Many authors use $1/(n-1)$ rather than $1/n$ in the definition of the sample variance. This gives S^2 as an *unbiased* estimator of the population variance σ^2. But our definition emphasizes the parallel between the bar, or average, for sample quantities and the expectation for the corresponding population quantities:

$$\overline{X} = \frac{1}{n}\sum\nolimits_1^n X_i \leftrightarrow EX,$$

$$S^2 = \overline{\left(X - \overline{X}\right)^2} \;\leftrightarrow\; \sigma^2 = E\left[(X - EX)^2\right],$$

which is mathematically more convenient.

Theorem 2.4

If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$,
(i) the sample mean $\overline{X}$ and the sample variance S^2 are independent,
(ii) $\overline{X}$ is $N(\mu, \sigma^2/n)$,
(iii) nS^2/σ^2 is $\chi^2(n-1)$.

Proof

(i) Put $Z_i := (X_i - \mu)/\sigma$, $Z := (Z_1, \ldots, Z_n)^T$; then the Z_i are iid $N(0,1)$,

$$\overline{Z} = (\overline{X} - \mu)/\sigma, \qquad nS^2/\sigma^2 = \sum\nolimits_1^n (Z_i - \overline{Z})^2.$$

Also, since

$$\begin{aligned}\sum\nolimits_1^n (Z_i - \overline{Z})^2 &= \sum\nolimits_1^n Z_i^2 - 2\overline{Z}\sum\nolimits_1^n Z_i + n\overline{Z}^2 \\ &= \sum\nolimits_1^n Z_i^2 - 2\overline{Z}.n\overline{Z} + n\overline{Z}^2 = \sum\nolimits_1^n Z_i^2 - n\overline{Z}^2 : \\ \sum\nolimits_1^n Z_i^2 &= \sum\nolimits_1^n (Z_i - \overline{Z})^2 + n\overline{Z}^2.\end{aligned}$$

The terms on the right above are quadratic forms, with matrices A, B say, so we can write

$$\sum\nolimits_1^n Z_i^2 = Z^T A Z + Z^T B X. \qquad (*)$$

Put $W := PZ$ with P a Helmert transformation – P orthogonal with first row $(1, \ldots, 1)/\sqrt{n}$:

$$W_1 = \frac{1}{\sqrt{n}} \sum\nolimits_1^n Z_i = \sqrt{n}\overline{Z}; \qquad W_1^2 = n\overline{Z}^2 = Z^T B Z.$$

So

$$\sum_2^n W_i^2 = \sum_1^n W_i^2 - W_1^2 = \sum_1^n Z_i^2 - Z^T B Z = Z^T A Z = \sum_1^n (Z_i - \overline{Z})^2 = nS^2/\sigma^2.$$

But the W_i are independent (by the orthogonality of P), so W_1 is independent of $W_2, \ldots, W_n$. So W_1^2 is independent of $\sum_2^n W_i^2$. So nS^2/σ^2 is independent of $n(\overline{X} - \mu)^2/\sigma^2$, so S^2 is independent of $\overline{X}$, as claimed.
(ii) We have $\overline{X} = (X_1 + \ldots + X_n)/n$ with X_i independent, $N(\mu, \sigma^2)$, so with MGF $\exp(\mu t + \frac{1}{2}\sigma^2 t^2)$. So X_i/n has MGF $\exp(\mu t/n + \frac{1}{2}\sigma^2 t^2/n^2)$, and $\overline{X}$ has MGF

$$\prod_1^n \exp\left(\mu t/n + \frac{1}{2}\sigma^2 t^2/n^2\right) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2/n\right).$$

So $\overline{X}$ is $N(\mu, \sigma^2/n)$.
(iii) In $(*)$, we have on the left $\sum_1^n Z_i^2$, which is the sum of the squares of n standard normals Z_i, so is $\chi^2(n)$ with MGF $(1-2t)^{-\frac{1}{2}n}$. On the right, we have

two independent terms. As $\overline{Z}$ is $N(0, 1/n)$, $\sqrt{n}\overline{Z}$ is $N(0, 1)$, so $n\overline{Z}^2 = Z^T BZ$ is $\chi^2(1)$, with MGF $(1-2t)^{-\frac{1}{2}}$. Dividing (as in chi-square subtraction above), $Z^T AZ = \sum_1^n (Z_i - \overline{Z})^2$ has MGF $(1-2t)^{-\frac{1}{2}(n-1)}$. So $Z^T AZ = \sum_1^n (Z_i - \overline{Z})^2$ is $\chi^2(n-1)$. So nS^2/σ^2 is $\chi^2(n-1)$. □

Note 2.5

1. This is a remarkable result. We quote (without proof) that this property actually *characterises* the normal distribution: if the sample mean and sample variance are independent, then the population distribution is normal (Geary's Theorem: R. C. Geary (1896–1983) in 1936; see e.g. Kendall and Stuart (1977), Examples 11.9 and 12.7).
2. The fact that when we form the sample mean, the mean is unchanged, while the variance decreases by a factor of the sample size n, is true generally. The point of (ii) above is that normality is preserved. This holds more generally: it will emerge in Chapter 4 that normality is preserved under any linear operation.

Theorem 2.6 (Fisher's Lemma)

Let $X_1, \ldots, X_n$ be iid $N(0, \sigma^2)$. Let

$$Y_i = \sum\nolimits_{j=1}^{n} c_{ij} X_j \qquad (i = 1, \ldots, p, \quad p < n),$$

where the row-vectors $(c_{i1}, \ldots, c_{in})$ are orthogonal for $i = 1, \ldots, p$. If

$$S^2 = \sum\nolimits_1^n X_i^2 - \sum\nolimits_1^p Y_i^2,$$

then
(i) S^2 is independent of $Y_1, \ldots, Y_p$,
(ii) S^2 is $\chi^2(n-p)$.

Proof

Extend the $p \times n$ matrix (c_{ij}) to an $n \times n$ orthogonal matrix $C = (c_{ij})$ by Gram–Schmidt orthogonalisation. Then put

$$Y := CX,$$

so defining $Y_1, \ldots, Y_p$ (again) and $Y_{p+1}, \ldots, Y_n$. As C is orthogonal, $Y_1, \ldots, Y_n$ are iid $N(0, \sigma^2)$, and $\sum_1^n Y_i^2 = \sum_1^n X_i^2$. So

$$S^2 = \left(\sum\nolimits_1^n - \sum\nolimits_1^p \right) Y_i^2 = \sum\nolimits_{p+1}^n Y_i^2$$

is independent of $Y_1, \ldots, Y_p$, and S^2/σ^2 is $\chi^2(n-p)$. □

2.6 One-Way Analysis of Variance

To compare two normal means, we use the Student t-test, familiar from your first course in Statistics. What about comparing r means for $r > 2$?

Analysis of Variance goes back to early work by Fisher in 1918 on mathematical genetics and was further developed by him at Rothamsted Experimental Station in Harpenden, Hertfordshire in the 1920s. The convenient acronym ANOVA was coined much later, by the American statistician John W. Tukey (1915–2000), the pioneer of exploratory data analysis (EDA) in Statistics (Tukey (1977)), and coiner of the terms hardware, software and bit from computer science.

Fisher's motivation (which arose directly from the agricultural field trials carried out at Rothamsted) was to compare yields of several varieties of crop, say – or (the version we will follow below) of one crop under several fertiliser *treatments*. He realised that if there was more variability between groups (of yields with different treatments) than within groups (of yields with the same treatment) than one would expect if the treatments were the same, then this would be evidence against believing that they were the same. In other words, Fisher set out to *compare means by analysing variability* ('variance' – the term is due to Fisher – is simply a short form of 'variability').

We write μ_i for the mean yield of the ith variety, for $i = 1, \ldots, r$. For each i, we draw n_i independent readings X_{ij}. The X_{ij} are independent, and we assume that they are normal, all with the same unknown variance σ^2:

$$X_{ij} \sim N(\mu_i, \sigma^2) \qquad (j = 1, \ldots, n_i, \quad i = 1, \ldots, r).$$

We write

$$n := \sum\nolimits_1^r n_i$$

for the total sample size.

With two suffices i and j in play, we use a bullet to indicate that the suffix in that position has been averaged out. Thus we write

$$X_{i\bullet}, \quad \text{or} \quad \overline{X}_i, := \frac{1}{n_i}\sum\nolimits_{j=1}^{n_i} X_{ij} \qquad (i = 1, \ldots, r)$$

for the ith *group mean* (the sample mean of the ith sample)

$$X_{\bullet\bullet}, \quad \text{or} \quad \overline{X}, := \frac{1}{n}\sum\nolimits_{i=1}^{r}\sum\nolimits_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum\nolimits_{i=1}^{r} n_i X_{i\bullet}$$

for the *grand mean* and,

$$S_i^2 := \frac{1}{n_i}\sum\nolimits_{j=1}^{n_i}(X_{ij} - X_{i\bullet})^2$$

for the ith sample variance.

Define the *total sum of squares*

$$SS := \sum\nolimits_{i=1}^{r}\sum\nolimits_{j=1}^{n_i}(X_{ij} - X_{\bullet\bullet})^2 = \sum\nolimits_i\sum\nolimits_j[(X_{ij} - X_{i\bullet}) + (X_{i\bullet} - X_{\bullet\bullet})]^2.$$

As

$$\sum\nolimits_j(X_{ij} - X_{i\bullet}) = 0$$

(from the definition of $X_{i\bullet}$ as the average of the X_{ij} over j), if we expand the square above, the cross terms vanish, giving

$$\begin{aligned} SS &= \sum\nolimits_i\sum\nolimits_j(X_{ij} - X_{i\bullet})^2 \\ &\quad +\sum\nolimits_i\sum\nolimits_j(X_{ij} - X_{i\bullet})(X_{i\bullet} - X_{\bullet\bullet}) \\ &\quad +\sum\nolimits_i\sum\nolimits_j(X_{i\bullet} - X_{\bullet\bullet})^2 \\ &= \sum\nolimits_i\sum\nolimits_j(X_{ij} - X_{i\bullet})^2 + \sum\nolimits_i\sum\nolimits_j X_{i\bullet} - X_{\bullet\bullet})^2 \\ &= \sum\nolimits_i n_i S_i^2 + \sum\nolimits_i n_i(X_{i\bullet} - X_{\bullet\bullet})^2. \end{aligned}$$

The first term on the right measures the amount of variability *within* groups. The second measures the variability *between* groups. We call them the *sum of squares for error* (or *within groups*), SSE, also known as the *residual sum of squares*, and the *sum of squares for treatments* (or *between groups*), respectively:

$$SS = SSE + SST,$$

where

$$SSE := \sum\nolimits_i n_i S_i^2, \qquad SST := \sum\nolimits_i n_i(X_{i\bullet} - X_{\bullet\bullet})^2.$$

Let H_0 be the null hypothesis of no treatment effect:

$$H_0: \qquad \mu_i = \mu \qquad (i = 1, \ldots, r).$$

If H_0 is true, we have merely one large sample of size n, drawn from the distribution $N(\mu, \sigma^2)$, and so

$$SS/\sigma^2 = \frac{1}{\sigma^2}\sum\nolimits_i\sum\nolimits_j(X_{ij} - X_{\bullet\bullet})^2 \sim \chi^2(n-1) \qquad \text{under } H_0.$$

In particular,

$$E[SS/(n-1)] = \sigma^2 \qquad \text{under } H_0.$$

Whether or not H_0 is true,

$$n_i S_i^2/\sigma^2 = \frac{1}{\sigma^2}\sum\nolimits_j (X_{ij} - X_{i\bullet})^2 \sim \chi^2(n_i - 1).$$

So by the Chi-Square Addition Property

$$SSE/\sigma^2 = \sum\nolimits_i n_i S_i^2/\sigma^2 = \frac{1}{\sigma^2}\sum\nolimits_i \sum\nolimits_j (X_{ij} - X_{i\bullet})^2 \sim \chi^2(n - r),$$

since as $n = \sum_i n_i$,

$$\sum\nolimits_{i=1}^{r} (n_i - 1) = n - r.$$

In particular,

$$E[SSE/(n - r)] = \sigma^2.$$

Next,

$$SST := \sum_i n_i (X_{i\bullet} - X_{\bullet\bullet})^2, \quad \text{where} \quad X_{\bullet\bullet} = \frac{1}{n}\sum_i n_i X_{i\bullet}, \quad SSE := \sum_i n_i S_i^2.$$

Now S_i^2 is independent of $X_{i\bullet}$, as these are the sample variance and sample mean from the ith sample, whose independence was proved in Theorem 2.4. Also S_i^2 is independent of $X_{j\bullet}$ for $j \neq i$, as they are formed from different independent samples. Combining, S_i^2 is independent of all the $X_{j\bullet}$, so of their (weighted) average $X_{\bullet\bullet}$, so of SST, a function of the $X_{j\bullet}$ and of $X_{\bullet\bullet}$. So $SSE = \sum_i n_i S_i^2$ is also independent of SST.

We can now use the Chi-Square Subtraction Property. We have, under H_0, the independent sum

$$SS/\sigma^2 = SSE/\sigma^2 +_{ind} SST/\sigma^2.$$

By above, the left-hand side is $\chi^2(n - 1)$, while the first term on the right is $\chi^2(n - r)$. So the second term on the right must be $\chi^2(r - 1)$. This gives:

Theorem 2.7

Under the conditions above and the null hypothesis H_0 of no difference of treatment means, we have the sum-of-squares decomposition

$$SS = SSE +_{ind} SST,$$

where $SS/\sigma^2 \sim \chi^2(n - 1)$, $SSE/\sigma^2 \sim \chi^2(n - r)$ and $SSE/\sigma^2 \sim \chi^2(r - 1)$.

When we have a sum of squares, chi-square distributed, and we divide by its degrees of freedom, we will call the resulting ratio a *mean sum of squares*, and denote it by changing the SS in the name of the sum of squares to MS. Thus the mean sum of squares is

$$MS := SS/\mathrm{df}(SS) = SS/(n-1)$$

and the mean sums of squares for treatment and for error are

$$\begin{aligned} MST &:= SST/\mathrm{df}(SST) = SST/(r-1), \\ MSE &:= SSE/\mathrm{df}(SSE) = SSE/(n-r). \end{aligned}$$

By the above,

$$SS = SST + SSE;$$

whether or not H_0 is true,

$$E[MSE] = E[SSE]/(n-r) = \sigma^2;$$

under H_0,

$$E[MS] = E[SS]/(n-1) = \sigma^2, \qquad \text{and so also} \qquad E[MST]/(r-1) = \sigma^2.$$

Form the F-statistic

$$F := MST/MSE.$$

Under H_0, this has distribution $F(r-1, n-r)$. Fisher realised that comparing the size of this F-statistic with percentage points of this F-distribution gives us a way of testing the truth or otherwise of H_0. Intuitively, if the treatments do differ, this will tend to inflate SST, hence MST, hence $F = MST/MSE$. To justify this intuition, we proceed as follows. Whether or not H_0 is true,

$$\begin{aligned} SST &= \sum\nolimits_i n_i (X_{i\bullet} - X_{\bullet\bullet})^2 = \sum\nolimits_i n_i X_{i\bullet}^2 - 2X_{\bullet\bullet} \sum\nolimits_i n_i X_{i\bullet} + X_{\bullet\bullet}^2 \sum\nolimits_i n_i \\ &= \sum\nolimits_i n_i X_{i\bullet}^2 - nX_{\bullet\bullet}^2, \end{aligned}$$

since $\sum_i n_i X_{i\bullet} = nX_{\bullet\bullet}$ and $\sum_i n_i = n$. So

$$\begin{aligned} E[SST] &= \sum\nolimits_i n_i E\left[X_{i\bullet}^2\right] - nE\left[X_{\bullet\bullet}^2\right] \\ &= \sum\nolimits_i n_i \left[\mathrm{var}(X_{i\bullet}) + (EX_{i\bullet})^2\right] - n\left[\mathrm{var}(X_{\bullet\bullet}) + (EX_{\bullet\bullet})^2\right]. \end{aligned}$$

But $\mathrm{var}(X_{i\bullet}) = \sigma^2/n_i$,

$$\begin{aligned} \mathrm{var}(X_{\bullet\bullet}) &= \mathrm{var}(\frac{1}{n}\sum\nolimits_{i=1}^r n_i X_{i\bullet}) = \frac{1}{n^2}\sum\nolimits_1^r n_i^2 \mathrm{var}(X_{i\bullet}), \\ &= \frac{1}{n^2}\sum\nolimits_1^r n_i^2 \sigma^2/n_i = \sigma^2/n \end{aligned}$$

(as $\sum_i n_i = n$). So writing

$$\overline{\mu} := \frac{1}{n}\sum\nolimits_i n_i \mu_i = EX_{\bullet\bullet} = E\frac{1}{n}\sum\nolimits_i n_i X_{i\bullet},$$

$$\begin{aligned} E(SST) &= \sum\nolimits_1^r n_i \left[\frac{\sigma^2}{n_i} + \mu_i^2\right] - n\left[\frac{\sigma^2}{n} + \overline{\mu}^2\right] \\ &= (r-1)\sigma^2 + \sum\nolimits_i n_i \mu_i^2 - n\overline{\mu}^2 \\ &= (r-1)\sigma^2 + \sum\nolimits_i n_i (\mu_i - \overline{\mu})^2 \end{aligned}$$

(as $\sum_i n_i = n$, $n\overline{\mu} = \sum_i n_i \mu_i$). This gives the inequality

$$E[SST] \geq (r-1)\sigma^2,$$

with equality iff

$$\mu_i = \overline{\mu} \quad (i = 1, \ldots, r), \qquad \text{i.e.} \qquad H_0 \text{ is true.}$$

Thus when H_0 is *false*, the mean of SST *increases*, so *larger* values of SST, so of MST and of $F = MST/MSE$, are evidence *against* H_0. It is thus appropriate to use a *one-tailed* F-test, rejecting H_0 if the value F of our F-statistic is *too big*. How big is too big depends, of course, on our chosen significance level α, and hence on the tabulated value $F_{tab} := F_\alpha(r-1, n-r)$, the upper α-point of the relevant F-distribution. We summarise:

Theorem 2.8

When the null hypothesis H_0 (that all the treatment means $\mu_1, \ldots, \mu_r$ are equal) is true, the F-statistic $F := MST/MSE = (SST/(r-1))/(SSE/(n-r))$ has the F-distribution $F(r-1, n-r)$. When the null hypothesis is false, F increases. So large values of F are evidence against H_0, and we test H_0 using a one-tailed test, rejecting at significance level α if F is too big, that is, with critical region

$$F > F_{tab} = F_\alpha(r-1, n-r).$$

Model Equations for One-Way ANOVA.

$$X_{ij} = \mu_i + \epsilon_{ij} \qquad (i = 1, \ldots, r, \quad j = 1, \ldots, r), \qquad \epsilon_{ij} \quad \text{iid} \quad N(0, \sigma^2).$$

Here μ_i is the *main effect* for the ith treatment, the null hypothesis is H_0: $\mu_1 = \ldots = \mu_r = \mu$, and the unknown variance σ^2 is a nuisance parameter. The point of forming the ratio in the F-statistic is to cancel this nuisance parameter σ^2, just as in forming the ratio in the Student t-statistic in one's first course in Statistics. We will return to nuisance parameters in §5.1.1 below.

Calculations.

In any calculation involving variances, there is cancellation to be made, which is worthwhile and important numerically. This stems from the definition and 'computing formula' for the variance,

$$\sigma^2 := E\left[(X - EX)^2\right] = E\left[X^2\right] - (EX)^2$$

and its sample counterpart

$$S^2 := \overline{(X - \overline{X})^2} = \overline{X^2} - \overline{X}^2.$$

Writing T, T_i for the grand total and group totals, defined by

$$T := \sum\nolimits_i \sum\nolimits_j X_{ij}, \qquad T_i := \sum\nolimits_j X_{ij},$$

so $X_{\bullet\bullet} = T/n$, $nX_{\bullet\bullet}^2 = T^2/n$:

$$SS = \sum\nolimits_i \sum\nolimits_j X_{ij}^2 - T^2/n,$$

$$SST = \sum\nolimits_i T_i^2/n_i - T^2/n,$$

$$SSE = SS - SST = \sum\nolimits_i \sum\nolimits_j X_{ij}^2 - \sum\nolimits_i T_i^2/n_i.$$

These formulae help to reduce rounding errors and are easiest to use if carrying out an Analysis of Variance by hand.

It is customary, and convenient, to display the output of an Analysis of Variance by an ANOVA table, as shown in Table 2.1. (The term 'Error' can be used in place of 'Residual' in the 'Source' column.)

Source	df	SS	Mean Square	F
Treatments	$r-1$	SST	$MST = SST/(r-1)$	MST/MSE
Residual	$n-r$	SSE	$MSE = SSE/(n-r)$	
Total	$n-1$	SS		

Table 2.1 One-way ANOVA table.

Example 2.9

We give an example which shows how to calculate the Analysis of Variance tables by hand. The data in Table 2.2 come from an agricultural experiment. We wish to test for different mean yields for the different fertilisers. We note that

Fertiliser	Yield
A	14.5, 12.0, 9.0, 6.5
B	13.5, 10.0, 9.0, 8.5
C	11.5, 11.0, 14.0, 10.0
D	13.0, 13.0, 13.5, 7.5
E	15.0, 12.0, 8.0, 7.0
F	12.5, 13.5, 14.0, 8.0

Table 2.2 Data for Example 2.9

we have six treatments so $6-1=5$ degrees of freedom for treatments. The total number of degrees of freedom is the number of observations minus one, hence 23. This leaves 18 degrees of freedom for the within-treatments sum of squares. The total sum of squares can be calculated routinely as $\sum(y_{ij}-\overline{y}^2)=\sum y_{ij}^2 - n\overline{y}^2$, which is often most efficiently calculated as $\sum y_{ij}^2 - (1/n)\left(\sum y_{ij}\right)^2$. This calculation gives $SS = 3119.25 - (1/24)(266.5)^2 = 159.990$. The easiest next step is to calculate SST, which means we can then obtain SSE by subtraction as above. The formula for SST is relatively simple and reads $\sum_i T_i/n_i - T^2/n$, where T_i denotes the sum of the observations corresponding to the ith treatment and $T=\sum_{ij} y_{ij}$. Here this gives $SST = (1/4)(42^2+41^2+46.5^2+47^2+42^2+48^2)-1/24(266.5)^2 = 11.802$. Working through, the full ANOVA table is shown in Table 2.3.

Source	df	Sum of Squares	Mean Square	F
Between fertilisers	5	11.802	2.360	0.287
Residual	18	148.188	8.233	
Total	23	159.990		

Table 2.3 One-way ANOVA table for Example 2.9

This gives a non-significant p-value compared with $F_{3,16}(0.95) = 3.239$. R calculates the p-value to be 0.914. Alternatively, we may place bounds on the p-value by looking at statistical tables. In conclusion, we have no evidence for differences between the various types of fertiliser.

In the above example, the calculations were made more simple by having equal numbers of observations for each treatment. However, the same general procedure works when this no longer continues to be the case. For detailed worked examples with unequal sample sizes see Snedecor and Cochran (1989) §12.10.

S-Plus/R®.

We briefly describe implementation of one-way ANOVA in S-Plus/R®. For background and details, see e.g. Crawley (2002), Ch. 15. Suppose we are studying the dependence of yield on treatment, as above. [Note that this requires that we set treatment to be a *factor* variable, taking discrete rather than continuous values, which can be achieved by setting `treatment <- factor(treatment)`.] Then, using `aov` as short for 'Analysis of Variance', `<-` for the assignment operator in S-Plus (read as 'goes to' or 'becomes') and ∼ as short for 'depends on' or 'is regressed on', we use

```
model <- aov (yield ~ treatment)
```

to do the analysis, and ask for the summary table by

```
summary(model)
```

A complementary `anova` command is summarised briefly in Chapter 5.2.1.

2.7 Two-Way ANOVA; No Replications

In the agricultural experiment considered above, problems may arise if the growing area is not homogeneous. The plots on which the different treatments are applied may differ in fertility – for example, if a field slopes, nutrients tend to leach out of the soil and wash downhill, so lower-lying land may give higher yields than higher-lying land. Similarly, differences may arise from differences in drainage, soil conditions, exposure to sunlight or wind, crops grown in the past, etc. If such differences are not taken into account, we will be unable to distinguish between differences in yield resulting from differences in *treatment*, our object of study, and those resulting from differences in growing conditions – *plots*, for short – which are not our primary concern. In such a case, one says that treatments are *confounded* with plots – we would have no way of separating the effect of one from that of the other.

The only way out of such difficulties is to subdivide the growing area into plots, each of which can be treated as a homogeneous growing area, and then subdivide each plot and apply different treatments to the different sub-plots or blocks. In this way we will be 'comparing like with like', and avoid the pitfalls of confounding.

When allocating treatments to blocks, we may wish to *randomise*, to avoid the possibility of inadvertently introducing a treatment-block linkage. Relevant here is the subject of *design of experiments*; see §9.3.

In the sequel, we assume for simplicity that the block sizes are the same and the number of treatments is the same for each block. The model equations will now be of the form

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad (i = 1, \ldots, r, \quad j = 1, \ldots, n).$$

Here μ is the *grand mean* (or *overall mean*); α_i is the ith *treatment effect* (we take $\sum_i \alpha_i = 0$, otherwise this sum can – and so should – be absorbed into μ; β_j is the jth *block effect* (similarly, we take $\sum_j \beta_j = 0$); the errors ϵ_{ij} are iid $N(0, \sigma^2)$, as before.

Recall the terms $X_{i\bullet}$ from the one-way case; their counterparts here are similarly denoted $X_{\bullet j}$. Start with the algebraic identity

$$(X_{ij} - X_{\bullet\bullet}) = (X_{ij} - X_{i\bullet} - X_{\bullet j} + X_{\bullet\bullet}) + (X_{i\bullet} - X_{\bullet\bullet}) + (X_{\bullet j} - X_{\bullet\bullet}).$$

Square and add. One can check that the cross terms cancel, leaving only the squared terms. For example, $(X_{ij} - X_{i\bullet} - X_{\bullet j} + X_{\bullet\bullet})$ averages over i to $-(X_{\bullet j} - X_{\bullet\bullet})$, and over j to $-(X_{\bullet j} - X_{\bullet\bullet})$, while each of the other terms on the right involves only one of i and j, and so is unchanged when averaged over the other. One is left with

$$\begin{aligned}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - X_{\bullet\bullet})^2 \quad = \quad & \sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - X_{i\bullet} - X_{\bullet j} + X_{\bullet\bullet})^2 \\ & + n\sum_{i=1}^{r}(X_{i\bullet} - X_{\bullet\bullet})^2 \\ & + r\sum_{j=1}^{n}(X_{\bullet j} - X_{\bullet\bullet})^2.\end{aligned}$$

We write this as

$$SS = SSE + SST + SSB,$$

giving the total sum of squares SS as the sum of the sum of squares for error (SSE), the sum of squares for treatments (SST) (as before) and a new term, the sum of squares for *blocks*, (SSB). The degrees of freedom are, respectively, $nr - 1$ for SS (the total sample size is nr, and we lose one df in estimating σ), $r - 1$ for treatments (as before), $n - 1$ for blocks (by analogy with treatments – or equivalently, there are n block parameters β_j, but they are subject to one constraint, $\sum_j \beta_j = 0$), and $(n - 1)(r - 1)$ for error (to give the correct total in the df column in the table below). Independence of the three terms on the right follows by arguments similar to those in the one-way case. We can accordingly construct a two-way ANOVA table, as in Table 2.4.

Here we have *two* F-statistics, $FT := MST/MSE$ for treatment effects and $FB := MSB/MSE$ for block effects. Accordingly, we can test *two* null hypotheses, one, $H_0(T)$, for presence of a treatment effect and one, $H_0(B)$, for presence of a block effect.

Source	df	SS	Mean Square	F
Treatments	$r-1$	SST	$MST = \frac{SST}{r-1}$	MST/MSE
Blocks	$n-1$	SSB	$MSB = \frac{SSB}{n-1}$	MSB/MSE
Residual	$(r-1)(n-1)$	SSE	$MSE = \frac{SSE}{(r-1)(n-1)}$	
Total	$rn-1$	SS		

Table 2.4 Two-way ANOVA table

Note 2.10

In educational psychology (or other behavioural sciences), 'treatments' might be different questions on a test, 'blocks' might be *individuals*. We take it for granted that individuals differ. So we need not calculate MSB nor test $H_0(B)$ (though packages such as S-Plus will do so automatically). Then $H_0(T)$ as above tests for differences between mean scores on questions in a test. (Where the questions carry equal credit, such differences are undesirable – but may well be present in practice!)

Implementation. In S-Plus, the commands above extend to

```
model <- aov(yield ~ treatment + block)

summary(model)
```

Example 2.11

We illustrate the two-way Analysis of Variance with an example. We return to the agricultural example in Example 2.9, but suppose that the data can be linked to growing areas as shown in Table 2.5. We wish to test the hypothesis that there are no differences between the various types of fertiliser. The

Fertiliser	Area 1	Area 2	Area 3	Area 4
A	14.5	12.0	9.0	6.5
B	13.5	10.0	9.0	8.5
C	11.5	11.0	14.0	10.0
D	13.0	13.0	13.5	7.5
E	15.0	12.0	8.0	7.0
F	12.5	13.5	14.0	8.0

Table 2.5 Data for Example 2.11

sum-of-squares decomposition for two-way ANOVA follows in an analogous way to the one-way case. There are relatively simple formulae for SS, SST, and SSB, meaning that SSE can easily be calculated by subtraction. In detail, these formulae are

$$\begin{aligned} SS &= \sum\nolimits_{ij} X_{ij}^2 - \frac{1}{nr}\left(\sum X_{ij}\right)^2, \\ SST &= \left(X_{1\bullet}^2 + \ldots + X_{r\bullet}^2\right)/n - \frac{1}{nr}\left(\sum X_{ij}\right)^2, \\ SSB &= \left(X_{\bullet 1}^2 + \ldots + X_{\bullet n}^2\right)/r - \frac{1}{nr}\left(\sum X_{ij}\right)^2, \end{aligned}$$

with $SSE = SS - SST - SSB$. Returning to our example, we see that

$$\begin{aligned} SS &= 3119.25 - (1/24)(266.5)^2 = 159.990, \\ SST &= (42^2 + 41^2 + 46.5^2 + 47^2 + 42^2 + 48^2)/4 - (1/24)(266.5)^2 = 11.802, \\ SSB &= (80^2 + 71.5^2 + 67.5^2 + 47.5^2)/6 - (1/24)(266.5)^2 = 94.865. \end{aligned}$$

By subtraction $SSE = 159.9896 - 11.80208 - 94.86458 = 53.323$. These calculations lead us to the ANOVA table in Table 2.6. Once again we have no evidence for differences amongst the 6 types of fertiliser. The variation that does occur is mostly due to the effects of different growing areas.

Source	df	S.S.	MS	F	p
Fertilisers	5	11.802	2.360	0.664	0.656
Area	3	94.865	31.622	8.895	0.001
Residual	15	53.323	3.555		
Total	23	159.990			

Table 2.6 Two-way ANOVA table for Example 2.11

2.8 Two-Way ANOVA: Replications and Interaction

In the above, we have one reading X_{ij} for each *cell*, or combination of the ith treatment and the jth block. But we may have more. Suppose we have m *replications* – independent readings – per cell. We now need three suffices rather than two. The model equations will now be of the form

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \qquad (i = 1, \ldots, r, \quad j = 1, \ldots, n, \quad k = 1, \ldots, m).$$

Here the new parameters γ_{ij} measure possible *interactions* between treatment and block effects. This allows one to study situations in which effects are *not additive.* Although we use the word interaction here as a technical term in Statistics, this is fully consistent with its use in ordinary English. We are all familiar with situations where, say, a medical treatment (e.g. a drug) may interact with some aspect of our diet (e.g. alcohol). Similarly, two drugs may interact (which is why doctors must be careful in checking what medication a patient is currently taking before issuing a new prescription). Again, different alcoholic drinks may interact (folklore wisely counsels against mixing one's drinks), etc.

Arguments similar to those above lead to the following sum-of-squares decomposition:

$$\begin{aligned}\sum\nolimits_{i=1}^{r}\sum\nolimits_{j=1}^{n}(X_{ijk}-X_{\bullet\bullet\bullet})^2 \;=\; & \sum\nolimits_{i}\sum\nolimits_{j}\sum\nolimits_{k}(X_{ijk}-X_{ij\bullet})^2\\ & +nm\sum\nolimits_{i}(X_{i\bullet\bullet}-X_{\bullet\bullet\bullet})^2\\ & +rm\sum\nolimits_{j}(X_{\bullet j\bullet}-X_{\bullet\bullet\bullet})^2\\ & +m\sum\nolimits_{i}\sum\nolimits_{j}(X_{ij\bullet}-X_{i\bullet\bullet}-X_{\bullet j\bullet}+X_{\bullet\bullet\bullet})^2.\end{aligned}$$

We write this as

$$SS = SSE + SST + SSB + SSI,$$

where the new term is the sum of squares for *interactions.* The degrees of freedom are $r-1$ for treatments as before, $n-1$ for blocks as before, $(r-1)(n-1)$ for interactions (the product of the effective number of parameters for treatments and for blocks), $rnm-1$ in total (there are rnm readings), and $rn(m-1)$ for error (so that the df totals on the right and left above agree).
Implementation. The S-Plus/R® commands now become

```
model <- aov(yield ~ treatment * block)

summary(model)
```

This notation is algebraically motivated, and easy to remember. With *additive* effects, we used a $+$. We now use a $*$, suggestive of the possibility of 'product' terms representing the interactions. We will encounter many more such situations in the next chapter, when we deal with multiple regression.

The summary table now takes the form of Table 2.7. We now have *three* F-statistics, FT and FB as before, and now FI also, which we can use to test for the presence of interactions.

Source	df	SS	Mean Square	F
Treatments	$r-1$	SST	$MST = \frac{SST}{r-1}$	MST/MSE
Blocks	$n-1$	SSB	$MSB = \frac{SSB}{n-1}$	MSB/MSE
Interaction	$(r-1)(n-1)$	SSI	$MSI = \frac{SSI}{(r-1)(n-1)}$	MSI/MSE
Residual	$rn(m-1)$	SSE	$MSE = \frac{SSE}{rn(m-1)}$	
Total	$rmn-1$	SS		

Table 2.7 Two-way ANOVA table with interactions

Example 2.12

The following example illustrates the procedure for two-way ANOVA with interactions. The data in Table 2.8 link the growth of hamsters of different coat colours when fed different diets.

	Light coat	Dark coat
Diet A	6.6, 7.2	8.3, 8.7
Diet B	6.9, 8.3	8.1, 8.5
Diet C	7.9, 9.2	9.1, 9.0

Table 2.8 Data for Example 2.12

The familiar formula for the total sum of squares gives $SS = 805.2 - (97.8^2/12) = 8.13$. In a similar manner to Example 2.11, the main effects sum-of-squares calculations give

$$\begin{aligned} SST &= \sum \frac{y_{i\bullet\bullet}^2}{nm} - \frac{\left(\sum_{ijk} y_{ijk}\right)^2}{rmn}, \\ SSB &= \frac{y_{\bullet j\bullet}^2}{rm} - \frac{\left(\sum_{ijk} y_{ijk}\right)^2}{rmn}, \end{aligned}$$

and in this case give $SST = (1/4)(30.8^2 + 31.8^2 + 35.2^2) - (97.8^2/12) = 2.66$ and $SSB = (1/6)(46.1^2 + 51.7^2) - (97.8^2/12) = 2.613$. The interaction sum of squares can be calculated as a sum of squares corresponding to every cell in the table once the main effects of SST and SSB have been accounted for. The calculation is

$$SSI = \frac{1}{m}\sum y_{ij\bullet}^2 - SST - SSB - \frac{\left(\sum_{ijk} y_{ijk}\right)^2}{rmn},$$

which in this example gives $SSI = (1/2)(13.8^2 + 17^2 + 15.2^2 + 16.6^2 + 17.1^2 + 18.1^2) - 2.66 - 2.613 - (97.8^2/12) = 0.687$. As before, SSE can be calculated by subtraction, and the ANOVA table is summarised in Table 2.9. The results suggest that once we take into account the different types of coat, the effect of the different diets is seen to become only borderline significant. The diet:coat interaction term is seen to be non-significant and we might consider in a subsequent analysis the effects of deleting this term from the model.

Source	df	SS	MS	F	p
Diet	2	2.66	1.33	3.678	0.091
Coat	1	2.613	2.613	7.226	0.036
Diet:Coat	2	0.687	0.343	0.949	0.438
Residual	5	2.17	0.362		
Total	11	8.13			

Table 2.9 Two-way ANOVA with interactions for Example 2.12.

Note 2.13 (Random effects)

The model equation for two-way ANOVA with interactions is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

with $\sum_i \alpha_i = \sum_j \beta_j = \sum_{ij} \gamma_{ij} = 0$. Here the α_i, β_j, γ_{ij} are constants, and the randomness is in the errors ϵ_{ijk}. Suppose, however, that the β_i were themselves random (in the examination set-up above, the suffix i might refer to the ith question, and suffix j to the jth candidate; the candidates might be chosen at random from a larger population). We would then use notation such as

$$y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + \epsilon_{ijk}.$$

Here we have both a fixed effect (for questions, i) and a random effect (for candidates, j). With both fixed and random effects, we speak of a *mixed* model; see §9.1.

With only random effects, we have a *random effects model*, and use notation such as

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + \epsilon_{ijk}.$$

We restrict for simplicity here to the model with no interaction terms:

$$y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}.$$

Assuming independence of the random variables on the right, the variances add (see e.g. Haigh (2002), Cor. 5.6):

$$\sigma_y^2 = \sigma_a^2 + \sigma_b^2 + \sigma_\epsilon^2,$$

in an obvious notation. The terms on the right are called *variance components*; see e.g. Searle, Casella and McCulloch (1992) for a detailed treatment.

Variance components can be traced back to work of Airy in 1861 on astronomical observations (recall that astronomy also led to the development of Least Squares by Legendre and Gauss).

EXERCISES

2.1. (i) Show that if X, Y are positive random variables with joint density $f(x, y)$ their quotient $Z := X/Y$ has density

$$h(z) = \int_0^\infty y f(yz, y)\, dy \quad (z > 0).$$

So if X, Y are independent with densities f, g,

$$h(z) = \int_0^\infty y f(yz) g(y)\, dy \quad (z > 0).$$

(ii) If X has density f and $c > 0$, show that X/c has density

$$f_{X/c}(x) = c f(cx).$$

(iii) Deduce that the Fisher F-distribution $F(m, n)$ has density

$$h(z) = m^{\frac{1}{2}m} n^{\frac{1}{2}n} \frac{\Gamma(\frac{1}{2}m + \frac{1}{2}n)}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} \cdot \frac{z^{\frac{1}{2}m-1}}{(n + mz)^{\frac{1}{2}(m+n)}} \quad (z > 0).$$

2.2. Using tables or S-Plus/R® produce bounds or calculate the exact probabilities for the following statements. [Note. In S-Plus/R® the command `pf` may prove useful.]
(i) $\mathrm{P}(X < 1.4)$ where $X \sim F_{5,17}$,
(ii) $\mathrm{P}(X > 1)$ where $X \sim F_{1,16}$,
(iii) $\mathrm{P}(X < 4)$ where $X \sim F_{1,3}$,
(iv) $\mathrm{P}(X > 3.4)$ where $X \sim F_{19,4}$,
(v) $\mathrm{P}(\ln\ X\ > -1.4)$ where $X \sim F_{10,4}$.

Fat 1	Fat 2	Fat 3	Fat 4
164	178	175	155
172	191	193	166
168	197	178	149
177	182	171	164
156	185	163	170
195	177	176	168

Table 2.10 Data for Exercise 2.3.

2.3. *Doughnut data.* Doughnuts absorb fat during cooking. The following experiment was conceived to test whether the amount of fat absorbed depends on the type of fat used. Table 2.10 gives the amount of fat absorbed per batch of doughnuts. Produce the one-way Analysis of Variance table for these data. What is your conclusion?

2.4. The data in Table 2.11 come from an experiment where growth is measured and compared to the variable *photoperiod* which indicates the length of daily exposure to light. Produce the one-way ANOVA table for these data and determine whether or not growth is affected by the length of daily light exposure.

Very short	Short	Long	Very long
2	3	3	4
3	4	5	6
1	2	1	2
1	1	2	2
2	2	2	2
1	1	2	3

Table 2.11 Data for Exercise 2.4

2.5. *Unpaired t-test with equal variances.* Under the null hypothesis the statistic t defined as

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\left(\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)\right)}{s}$$

should follow a t distribution with $n_1 + n_2 - 2$ degrees of freedom, where n_1 and n_2 denote the number of observations from samples 1 and 2 and s is the pooled estimate given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

where

$$\begin{aligned} s_1^2 &= \frac{1}{n_1-1}(\sum x_1^2 - (n_1-1)\overline{x}_1^2), \\ s_2^2 &= \frac{1}{n_2-1}(\sum x_2^2 - (n_2-1)\overline{x}_2^2). \end{aligned}$$

(i) Give the relevant statistic for a test of the hypothesis $\mu_1 = \mu_2$ and $n_1 = n_2 = n$.
(ii) Show that if $n_1 = n_2 = n$ then one-way ANOVA recovers the same results as the unpaired t-test. [Hint. Show that the F-statistic satisfies $F_{1,2(n-1)} = t^2_{2(n-1)}$].

2.6. Let Y_1, Y_2 be iid $N(0,1)$. Give values of a and b such that

$$a(Y_1 - Y_2)^2 + b(Y_1 + Y_2)^2 \sim \chi^2_2.$$

2.7. Let Y_1, Y_2, Y_3 be iid $N(0,1)$. Show that

$$\frac{1}{3}\left[(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + (Y_3 - Y_1)^2\right] \sim \chi^2_2.$$

Generalise the above result for a sample $Y_1, Y_2, \ldots, Y_n$ of size n.

2.8. The data in Table 2.12 come from an experiment testing the number of failures out of 100 planted soyabean seeds, comparing four different seed treatments, with no treatment ('check'). Produce the two-way ANOVA table for this data and interpret the results. (We will return to this example in Chapter 8.)

Treatment	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5
Check	8	10	12	13	11
Arasan	2	6	7	11	5
Spergon	4	10	9	8	10
Semesan, Jr	3	5	9	10	6
Fermate	9	7	5	5	3

Table 2.12 Data for Exercise 2.8

2.9. *Photoperiod example revisited.* When we add in knowledge of plant genotype the full data set is as shown in Table 2.13. Produce the two-way ANOVA table and revise any conclusions from Exercise 2.4 in the light of these new data as appropriate.

Genotype	Very short	Short	Long	Very Long
A	2	3	3	4
B	3	4	5	6
C	1	2	1	2
D	1	1	2	2
E	2	2	2	2
F	1	1	2	3

Table 2.13 Data for Exercise 2.9

2.10. *Two-way ANOVA with interactions.* Three varieties of potato are planted on three plots at each of four locations. The yields in bushels are given in Table 2.14. Produce the ANOVA table for these data. Does the interaction term appear necessary? Describe your conclusions.

Variety	Location 1	Location 2	Location 3	Location 4
A	15, 19, 22	17, 10, 13	9, 12, 6	14, 8, 11
B	20, 24, 18	24, 18, 22	12, 15, 10	21, 16, 14
C	22, 17, 14	26, 19, 21	10, 5, 8	19, 15, 12

Table 2.14 Data for Exercise 2.10

2.11. *Two-way ANOVA with interactions.* The data in Table 2.15 give the gains in weight of male rats from diets with different sources and different levels of protein. Produce the two-way ANOVA table with interactions for these data. Test for the presence of interactions between source and level of protein and state any conclusions that you reach.

Source	High Protein	Low Protein
Beef	73, 102, 118, 104, 81, 107, 100, 87, 117, 111	90, 76, 90, 64, 86, 51, 72, 90, 95, 78
Cereal	98, 74, 56, 111, 95, 88, 82, 77, 86, 92	107, 95, 97, 80, 98, 74, 74, 67, 89, 58
Pork	94, 79, 96, 98, 102, 102, 108, 91, 120, 105	49, 82, 73, 86, 81, 97, 106, 70, 61, 82

Table 2.15 Data for Exercise 2.11

3
Multiple Regression

3.1 The Normal Equations

We saw in Chapter 1 how the model

$$y_i = a + bx_i + \epsilon_i, \qquad \epsilon_i \quad \text{iid} \quad N(0, \sigma^2)$$

for simple linear regression occurs. We saw also that we may need to consider two or more regressors. We dealt with two regressors u and v, and could deal with three regressors u, v and w similarly. But in general we will need to be able to handle any number of regressors, and rather than rely on the finite resources of the alphabet it is better to switch to suffix notation, and use the language of vectors and matrices. For a random vector $\mathbf{X}$, we will write $E\mathbf{X}$ for its *mean vector* (thus the mean of the ith coordinate X_i is $E(X_i) = (E\mathbf{X})_i$), and var$(\mathbf{X})$ for its *covariance matrix* (whose (i, j) entry is $\text{cov}(X_i, X_j)$). We will use p regressors, called $x_1, \ldots, x_p$, each with a corresponding parameter $\beta_1, \ldots, \beta_p$ ('p for parameter'). In the equation above, regard a as short for $a.1$, with 1 as a regressor corresponding to a constant term (the intercept term in the context of linear regression). Then for one reading ('a sample of size 1') we have the model

$$y = \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon, \qquad \epsilon_i \quad \sim \quad N(0, \sigma^2).$$

In the general case of a sample of size n, we need two suffices, giving the model equations

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad \epsilon_i \quad \text{iid} \quad N(0, \sigma^2) \qquad (i = 1, \ldots, n).$$

N.H. Bingham and J.M. Fry, *Regression: Linear Models in Statistics*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-969-5_3,

Writing the typical term on the right as $x_{ij}\beta_j$, we recognise the form of a matrix product. Form $y_1, \ldots, y_n$ into a column vector $\mathbf{y}$, $\epsilon_1, \ldots, \epsilon_n$ into a column vector ϵ, $\beta_1, \ldots, \beta_p$ into a column vector β, and x_{ij} into a matrix X (thus $\mathbf{y}$ and ϵ are $n \times 1$, β is $p \times 1$ and X is $n \times p$). Then our system of equations becomes one matrix equation, the *model equation*

$$\mathbf{y} = X\beta + \epsilon. \qquad (ME)$$

This matrix equation, and its consequences, are the object of study in this chapter. Recall that, as in Chapter 1, n is the sample size – the larger the better – while p, the number of parameters, is small – as small as will suffice. We will have more to say on choice of p later. Typically, however, p will be at most five or six, while n could be some tens or hundreds. Thus we must expect n to be *much larger* than p, which we write as

$$n >> p.$$

In particular, the $n \times p$ matrix X has no hope of being invertible, as it is not even square (a common student howler).

Note 3.1

We pause to introduce the objects in the model equation (ME) by name. On the left is $\mathbf{y}$, the *data*, or *response vector*. The last term ϵ is the *error* or *error vector*; β is the *parameter* or *parameter vector*. Matrix X is called the *design matrix*. Although its (i, j) entry arose above as the ith value of the jth regressor, for most purposes from now on x_{ij} is just a *constant*. Emphasis shifts from these constants to the *parameters*, β_j.

Note 3.2

To underline this shift of emphasis, it is often useful to change notation and write A for X, when the model equation becomes

$$\mathbf{y} = A\beta + \epsilon. \qquad (ME)$$

Lest this be thought a trivial matter, we mention that Design of Experiments (initiated by Fisher) is a subject in its own right, on which numerous books have been written, and to which we return in §9.3.

We will feel free to use either notation as seems most convenient at the time. While X is the natural choice for straight regression problems, as in this chapter, it is less suitable in the general Linear Model, which includes related contexts such as Analysis of Variance (Chapter 2) and Analysis of Covariance (Chapter 5). Accordingly, we shall usually prefer A to X for use in developing theory.

We make a further notational change. As we shall be dealing from now on with vectors rather than scalars, there is no need to remind the reader of this by using boldface type. We may thus lighten the notation by using y for $\mathbf{y}$, etc.; thus we now have

$$y = A\beta + \epsilon, \qquad (ME)$$

for use in this chapter (in Chapter 4 below, where we again use x as a scalar variable, we use $\mathbf{x}$ for a vector variable).

From the model equation

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

the likelihood is

$$\begin{aligned} L &= \frac{1}{\sigma^n (2\pi)^{\frac{1}{2}n}} \prod\nolimits_{i=1}^{n} \exp\left\{-\frac{1}{2}\left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\beta_j\right)^2 / \sigma^2\right\} \\ &= \frac{1}{\sigma^n (2\pi)^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2}\sum\nolimits_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\beta_j\right)^2 / \sigma^2\right\}, \end{aligned}$$

and the log-likelihood is

$$\ell := \log L = \text{const} - n\log\sigma - \frac{1}{2}\left[\sum\nolimits_{i=1}^{n}\left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\beta_j\right)^2\right] / \sigma^2.$$

As before, we use Fisher's Method of Maximum Likelihood, and maximise with respect to β_r: $\partial\ell/\partial\beta_r = 0$ gives

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

or

$$\sum\nolimits_{j=1}^{p}\left(\sum\nolimits_{i=1}^{n} a_{ir}a_{ij}\right)\beta_j = \sum\nolimits_{i=1}^{n} a_{ir}y_i.$$

Write $C = (c_{ij})$ for the $p \times p$ matrix

$$C := A^T A,$$

(called the *information matrix* – see Definition 3.10 below), which we note is *symmetric*: $C^T = C$. Then

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

So this says

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

In matrix notation, this is

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

or combining,

$$C\beta = A^T y, \qquad C := A^T A. \tag{NE}$$

These are the *normal equations*, the analogues for the general case of the normal equations obtained in Chapter 1 for the cases of one and two regressors.

3.2 Solution of the Normal Equations

Our next task is to solve the normal equations for β. Before doing so, we need to check that there exists a unique solution, the condition for which is, from Linear Algebra, that the information matrix $C := A^T A$ should be non-singular (see e.g. Blyth and Robertson (2002a), Ch. 4). This imposes an important condition on the design matrix A. Recall that the *rank* of a matrix is the maximal number of independent rows or columns. If this is as big as it could be given the size of the matrix, the matrix is said to have *full rank*, otherwise it has *deficient rank*. Since A is $n \times p$ with $n >> p$, A has full rank if its rank is p.

Recall from Linear Algebra that a square matrix C is *non-negative definite* if

$$x^T C x \geq 0$$

for all vectors x, while C is *positive definite* if

$$x^T C x > 0 \qquad \forall x \neq 0$$

(see e.g. Blyth and Robertson (2002b), Ch. 8). A positive definite matrix is non-singular, so invertible; a non-negative definite matrix need not be.

Lemma 3.3

If A ($n \times p$, $n > p$) has full rank p, $C := A^T A$ is positive definite.

Proof

As A has full rank, there is no vector x with $Ax = 0$ other than the zero vector (such an equation would give a non-trivial linear dependence relation between the columns of A). So

$$(Ax)^T Ax = x^T A^T Ax = x^T Cx = 0$$

only for $x = 0$, and is > 0 otherwise. This says that C is positive definite, as required. □

Note 3.4

The same proof shows that $C := A^T A$ is always non-negative definite, regardless of the rank of A.

Theorem 3.5

For A full rank, the normal equations have the unique solution

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

Proof

In the full-rank case, C is positive definite by Lemma 3.3, so invertible, so we may solve the normal equations to obtain the solution above. □

From now on, we restrict attention to the full-rank case: the design matrix A, which is $n \times p$, has full rank p.

Note 3.6

The distinction between the full- and deficient-rank cases is the same as that between the general and singular cases that we encountered in Chapter 1 in connection with the bivariate normal distribution. We will encounter it again later in Chapter 4, in connection with the multivariate normal distribution. In fact, this distinction bedevils the whole subject. Linear dependence causes rank-deficiency, in which case we should identify the linear dependence relation, use it to express some regressors (or columns of the design matrix) in terms of others, eliminate the redundant regressors or columns, and begin again in a lower dimension, where the problem will have full rank. What is worse is that *near*-linear dependence – which when regressors are at all numerous is not uncommon – means that one is close to rank-deficiency, and this makes things numerically unstable. Remember that in practice, we work numerically, and when one is within rounding error of rank-deficiency, one is close to disaster. We shall return to this vexed matter later (§4.4), in connection with *multicollinearity*. We note in passing that Numerical Linear Algebra is a subject in its own right; for a monograph treatment, see e.g. Golub and Van Loan (1996).

Just as in Chapter 1, the functional form of the normal likelihood means that maximising the likelihood minimises the sum of squares

$$SS := (y - A\beta)^T(y - A\beta) = \sum\nolimits_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\beta_j\right)^2 .$$

Accordingly, we have as before the following theorem.

Theorem 3.7

The solutions $(\hat{\beta})$ to the normal equations (NE) are both the maximum-likelihood estimators and the least-squares estimators of the parameters β.

There remains the task of estimating the remaining parameter σ. At the maximum, $\beta = \hat{\beta}$. So taking $\partial SS/\partial\sigma = 0$ in the log-likelihood

$$\ell := \log L = \text{const} - n\log\sigma - \frac{1}{2}\left[\sum\nolimits_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\beta_j\right)^2\right]/\sigma^2$$

gives, at the maximum,

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum\nolimits_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\beta_j\right)^2 = 0.$$

At the maximum, $\beta = \hat{\beta}$; rearranging, we have at the maximum that

$$\sigma^2 = \frac{1}{n}\sum\nolimits_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\hat{\beta}_j\right)^2 .$$

This sum of squares is, by construction, the minimum value of the total sum of squares SS as the parameter β varies, the minimum being attained at the least-squares estimate $\hat{\beta}$. This minimised sum of squares is called the *sum of squares for error*, SSE:

$$SSE = \sum\nolimits_{i=1}^{n} \left(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\hat{\beta}_j\right)^2 = \left(y - A\hat{\beta}\right)^T \left(y - A\hat{\beta}\right),$$

so-called because, as we shall see in Corollary 3.23 below, the unbiased estimator of the error variance σ^2 is $\hat{\sigma}^2 = SSE/(n-p)$.

We call

$$\hat{y} := A\hat{\beta}$$

the *fitted values*, and

$$e := y - \hat{y},$$

the difference between the actual values (data) and fitted values, the *residual* vector. If $e = (e_1, \ldots, e_n)$, the e_i are the *residuals*, and the sum of squares for error

$$SSE = \sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum\nolimits_{i=1}^{n} e_i^2$$

is the *sum of squared residuals.*

Note 3.8

We pause to discuss unbiasedness and degrees of freedom (df). In a first course in Statistics, one finds the maximum-likelihood estimators (MLEs) $\hat{\mu}$, $\hat{\sigma}^2$ of the parameters μ, σ^2 in a normal distribution $N(\mu, \sigma^2)$. One finds

$$\hat{\mu} = \overline{x}, \qquad \hat{\sigma}^2 = s_x^2 := \frac{1}{n}\sum\nolimits_{i=1}^{n} (x_i - \overline{x})^2$$

(and the distributions are given by $\overline{x} \sim N(\mu, \sigma^2/n)$ and $n\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1)$). But this is a *biased* estimator of σ^2; to get an *unbiased* estimator, one has to replace n in the denominator above by $n-1$ (in distributional terms: the mean of a chi-square is its df). This is why many authors use $n-1$ in place of n in the denominator when they *define* the sample variance (and we warned, when we used n in Chapter 1, that this was not universal!), giving what we will call the *unbiased* sample variance,

$$s_u^2 := \frac{1}{(n-1)}\sum\nolimits_{i=1}^{n} (x_i - \overline{x})^2.$$

The problem is that to estimate σ^2, one has first to estimate μ by $\overline{x}$. *Every time one has to estimate a parameter from the data, one loses a degree of freedom.* In this one-dimensional problem, the df accordingly decreases from n to $n-1$.

Returning to the general case: here we have to estimate p parameters, $\beta_1, \ldots, \beta_p$. Accordingly, we lose p degrees of freedom, and to get an unbiased estimator we have to divide, not by n as above but by $n-p$, giving the estimator

$$\hat{\sigma}^2 = \frac{1}{(n-p)} SSE.$$

Since n is much larger than p, the difference between this (unbiased) estimator and the previous (maximum-likelihood) version is not large, but it is worthwhile, and so we shall work with the unbiased version unless otherwise stated. We find its distribution in §3.4 below (and check it is unbiased – Corollary 3.23).

Note 3.9 (Degrees of Freedom)

Recall that n is our sample size, that p is our number of parameters, and that n is much greater than p. The need to estimate p parameters, which reduces the degrees of freedom from n to $n-p$, thus effectively reduces the sample size by this amount. We can think of the degrees of freedom as a measure of the *amount of information available* to us.

This interpretation is in the minds of statisticians when they prefer one procedure to another because it 'makes more degrees of freedom available' for the task in hand. We should always keep the degrees of freedom of all relevant terms (typically, sums of squares, or quadratic forms in normal variates) in mind, and think of keeping this large as being desirable.

We rewrite our conclusions so far in matrix notation. The total sum of squares is

$$SS := \sum\nolimits_{i=1}^{n} \Big(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\beta_j\Big)^2 = (y - A\beta)^T (y - A\beta)\,;$$

its minimum value with respect to variation in β is the sum of squares for error

$$SSE = \sum\nolimits_{i=1}^{n} \Big(y_i - \sum\nolimits_{j=1}^{p} a_{ij}\hat{\beta}_j\Big)^2 = \Big(y - A\hat{\beta}\Big)^T \Big(y - A\hat{\beta}\Big)\,,$$

where $\hat{\beta}$ is the solution to the normal equations (NE). Note that SSE is a *statistic* – we can calculate it from the data y and $\hat{\beta} = C^{-1}A^T y$, unlike SS which contains unknown parameters β.

One feature is amply clear already. To carry through a regression analysis in practice, we must perform considerable matrix algebra – or, with actual data, numerical matrix algebra – involving in particular the inversion of the $p \times p$ matrix $C := A^T A$. With matrices of any size, the calculations may well be laborious to carry out by hand. In particular, *matrix inversion* to find C^{-1} will be unpleasant for matrices larger than 2×2, even though C – being symmetric and positive definite – has good properties. For matrices of any size, one needs computer assistance. The package MATLAB®[1] is specially designed with matrix operations in mind. General mathematics packages such as Mathematica®[2] or Maple®[3] have a matrix inversion facility; so too do a number of statistical packages – for example, the `solve` command in S-Plus/R®.

QR Decomposition

The numerical solution of the normal equations ((NE) in §3.1, $(\hat{\beta})$ in Theorem 3.5) is simplified if the design matrix A (which is $n \times p$, and of full rank p) is given its *QR decomposition*

$$A = QR,$$

where Q is $n \times p$ and has *orthonormal columns* – so

$$Q^T Q = I$$

[1] MATLAB®, Simulink® and Symbolic Math Toolbox™ are trademarks of The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA, 01760-2098, USA, http://www.mathworks.com

[2] Mathematica® is a registered trademark of Wolfram Research, Inc., 100 Trade Center Drive, Champaign, IL 61820-7237, USA, http://www.wolfram.com

[3] Maple™ is a trademark of Waterloo Maple Inc., 615 Kumpf Drive, Waterloo, Ontario, Canada N2V 1K8, http://www.maplesoft.com

– and R is $p \times p$, *upper triangular*, and non-singular (has no zeros on the diagonal). This is always possible; see below. The normal equations $A^T A\hat{\beta} = A^T y$ then become

$$R^T Q^T Q R\hat{\beta} = R^T Q^T y,$$

or

$$R^T R\hat{\beta} = R^T Q^T y,$$

as $Q^T Q = I$, or

$$R\hat{\beta} = Q^T y,$$

as R, and so also R^T, is non-singular. This system of linear equations for $\hat{\beta}$ has an upper triangular matrix R, and so may be solved simply by back-substitution, starting with the bottom equation and working upwards.

The QR decomposition is just the expression in matrix form of the process of *Gram–Schmidt orthogonalisation*, for which see e.g. Blyth and Robertson (2002b), Th. 1.4. Write A as a row of its columns,

$$A = (a_1, \ldots, a_p);$$

the n-vectors a_i are linearly independent as A has full rank p. Write $q_1 := a_1/\|a_1\|$, and for $j = 2, \ldots, p$,

$$q_j := w_j/\|w_j\|, \qquad \text{where} \qquad w_j := a_j - \sum\nolimits_{k=1}^{j-1} (a_k^T q_k) q_k.$$

Then the q_j are orthonormal (are mutually orthogonal unit vectors), which span the column-space of A (Gram-Schmidt orthogonalisation is this process of passing from the a_j to the q_j). Each q_j is a linear combination of $a_1, \ldots, a_j$, and the construction ensures that, conversely, each a_j is a linear combination of $q_1, \ldots, q_j$. That is, there are scalars r_{kj} with

$$a_j = \sum\nolimits_{k=1}^{j} r_{kj} q_k \qquad (j = 1, \ldots, p).$$

Put $r_{kj} = 0$ for $k > j$. Then assembling the p columns a_j into the matrix A as above, this equation becomes

$$A = QR,$$

as required.

Note 3.10

Though useful as a theoretical tool, the Gram–Schmidt orthogonalisation process is not numerically stable. For numerical implementation, one needs a stable variant, the modified Gram-Schmidt process. For details, see Golub and Van Loan (1996), §5.2. They also give other forms of the QR decomposition (Householder, Givens, Hessenberg etc.).

3.3 Properties of Least-Squares Estimators

We have assumed *normal errors* in our model equations, (ME) of §3.1. But (until we need to assume normal errors in §3.5.2), we may work more generally, and assume only

$$Ey = A\beta, \qquad \mathrm{var}(y) = \sigma^2 I. \tag{ME^*}$$

We must then restrict ourselves to the Method of Least Squares, as without distributional assumptions we have no likelihood function, so cannot use the Method of Maximum Likelihood.

Linearity. The least-squares estimator

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

is *linear* in the data y.

Unbiasedness.

$$E\hat{\beta} = C^{-1}A^T Ey = C^{-1}A^T A\beta = C^{-1}C\beta = \beta :$$

$\hat{\beta}$ is an unbiased estimator of β.

Covariance matrix.

$$\begin{aligned}
\mathrm{var}(\hat{\beta}) = \mathrm{var}(C^{-1}A^T y) &= C^{-1}A^T(\mathrm{var}(y))(C^{-1}A^T)^T \\
&= C^{-1}A^T.\sigma^2 I.AC^{-1} \qquad (C = C^T) \\
&= \sigma^2.C^{-1}A^T.AC^{-1} \\
&= \sigma^2 C^{-1} \qquad (C = A^T A).
\end{aligned}$$

We wish to keep the variances of our estimators of our p parameters β_i small, and these are the diagonal elements of the covariance matrix above; similarly for the covariances (off-diagonal elements). The smaller the variances, the more precise our estimates, and the more information we have. This motivates the next definition.

Definition 3.11

The matrix $C := A^T A$, with A the design matrix, is called the *information matrix*.

Note 3.12

1. The variance σ^2 in our errors ϵ_i (which we of course wish to keep small) is usually beyond our control. However, at least at the stage of design and planning of the experiment, the design matrix A may well be within our control; hence so will be the information matrix $C := A^T A$, which we wish to maximise (in some sense), and hence so will be C^{-1}, which we wish to minimise in some sense. We return to this in §9.3 in connection with Design of Experiments.
2. The term information matrix is due to Fisher. It is also used in the context of parameter estimation by the *method of maximum likelihood*. One has the likelihood $L(\theta)$, with θ a vector parameter, and the log-likelihood $\ell(\theta) := \log L(\theta)$. The information matrix is the negative of the Hessian (matrix of second derivatives) of the log-likelihood: $I(\theta) := (I_{ij}(\theta))_{i,j=1}^p$, when

$$I_{ij}(\theta) := -\frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta).$$

Under suitable regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is asymptotically normal and unbiased, with variance matrix $(nI(\theta))^{-1}$; see e.g. Rao (1973), 5a.3, or Cramér (1946), §33.3.

Unbiased linear estimators. Now let $\tilde{\beta} := By$ be *any* unbiased linear estimator of β (B a $p \times n$ matrix). Then

$$E\tilde{\beta} = BEy = BA\beta = \beta$$

– and so $\tilde{\beta}$ is an unbiased estimator for β – iff

$$BA = I.$$

Note that

$$\mathrm{var}(\tilde{\beta}) = B\mathrm{var}(y)B^T = B.\sigma^2 I.B^T = \sigma^2 BB^T.$$

In the context of linear regression, as here, it makes sense to restrict attention to linear estimators. The two most obviously desirable properties of such estimators are unbiasedness (to get the mean right), and being minimum variance (to get maximum precision). An estimator with both these desirable properties may be termed a *best estimator*. A linear one is then a best linear unbiased estimator or BLUE (such acronyms are common in Statistics, and useful; an alternative usage is minimum variance unbiased linear estimate, or MVULE, but this is longer and harder to say). It is remarkable that the least-squares estimator that we have used above is best in this sense, or BLUE.

Theorem 3.13 (Gauss–Markov Theorem)

Among all unbiased linear estimators $\tilde{\beta} = By$ of β, the least-squares estimator $\hat{\beta} = C^{-1}A^T y$ has the minimum variance in each component. That is $\hat{\beta}$ is the BLUE.

Proof

By above, the covariance matrix of an arbitrary unbiased linear estimate $\tilde{\beta} = By$ and of the least-squares estimator $\hat{\beta}$ are given by

$$\mathrm{var}(\tilde{\beta}) = \sigma^2 BB^T \quad \text{and} \quad \mathrm{var}(\hat{\beta}) = \sigma^2 C^{-1}.$$

Their difference (which we wish to show is non-negative) is

$$\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}) = \sigma^2[BB^T - C^{-1}].$$

Now using symmetry of C, C^{-1}, and $BA = I$ (so $A^T B^T = I$) from above,

$$(B - C^{-1}A^T)(B - C^{-1}A^T)^T = (B - C^{-1}A^T)(B^T - AC^{-1}).$$

Further,

$$\begin{aligned}(B - C^{-1}A^T)(B^T - AC^{-1}) &= BB^T - BAC^{-1} - C^{-1}A^T B^T + C^{-1}A^T AC^{-1} \\ &= BB^T - C^{-1} - C^{-1} + C^{-1} \qquad (C = A^T A) \\ &= BB^T - C^{-1}.\end{aligned}$$

Combining,

$$\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}) = \sigma^2 (B - C^{-1}A^T)(B - C^{-1}A^T)^T.$$

Now for a matrix $M = (m_{ij})$,

$$(MM^T)_{ii} = \sum\nolimits_k m_{ik}(M^T)_{ki} = \sum\nolimits_k m_{ik}^2,$$

the sum of the squares of the elements on the ith row of matrix M. So the ith diagonal entry above is

$$\mathrm{var}(\tilde{\beta}_i) = \mathrm{var}(\hat{\beta}_i) + \sigma^2(\text{sum of squares of elements on } i\text{th row of } B - C^{-1}A^T).$$

So

$$\mathrm{var}(\tilde{\beta}_i) \geq \mathrm{var}(\hat{\beta}_i),$$

and

$$\mathrm{var}(\tilde{\beta}_i) = \mathrm{var}(\hat{\beta}_i)$$

iff $B - C^{-1}A^T$ has ith row zero. So *some* $\tilde{\beta}_i$ has greater variance than $\hat{\beta}_i$ unless $B = C^{-1}A^T$ (i.e., unless *all* rows of $B - C^{-1}A^T$ are zero) – that is, unless $\tilde{\beta} = By = C^{-1}A^T y = \hat{\beta}$, the least-squares estimator, as required. □

One may summarise all this as: whether or not errors are assumed normal, **LEAST SQUARES IS BEST.**

Note 3.14

The Gauss–Markov theorem is in fact a misnomer. It is due to Gauss, in the early eighteenth century; it was treated in the book Markov (1912) by A. A. Markov (1856–1922). A misreading of Markov's book gave rise to the impression that he had rediscovered the result, and the name Gauss–Markov theorem has stuck (partly because it is useful!).

Estimability. A linear combination $c^T\beta = \sum_{i=1}^{p} c_i\beta_i$, with $c = (c_1, \ldots, c_p)^T$ a known p-vector, is called *estimable* if it has an unbiased linear estimator, $b^T y = \sum_{i=1}^{n} b_i y_i$, with $b = (b_1, \ldots, b_n)^T$ a known n-vector. Then

$$E(b^T y) = b^T E(y) = b^T A\beta = c^T\beta.$$

This can hold identically in the unknown parameter β iff

$$c^T = b^T A,$$

that is, c is a linear combination (by the n-vector b) of the n rows (p-vectors) of the design matrix A. The concept is due to R. C. Bose (1901–1987) in 1944.

In the full-rank case considered here, the rows of A span a space of full dimension p, and so all linear combinations are estimable. But in the defective rank case with rank $k < p$, the estimable functions span a space of dimension k, and non-estimable linear combinations exist.

3.4 Sum-of-Squares Decompositions

We define the *sum of squares for regression*, SSR, by

$$SSR := (\hat\beta - \beta)^T C(\hat\beta - \beta).$$

Since this is a quadratic form with matrix C which is positive definite, we have $SSR \geq 0$, and $SSR > 0$ unless $\hat\beta = \beta$, that is, unless the least-squares estimator is exactly right (which will, of course, never happen in practice).

Theorem 3.15 (Sum-of-Squares Decomposition)

$$SS = SSR + SSE. \qquad (SSD)$$

Proof

Write

$$y - A\beta = (y - A\hat{\beta}) + A(\hat{\beta} - \beta).$$

Now multiply the vector on each side by its transpose (that is, form the sum of squares of the coordinates of each vector). On the left, we obtain

$$SS = (y - A\beta)^T (y - A\beta),$$

the total sum of squares. On the right, we obtain three terms. The first squared term is

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

the sum of squares for error. The second squared term is

$$(A(\hat{\beta} - \beta))^T A(\hat{\beta} - \beta) = (\hat{\beta} - \beta)^T A^T A(\hat{\beta} - \beta) = (\hat{\beta} - \beta)^T C(\hat{\beta} - \beta) = SSR,$$

the sum of squares for regression. The cross terms on the right are

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

and its transpose, which are the same as both are scalars. But

$$A^T (y - A\hat{\beta}) = A^T y - A^T A\hat{\beta} = A^T y - C\hat{b} = 0,$$

by the normal equations (NE) of §3.1-3.2. Transposing,

$$(y - A\hat{\beta})^T A = 0.$$

So both cross terms vanish, giving $SS = SSR + SSE$, as required. □

Corollary 3.16

We have that

$$SSE = \min_{\beta} SS,$$

the minimum being attained at the least-squares estimator $\hat{\beta} = C^{-1} A^T y$.

Proof

$SSR \geq 0$, and $= 0$ iff $\beta = \hat{\beta}$. □

We now introduce the geometrical language of *projections*, to which we return in e.g. §3.5.3 and §3.6 below. The relevant mathematics comes from Linear Algebra; see the definition below. As we shall see, doing regression with p regressors amounts to an *orthogonal projection* on an appropriate p-dimensional subspace in n-dimensional space. The sum-of-squares decomposition involved can be visualised geometrically as an instance of *Pythagoras's Theorem*, as in the familiar setting of plane or solid geometry.

Definition 3.17

Call a linear transformation $P : V \rightarrow V$ a *projection* onto V_1 along V_2 if V is the direct sum $V = V_1 \oplus V_2$, and if $x = (x_1, x_2)^T$ with $Px = x_1$.

Then (Blyth and Robertson (2002b), Ch.2, Halmos (1979), §41) $V_1 = \text{Im } P = \text{Ker } (I - P)$, $V_2 = \text{Ker } P = \text{Im } (I - P)$.

Recall that a square matrix is *idempotent* if it is its own square $M^2 = M$. Then (Halmos (1979), §41), M is idempotent iff it is a projection.

For use throughout the rest of the book, with A the design matrix and $C := A^T A$ the information matrix, we write

$$P := AC^{-1}A^T$$

('P for projection' – see below). We note that P is symmetric. Note also

$$Py = AC^{-1}A^T y = A\hat{\beta},$$

by the normal equations (NE).

Lemma 3.18

P and $I - P$ are idempotent, and so are projections.

Proof

$P^2 = AC^{-1}A^T.AC^{-1}A^T = AC^{-1}A^T = P$:

$$P^2 = P.$$

$$(I - P)^2 = I - 2P + P^2 = I - 2P + P = I - P. \qquad \square$$

We now rewrite the two terms SSR and SSE on the right in Theorem 3.15 in the language of projections. Note that the first expression for SSE below shows again that it is a *statistic* – a function of the data (not involving unknown parameters), and so can be *calculated* from the data.

Theorem 3.19

$$SSE = y^T(I-P)y = (y-A\beta)^T(I-P)(y-A\beta),$$
$$SSR = (y-A\beta)^T P(y-A\beta).$$

Proof

As $SSE := \left(y-A\hat{\beta}\right)^T\left(y-A\hat{\beta}\right)$, and $A\hat{\beta} = Py$,

$$\begin{aligned} SSE &= \left(y-A\hat{\beta}\right)^T\left(y-A\hat{\beta}\right) \\ &= (y-Py)^T(y-Py) = y^T(I-P)(I-P)y = y^T(I-P)y, \end{aligned}$$

as $I-P$ is a projection.

For SSR, we have that

$$SSR := \left(\hat{\beta}-\beta\right)^T C\left(\hat{\beta}-\beta\right) = \left(\hat{\beta}-\beta\right)^T A^T A\left(\hat{\beta}-\beta\right).$$

But

$$\left(\hat{\beta}-\beta\right) = C^{-1}A^Ty - \beta = C^{-1}A^Ty - C^{-1}A^TA\beta = C^{-1}A^T(y-A\beta),$$

so

$$\begin{aligned} SSR &= (y-A\beta)^T AC^{-1}.A^TA.C^{-1}A^T(y-A\beta) \\ &= (y-A\beta)^T AC^{-1}A^T(y-A\beta) \qquad (A^TA = C) \\ &= (y-A\beta)^T P(y-A\beta), \end{aligned}$$

as required. The second formula for SSE follows from this and (SSD) by subtraction. □

Coefficient of Determination

The *coefficient of determination* is defined as R^2, where R is the (sample) correlation coefficient of the data and the fitted values that is of the pairs $(y_i, \hat{y}_i)$:

$$R := \sum (y_i - \overline{y}) \left(\hat{y}_i - \overline{\hat{y}}\right) / \sqrt{\sum (y_i - \overline{y})^2 \sum \left(\hat{y}_i - \overline{\hat{y}}\right)^2}.$$

Thus $-1 \leq R \leq 1$, $0 \leq R^2 \leq 1$, and R^2 is a measure of the *goodness of fit* of the fitted values to the data.

Theorem 3.20

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \overline{y})^2}.$$

For reasons of continuity, we postpone the proof to §3.4.1 below. Note that $R^2 = 1$ iff $SSE = 0$, that is, all the residuals are 0, and the fitted values are the exact values. As noted above, we will see in §3.6 that regression (estimating p parameters from n data points) amounts to a *projection* of the n-dimensional data space onto an p-dimensional hyperplane. So $R^2 = 1$ iff the data points lie in an p-dimensional hyperplane (generalising the situation of Chapter 1, where $R^2 = 1$ iff the data points lie on a line). In our full-rank (non-degenerate) case, this will not happen (see Chapter 4 for the theory of the relevant multivariate normal distribution), but the bigger R^2 is (or the smaller SSE is), the better the fit of our regression model to the data.

Note 3.21

R^2 provides a useful summary of the proportion of the variation in a data set explained by a regression. However, as discussed in Chapters 5 and 11 of Draper and Smith (1998) high values of R^2 can be misleading. In particular, we note that the values R^2 will tend to increase as additional terms are added to the model, irrespective of whether those terms are actually needed. An adjusted R^2 statistic which adds a penalty to complex models can be defined as

$$R_a^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right),$$

where n is the number of parameters and $n-p$ is the number of residual degrees of freedom; see Exercises 3.3, and §5.2 for a treatment of models penalised for complexity.

We note a result for later use.

Proposition 3.22 (Trace Formula)

$$E(x^T Ax) = \text{trace}(A.\text{var}(x)) + Ex^T.A.Ex.$$

Proof

$$x^T Ax = \sum\nolimits_{ij} a_{ij} x_i x_j,$$

so by linearity of E,

$$E[x^T Ax] = \sum\nolimits_{ij} a_{ij} E[x_i x_j].$$

Now $\text{cov}(x_i, x_j) = E(x_i x_j) - (Ex_i)(Ex_j)$, so

$$\begin{aligned} E\left[x^T Ax\right] &= \sum\nolimits_{ij} a_{ij}\left[\text{cov}(x_i x_j) + Ex_i.Ex_j\right] \\ &= \sum\nolimits_{ij} a_{ij}\text{cov}(x_i x_j) + \sum\nolimits_{ij} a_{ij}.Ex_i.Ex_j. \end{aligned}$$

The second term on the right is $Ex^T AEx$. For the first, note that

$$\text{trace}(AB) = \sum\nolimits_i (AB)_{ii} = \sum\nolimits_{ij} a_{ij} b_{ji} = \sum\nolimits_{ij} a_{ij} b_{ij},$$

if B is symmetric. But covariance matrices are symmetric, so the first term on the right is $\text{trace}(A\ \text{var}(x))$, as required. □

Corollary 3.23

$$\text{trace}(P) = p, \qquad \text{trace}(I - P) = n - p, \qquad E(SSE) = (n - p)\sigma^2.$$

So $\hat{\sigma}^2 := SSE/(n - p)$ is an unbiased estimator for σ^2.

Proof

By Theorem 3.19, SSE is a quadratic form in $y - A\beta$ with matrix $I - P = I - AC^{-1}A^T$. Now

$$\text{trace}(I - P) = \text{trace}(I - AC^{-1}A^T) = \text{trace}(I) - \text{trace}(AC^{-1}A^T).$$

But $\text{trace}(I) = n$ (as here I is the $n \times n$ identity matrix), and as $\text{trace}(AB) = \text{trace}(BA)$ (see Exercise 3.12),

$$\text{trace}(P) = \text{trace}(AC^{-1}A^T) = \text{trace}(C^{-1}A^T A) = \text{trace}(I) = p,$$

as here I is the $p \times p$ identity matrix. So

$$\text{trace}(I - P) = \text{trace}(I - AC^{-1}A^T) = n - p.$$

Since $Ey = A\beta$ and $\text{var}(y) = \sigma^2 I$, the Trace Formula gives

$$E(SSE) = (n - p)\sigma^2.$$

□

This last formula is analogous to the corresponding ANOVA formula $E(SSE) = (n - r)\sigma^2$ of §2.6. In §4.2 we shall bring the subjects of regression and ANOVA together.

3.4.1 Coefficient of determination

We now give the proof of Theorem 3.20, postponed in the above.

Proof

As at the beginning of Chapter 3 we may take our first regressor as 1, corresponding to the intercept term (this is not always present, but since R is translation-invariant, we may add an intercept term without changing R). The first of the normal equations then results from differentiating

$$\sum (y_i - \beta_1 - a_{2i}\beta_2 - \ldots - a_{pi}\beta_p)^2 = 0$$

with respect to β_1, giving

$$\sum (y_i - \beta_1 - a_{2i}\beta_2 - \ldots - a_{pi}\beta_p) = 0.$$

At the minimising values $\hat{\beta}_j$, this says

$$\sum (y_i - \hat{y}_i) = 0.$$

So

$$\overline{y} = \overline{\hat{y}}, \tag{a}$$

and also

$$\begin{aligned} \sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) &= \sum (y_i - \hat{y}_i)\hat{y}_i \\ &= (y - \hat{y})^T \hat{y} \\ &= (y - Py)^T Py \\ &= y^T (I - P) Py \\ &= y^T (P - P^2) y, \end{aligned}$$

so

$$\sum(y_i - \hat{y_i})(\hat{y_i} - \overline{y}) = 0, \qquad (b)$$

as P is a projection. So

$$\sum(y_i - \overline{y})^2 = \sum[(y_i - \hat{y_i}) + (\hat{y_i} - \overline{y})]^2 = \sum(y_i - \hat{y_i})^2 + \sum(\hat{y_i} - \overline{y})^2, \quad (c)$$

since the cross-term is 0. Also, in the definition of R,

$$\begin{aligned}\sum(y_i - \overline{y})(\hat{y_i} - \overline{\hat{y}}) &= \sum(y_i - \overline{y})(\hat{y_i} - \overline{y}) \quad \text{(by (a))} \\ &= \sum[(y_i - \hat{y_i}) + (\hat{y_i} - \overline{y})](\hat{y_i} - \overline{y}) \\ &= \sum(\hat{y_i} - \overline{y})^2 \qquad \text{(by } (b)\text{)}.\end{aligned}$$

So

$$R^2 = \frac{\left[\sum(\hat{y_i} - \overline{y})^2\right]^2}{(\sum(y_i - \overline{y})^2 \sum(\hat{y_i} - \overline{y})^2)} = \frac{\sum(\hat{y_i} - \overline{y})^2}{\sum(y_i - \overline{y})^2}.$$

By (c),

$$\begin{aligned}R^2 &= \frac{\sum(\hat{y_i} - \overline{y})^2}{\sum(y_i - \hat{y_i})^2 + \sum(\hat{y_i} - \overline{y})^2} \\ &= 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \hat{y_i})^2 + \sum(\hat{y_i} - \overline{y})^2} \\ &= 1 - \frac{SSE}{\sum(y_i - \overline{y})^2},\end{aligned}$$

by (c) again and the definition of SSE. □

3.5 Chi-Square Decomposition

Recall (Theorem 2.2) that if $x = x_1, \ldots, x_n$ is $N(0, I)$ – that is, if the x_i are iid $N(0, 1)$ – and we change variables by an orthogonal transformation B to

$$y := Bx,$$

then also $y \sim N(0, I)$. Recall from Linear Algebra (e.g. Blyth and Robertson (2002a) Ch. 9) that λ is an *eigenvalue* of a matrix A with *eigenvector* x $(\neq 0)$ if

$$Ax = \lambda x$$

(x is *normalised* if $x^T x = \Sigma_i x_i^2 = 1$, as is always possible).

Recall also (see e.g. Blyth and Robertson (2002b), Corollary to Theorem 8.10) that if A is a real symmetric matrix, then A can be diagonalised by an orthogonal transformation B, to D, say:

$$B^T AB = D$$

(see also Theorem 4.12 below, Spectral Decomposition) and that (see e.g. Blyth and Robertson (2002b), Ch. 9) if λ is an eigenvalue of A,

$$|D - \lambda I| = \left|B^T AB - \lambda I\right| = \left|B^T AB - \lambda B^T B\right| = \left|B^T\right| |A - \lambda I| \, |B| = 0.$$

Then a quadratic form in normal variables with matrix A is also a quadratic form in normal variables with matrix D, as

$$x^T Ax = x^T BDB^T x = y^T Dy, \qquad y := B^T x.$$

3.5.1 Idempotence, Trace and Rank

Recall that a (square) matrix M is *idempotent* if $M^2 = M$.

Proposition 3.24

If B is idempotent,

(i) its eigenvalues λ are 0 or 1,

(ii) its trace is its rank.

Proof

(i) If λ is an eigenvalue of B, with eigenvector x, $Bx = \lambda x$ with $x \neq 0$. Then

$$B^2 x = B(Bx) = B(\lambda x) = \lambda(Bx) = \lambda(\lambda x) = \lambda^2 x,$$

so λ^2 is an eigenvalue of B^2 (always true – that is, does not need idempotence). So

$$\lambda x = Bx = B^2 x = \ldots = \lambda^2 x,$$

and as $x \neq 0$, $\lambda = \lambda^2$, $\lambda(\lambda - 1) = 0$: $\lambda = 0$ or 1.

(ii)

$$\begin{aligned} \text{trace}(B) &= \text{sum of eigenvalues} \\ &= \#\text{ non-zero eigenvalues} \\ &= rank(B). \end{aligned}$$

□

Corollary 3.25

$$rank(P) = p, \qquad rank(I - P) = n - p.$$

Proof

This follows from Corollary 3.23 and Proposition 3.24. □

Thus $n = p + (n - p)$ is an instance of the Rank–Nullity Theorem ('dim source =dim Ker + dim Im'): Blyth and Robertson (2002a), Theorem 6. 4) applied to P, $I - P$.

3.5.2 Quadratic forms in normal variates

We will be interested in symmetric projection (so idempotent) matrices P. Because their eigenvalues are 0 and 1, we can diagonalise them by orthogonal transformations to a diagonal matrix of 0s and 1s. So if P has rank r, a quadratic form $x^T Px$ can be reduced to a sum of r squares of standard normal variates. By relabelling variables, we can take the 1s to precede the 0s on the diagonal, giving

$$x^T Px = y_1^2 + \ldots + y_r^2, \qquad y_i \quad \text{iid} \quad N(0, \sigma^2).$$

So $x^T Px$ is σ^2 times a $\chi^2(r)$-distributed random variable.

To summarise:

Theorem 3.26

If P is a symmetric projection of rank r and the x_i are independent $N(0, \sigma^2)$, the quadratic form

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

3.5.3 Sums of Projections

As we shall see below, a sum-of-squares decomposition, which expresses a sum of squares (chi-square distributed) as a sum of independent sums of squares (also chi-square distributed) corresponds to a decomposition of the identity I

as a sum of orthogonal projections. Thus Theorem 3.13 corresponds to $I = P + (I - P)$, but in Chapter 2 we encountered decompositions with more than two summands (e.g., $SS = SSB + SST + SSI$ has three). We turn now to the general case.

Suppose that $P_1, \ldots, P_k$ are symmetric projection matrices with sum the identity:

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

Take the trace of both sides: the $n \times n$ identity matrix I has trace n. Each P_i has trace its rank n_i, by Proposition 3.24, so

$$n = n_1 + \ldots + n_k.$$

Then squaring,

$$I = I^2 = \sum\nolimits_i P_i^2 + \sum\nolimits_{i<j} P_i P_j = \sum\nolimits_i P_i + \sum\nolimits_{i<j} P_i P_j.$$

Taking the trace,

$$n = \sum n_i + \sum\nolimits_{i<j} \text{trace}(P_i P_j) = n + \sum\nolimits_{i<j} \text{trace}(P_i P_j) :$$

$$\sum\nolimits_{i<j} \text{trace}(P_i P_j) = 0.$$

Hence

$$\begin{aligned} \text{trace}(P_i P_j) &= \text{trace}(P_i^2 P_j^2) && \text{(since } P_i,\ P_j \text{ projections)} \\ &= \text{trace}((P_j P_i).(P_i P_j)) && (\text{trace}(AB) = \text{trace}(BA)) \\ &= \text{trace}((P_i P_j)^T.(P_i P_j)), \end{aligned}$$

since $(AB)^T = B^T A^T$ and P_i, P_j symmetric and where we have defined $A = P_i P_i P_j$, $B = P_j$. Hence we have that

$$\text{trace}(P_i P_j) \geq 0,$$

since for a matrix M

$$\begin{aligned} \text{trace}(M^T M) &= \sum\nolimits_i (M^T M)_{ii} \\ &= \sum\nolimits_i \sum\nolimits_j (M^T)_{ij} (M)_{ji} \\ &= \sum\nolimits_i \sum\nolimits_j m_{ij}^2 \\ &\geq 0. \end{aligned}$$

So we have a sum of non-negative terms being zero. So each term must be zero. That is, the square of each element of $P_i P_j$ must be zero. So each element of $P_i P_j$ is zero, so matrix $P_i P_j$ is zero:

$$P_i P_j = 0 \qquad (i \neq j).$$

This is the condition that the *linear forms* $P_1x, \ldots, P_kx$ be independent (Theorem 4.15 below). Since the P_ix are independent, so are the $(P_ix)^T(P_ix) = x^TP_i^TP_ix$, that is, x^TP_ix as P_i is symmetric and idempotent. That is, the *quadratic forms* $x^TP_1x, \ldots, x^TP_kx$ are also independent.

We now have

$$x^Tx = x^TP_1x + \ldots + x^TP_kx.$$

The left is $\sigma^2\chi^2(n)$; the ith term on the right is $\sigma^2\chi^2(n_i)$.

We summarise our conclusions.

Theorem 3.27 (Chi-Square Decomposition Theorem)

If

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

with each P_i a symmetric projection matrix with rank n_i, then

(i) the ranks sum:

$$n = n_1 + \ldots + n_k;$$

(ii) each quadratic form $Q_i := x^TP_ix$ is chi-square:

$$Q_i \sim \sigma^2\chi^2(n_i);$$

(iii) the Q_i are mutually independent.

(iv)

$$P_iP_j = 0 \quad (i \neq j).$$

Property (iv) above is called *orthogonality* of the projections P_i; we study orthogonal projections in §3.6 below.

This fundamental result gives all the distribution theory that we shall use. In particular, since F-distributions are defined in terms of distributions of independent chi-squares, it explains why we constantly encounter F-statistics, and why all the tests of hypotheses that we encounter will be F-tests. This is so throughout the Linear Model – Multiple Regression, as here, Analysis of Variance, Analysis of Covariance and more advanced topics.

Note 3.28

The result above generalises beyond our context of projections. With the projections P_i replaced by symmetric matrices A_i of rank n_i with sum I, the corresponding result (Cochran's Theorem) is that (i), (ii) and (iii) are *equivalent*. The proof is harder (one needs to work with *quadratic* forms, where we were able to work with *linear* forms). For monograph treatments, see e.g. Rao (1973), §1c.1 and 3b.4 and Kendall and Stuart (1977), §15.16 – 15.21.

3.6 Orthogonal Projections and Pythagoras's Theorem

The least-squares estimators (LSEs) are the *fitted values*

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

with P the projection matrix (idempotent, symmetric) above. In the alternative notation, since P takes the data y into $\hat{y}$, P is called the *hat matrix*, and written H instead. Then

$$e := y - \hat{y} = y - Py = (I - P)y$$

('e for error') is the *residual vector*. Thus

$$y = A\beta + \epsilon = A\hat{\beta} + e = \hat{y} + e,$$

or in words,

data = true value + error = fitted value + residual.

Now

$$\begin{aligned} e^T \hat{y} &= y^T (I - P)^T P y \\ &= y^T (I - P) P y \qquad (P \text{ symmetric}) \\ &= y^T (P - P^2) y \\ &= 0, \end{aligned}$$

as P is idempotent. This says that e, $\hat{y}$ are *orthogonal.* They are also both Gaussian (= multinormal, §4.3), as linear combinations of Gaussians are Gaussian (§4.3 again). For Gaussians, orthogonal = uncorrelated = independent (see § 4.3):

The residuals e and the fitted values $\hat{y}$ are independent

(see below for another proof). This result is of great practical importance, in the context of residual plots, to which we return later. It says that residual values e_i plotted against fitted values $\hat{y}_i$ should be *patternless.* If such a residual plot shows clear pattern on visual inspection, this suggests that our model may be wrong – see Chapter 7.

The data vector y is thus the hypotenuse of a right-angled triangle in n-dimensional space with other two sides the fitted values $\hat{y} = (I - P)y$ and the residual $e = Py$. The lengths of the vectors are thus related by Pythagoras's Theorem in n-space (Pythagoras of Croton, d. c497 BC):

$$\|y\|^2 = \|\hat{y}\|^2 + \|e\|^2.$$

In particular, $\|\hat{y}\|^2 \leq \|y\|^2$:

$$\|\hat{P}y\|^2 \leq \|y\|^2$$

for all y. We summarise this by saying that

$$\|P\| \leq 1$$

that is P has *norm* < 1, or P is *length-diminishing.* It is a projection from data-space (y-space) onto the vector subspace spanned by the least-squares estimates $\hat{\beta}$.

Similarly for $I - P$: as we have seen, it is also a projection, and by above, it too is length-diminishing. It projects from y-space onto the *orthogonal complement* of the vector subspace spanned by the LSEs.

For real vector spaces (as here), a projection P is *symmetric* ($P = P^T$) iff P is *length-diminishing* ($\|P\| \leq 1$) iff P is an *orthogonal*, or *perpendicular*, projection – the subspaces Im P and Ker P are orthogonal, or perpendicular, subspaces (see e.g. Halmos (1979), §75). Because our $P := AC^{-1}A^T$ ($C := A^TA$) is automatically symmetric and idempotent (a projection), this is the situation relevant to us.

Note 3.29

1. The use of the language, results and viewpoint of geometry – here in n dimensions – in statistics is ubiquitous in the Linear Model. It is very valuable, because it enables us to draw pictures and visualise, or 'see', results.

2. The situation in the Chi-Square Decomposition Theorem takes this further. There we have k (≥ 2) projections P_i summing to I, and satisfying the conditions
$$P_iP_j = 0 \qquad (i \neq j).$$
This says that the projections P_i are mutually *orthogonal*: if we perform two different projections, we reduce any vector to 0 (while if we perform the same projection *twice*, this is the same as doing it *once*). The P_i are *orthogonal projections*; they project onto *orthogonal subspaces*, L_i say, whose linear span is the whole space, L say:
$$L = L_1 \oplus \ldots \oplus L_k,$$
in the 'direct sum' notation $\oplus$ of Linear Algebra.

3. The case $k = 2$ is that treated above, with P, $I - P$ orthogonal projections and $L = L_1 \oplus L_2$, with $L_1 = \text{Im } P = \ker\ (I - P)$ and $L_2 = \text{Im } (I - P) = \ker\ P$.

Theorem 3.30

(i) $\hat{y} = Py \sim N(A\beta, \sigma^2 P)$.