Statistical Inference

S.D. Silvey

Late Professor of Statistics,
University of Glasgow

CHAPMAN & HALL/CRC

A CRC Press Company
Boca Raton London New York Washington, D.C.

3 The Method of Least Squares

3.1 Examples

The intuitive appeal of the method of least squares may be illustrated by two simple examples.

3.1.1 Suppose that $x_1, x_2, \ldots, x_n$ is a random sample from a distribution on the line and that we are interested in estimating the mean, θ, of this distribution. We may write

$$x_i = \theta + \varepsilon_i,$$

where ε_i is the deviation or error of the observation x_i from the mean of the underlying distribution. A reasonably optimistic attitude here is to assume that the observations are trying to give us information about θ and that the deviations $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are in some sense small. Therefore a plausible method of estimating θ is to choose as estimate a number for which these deviations are small, and one way of doing so is to choose as estimate a value of θ for which $\sum_{i=1}^{n} \varepsilon_i^2$ is as small as possible: that is, to estimate θ by the number $\hat{\theta}(x_1, x_2, \ldots, x_n)$ which minimizes $\sum_{i=1}^{n} (x_i - \theta)^2$, regarded as a function of θ.

If we do this, then of course

$$\hat{\theta}(x_1, x_2, \ldots, x_n) = \frac{1}{n}(x_1 + x_2 + \ldots, + x_n) = \bar{x},$$

the mean of the observations, so that our plausible method leads to an intuitively appealing result in this case.

3.1.2 As another example, suppose that observations $x_1, x_2, \ldots, x_n$ are made at different values $a_1, a_2, \ldots, a_n$, respectively, of a 'concomitant variable' a. The xs are random variables and the as are known, non-random numbers. For instance the a_is might be different levels of a fertilizer and the x_is corresponding yields of a crop. The as are then controlled by the observer but there are factors affecting the xs outwith the observer's control – weather for instance. Suppose we know that 'the mean of x varies linearly with a', but we do not know the exact form of this relationship. In other words we know that

$E(x_i) = \alpha + \beta a_i$, though α and β are unknown. Then we may write

$$x_i = \alpha + \beta a_i + \varepsilon_i,$$

and, acting on the same optimistic principle as previously, estimate α and β by these numbers which minimize

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (x_i - \alpha - \beta a_i)^2.$$

This method is again appealing provided that the deviations $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ all have 'the same chance of being small', but if some have more chance of being small than others it might seem more sensible to estimate α and β by minimizing some weighted sum of squares

$$\sum_{i=1}^{n} w_i (x_i - \alpha - \beta a_i)^2,$$

the ws being weights which are large for those is for which ε_i is liable to be small and small for ε_is liable to be large. Another complicating factor which might lead us to think again about this method of estimation is the possibility of interdependencies among the ε_is. It is not then so obvious how we might adjust the method. However a mathematical investigation of the properties of this method will suggest an appropriate adjustment.

3.2 Normal equations

These two examples are particular cases of the following general situation. A random vector $x = (x_1, x_2, \ldots, x_n)'$ is such that

$$x = A\beta + \varepsilon,$$

where A is a known matrix of order $n \times p$ with $p < n$, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of unknown parameters and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ is a vector of 'deviations from the mean' or 'errors', that is, a vector whose expected value is zero. In this general situation we may apply the principle used in the examples above and estimate β by minimizing the sum of squares

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (x - A\beta)'(x - A\beta).$$

This method of estimation is called the *method of least squares*, for obvious reasons, and any minimizing value $\hat{\beta}(x)$ is called a *least-square estimate* of β. The function $\hat{\beta}$, a function from $\mathbf{R}^n$ (Euclidean n-space) into $\mathbf{R}^p$, is a least-squares estimator. However we shall not maintain the notational distinction between $\hat{\beta}(x)$ and $\hat{\beta}$ and we shall use the latter for both. The context will make clear the sense in which it is used.

Determination of a least-squares estimate is not a difficult problem. We

have to choose $\hat{\beta}$ to minimize the quadratic form

$$(x-A\beta)'(x-A\beta)$$

in the components $\beta_1, \beta_2, \ldots, \beta_p$ of β. Differentiation of this quadratic form with respect to $\beta_1, \beta_2, \ldots, \beta_p$ leads to the so-called *normal equations* satisfied by a least-squares estimate, namely

$$A'A\beta = A'x,$$

and any solution of these equations does in fact minimize $(x-A\beta)'(x-A\beta)$ and so is a least-squares estimate (see Appendix A). If rank $A = p$, then $A'A$, which has the same rank as A (see Appendix A), is non-singular and there is a unique least-squares estimate

$$\hat{\beta} = (A'A)^{-1}A'x.$$

If rank $A < p$, then $A'A$ is singular, the normal equations do not have a unique solution, and there is a family of least-squares estimates which may be determined in any particular case by the usual methods for solving a system of linear equations.

3.3 Geometric interpretation

The intuitive appreciation of linear algebra is greatly aided by a geometrical interpretation in which vectors are represented by points and matrices are regarded as representations of linear transformations or functions (Hohn, 1964, p. 182). This applies equally to the *linear statistical model*, $x = A\beta+\varepsilon$, which we are investigating.

The sample space here is R^n and there is a true distribution on this space which we do not know. We do have some knowledge about the mean vector or *centre* θ of this distribution, for we know that $\theta = E(x)$ can be expressed in the form $A\beta$; in other words that θ lies in a subspace ω of R^n, the subspace spanned by the columns of A, which we shall refer to as the range of A. Given an observation x, we estimate θ by $\hat{\theta} = A\hat{\beta}$.

Now $x-\hat{\theta}$ is orthogonal to ω, since $A'(x-\hat{\theta}) = A'x-A'A\hat{\beta} = 0$, so that $x-\hat{\theta}$ is orthogonal to every vector of the form $A\beta$. This means that we estimate θ by the projection of x on ω, or by the point of ω nearest to x; and this seems reasonable on the grounds that x is probably near the centre of the distribution. Here we have the geometric essence of the method of least squares.

Of course $\hat{\theta}$, the projection of x on ω, is always unique whatever the rank of the matrix A. On the other hand, any point of R^p which is mapped by A into $\hat{\theta}$ is a least-squares estimate of β. If A has rank p it establishes a one-to-one correspondence between points in R^p and points in ω and then $\hat{\beta}$ is unique. In this case $\hat{\theta} = A(A'A)^{-1}A'x$. The matrix $A(A'A)^{-1}A'$ is symmetric and idempotent and it represents the orthogonal projection of R^n on to the range ω of A (see Appendix A). If rank $A < p$, then A establishes a many-to-one

correspondence between points β of R^p and points θ of ω, whose dimension is rank A, so that while $\hat{\theta}$ is still unique, $\hat{\beta}$ is not. There is not then a simple matrix representation for the projection of R^n on ω.

For the case $n = 3, p = 2$, we can actually draw pictures for this geometric interpretation.

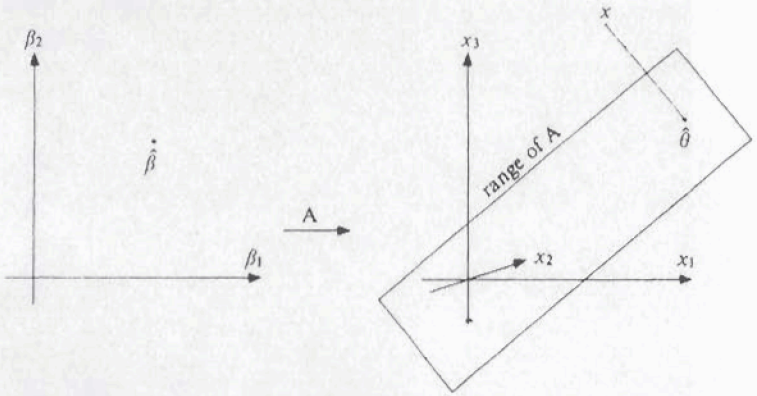

Figure 1 The matrix A of order 3×2 and rank 2 maps R^2 onto a two-dimensional subspace of R^3

Figure 1 gives a geometric picture of the case where the observation vector x has dimension 3, β has dimension 2 and the matrix A (which then has order 3×2) is of rank 2. A may be regarded as representing a linear transformation from R^2 into R^3, and, because its rank is 2, its range is represented by a plane in R^3 (see Hohn, 1964, p. 182). $\hat{\theta}$ is the projection of x on this plane, and there is a unique $\hat{\beta}$ mapped by A onto $\hat{\theta}$.

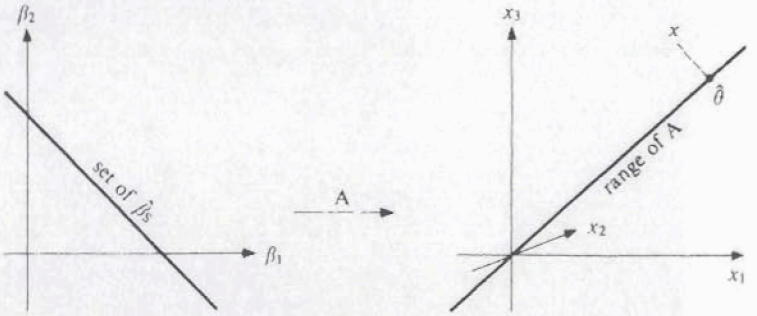

Figure 2 The matrix A of order 3×2 and rank 1 maps R^2 onto a one-dimensional subspace of R^3

Figure 2 illustrates the same case except that now, A has rank 1. Its range then is represented by a line through the origin. $\hat{\theta}$ is the projection of x on this line. But now there is not a unique β mapped by A onto $\hat{\theta}$, and the set of β such that $A\beta = \hat{\theta}$ is represented by a line in R^2 as indicated.

3.4 Identifiability

The case where rank A $< p$ in the linear model $x = A\beta + \varepsilon$ raises for the first time an issue which will be of concern later. If we are given a distribution for ε, the distribution on the sample space depends on β, as this distribution is centred on $A\beta$. However when rank A $< p$, different values of β yield the same distribution on the sample space because different values of β correspond to the same value of $A\beta$. It is clear that in this case, while an observation x may give us some information about $A\beta$, it can give no discriminatory information whatsoever between different values of β corresponding to the same value of $A\beta$. The parameter β is said to be unidentifiable. More generally, if different values of some parameter give the same distribution on the sample space, this parameter is not identifiable.

When a parameter is not identifiable we may say that two values of it are equivalent if and only if they yield the same distribution on the sample space. This defines an equivalence relation which partitions the parameter space into equivalence classes. Usually then an observation will give information about which equivalence class the true parameter belongs to but no information about which member of this equivalence class the true parameter is. This difficulty really arises because of our specification of the statistical model describing the situation in which observations are made, and it can be avoided by a different specification of the model. For instance, in the above linear model if rank A $= q < p$, then $p-q$ of the columns of A, say the last $p-q$, are linear combinations of the remaining q. It follows that, if $\mathbf{a}_i$ is the ith column of A, then

$$E(x) = \beta_1 \mathbf{a}_1 + \ldots + \beta_p \mathbf{a}_p$$

can be expressed in the form

$$E(x) = \gamma_1 \mathbf{a}_1 + \ldots + \gamma_q \mathbf{a}_q = A_q \gamma,$$

where A_q is the sub-matrix of A consisting of the first q columns of A. Now A_q has rank q and γ, a q-vector, is then identifiable. Had we specified the model in this way there would have been no identifiability problem; but the parameter β may have some significance in the practical situation which the statistical model is describing, whereas the parameter γ is not so easy to interpret. 'Natural' parametrization in the model set up may lead to non-identifiability, which is more in the nature of an irritant than a source of deep problems.

3.5 The Gauss–Markov theorem

The method of least squares has been introduced on the purely intuitive basis that if we estimate the mean of a distribution by the parameter nearest the observation made, this estimate should be quite good, since the observation is probably near the true mean. Some stronger justification of the method is really desirable and such a justification is provided by the celebrated theorem which we shall discuss in this section.

Whereas in chapter 2 we were dealing with real-valued parameters and estimates of these, we are concerned here with a vector-valued parameter β. Our criterion for choosing among unbiased estimates of a real parameter was that of minimum variance; in other words if $\hat{\theta}$ and $\tilde{\theta}$ were two unbiased estimates, $\hat{\theta}$ was regarded as better than $\tilde{\theta}$ if $\mathrm{var}_\theta(\tilde{\theta}) - \mathrm{var}_\theta(\hat{\theta})$ were greater than 0. What is the analogue of this criterion for a vector-valued parameter? It is that if $\hat{\beta}$ and $\tilde{\beta}$ are unbiased estimates of the vector β and their *variance matrices* are $\mathrm{var}_\beta(\hat{\beta}) = E_\beta(\hat{\beta}-\beta)(\hat{\beta}-\beta)'$ and $\mathrm{var}_\beta(\tilde{\beta})$, similarly defined, then $\hat{\beta}$ is a better estimate than $\tilde{\beta}$ if the *matrix* $\mathrm{var}_\beta(\tilde{\beta}) - \mathrm{var}_\beta(\hat{\beta})$ is positive semi-definite for all β. The dispersion of the random vector $\hat{\beta}$ about its mean β is then smaller than that of $\tilde{\beta}$. Another way of putting this is to say that the variance of any linear combination of the components of $\hat{\beta}$ is no larger than that of the same linear combination of the components of $\tilde{\beta}$; symbolically that, for every p-vector c,

$$\mathrm{var}_\beta(c'\hat{\beta}) \leqslant \mathrm{var}_\beta(c'\tilde{\beta}).$$

Now, in the identifiable case at least, the least-squares estimate $\hat{\beta}$ of β is a linear estimate in the sense that $\hat{\beta}_i(x)$ is a linear combination of the components of the observation x. It is also unbiased, since

$$E_\beta(\hat{\beta}) = E_\beta\{(A'A)^{-1}A'x\} = (A'A)^{-1}A'E_\beta(x) = (A'A)^{-1}A'A\beta = \beta.$$

The Gauss–Markov theorem proves that subject to certain conditions on the error-vector ε the least squares estimate $\hat{\beta}$ is better in the above sense than any other unbiased *linear* estimate. This of course is a weaker result than one which states that an estimator is best in the class of *all* unbiased estimates, but the input in the way of assumptions concerning the error vector is weak as we shall see, and as a general rule in the theory of inference, the weaker the assumptions, or the wider the family of possible distributions is allowed to be, the weaker are the results which can be obtained.

The examples in section 3.1 suggest that in order that the least-squares estimate be best, it may be necessary to require that the components of the error vector be independent and identically distributed. In fact the Gauss–Markov theorem requires less than this – only that these components have the same variance and are uncorrelated. Since the possibility of non-identifiability of β complicates matters considerably, we give first a proof for the case where rank A $= p$ so that β is identifiable and the least-squares estimate is unique and equal to $(A'A)^{-1}A'x$.

3.5.1 *The case where β is identifiable*

Let x be a random n-vector expressible in the form $x = A\beta + \varepsilon$ where A is a known $n \times p$ matrix of rank p, β is an unknown p-vector and ε is an error vector with $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \sigma^2 I$, where σ^2 is unknown, that is the components of ε have the same unknown variance σ^2 and are uncorrelated. Let $\hat{\beta}$ be the unique least-squares estimator of β and let $\phi = c'\beta$ be a linear parametric function. Then $c'\hat{\beta}$ is an unbiased estimator of ϕ and, if $\tilde{\phi}$ is any other linear unbiased estimator of ϕ, we have $\text{var}_\beta(c'\hat{\beta}) \leqslant \text{var}_\beta(\phi)$.

Proof. $E_\beta(c'\hat{\beta}) = c'E_\beta(\hat{\beta}) = c'\beta = \phi$ so that $c'\hat{\beta}$ is unbiased.

Since $\tilde{\phi}$ is a linear estimator it is expressible in the form $b'x$ and since $\tilde{\phi}$ is an unbiased estimator of ϕ, we have

$$b'A\beta = b'\{E_\beta(x)\} = E_\beta(b'x) = E_\beta(\phi) = c'\beta \quad \text{for every } \beta.$$

Hence $b'A = c'$.

Now $\text{var}_\beta(\tilde{\phi}) = \text{var}_\beta(b'x) = b'(\text{var}_\beta x)b = \sigma^2 b'b$,

and similarly

$$\text{var}_\beta(\hat{\beta}) = \text{var}_\beta\{(A'A)^{-1}A'x\} = (A'A)^{-1}A' \,\text{var}_\beta(x)A(A'A)^{-1} = \sigma^2(A'A)^{-1}.$$

It follows that $\text{var}_\beta(c'\hat{\beta}) = \sigma^2 c'(A'A)^{-1}c = \sigma^2 b'A(A'A)^{-1}A'b$.

To prove the theorem we must therefore show that

$$b'b \geqslant b'A(A'A)^{-1}A'b,$$

or that $I - A(A'A)^{-1}A'$ is positive semi-definite. This follows from the easily verifiable fact that this matrix is idempotent. (Incidentally it represents the orthogonal projection of R^n on to the orthogonal complement of the range of A and $b'\{I - A(A'A)^{-1}A'\}b$ is the square of the distance of the vector b from the range of A.)

This completes the proof.

3.5.2 *The general case*

If rank $A < p$, two complications arise in the above proof. The more obvious of these is that $A'A$ is then singular and we do not have the previous simple expression for a least-squares estimate $\hat{\beta}$. It remains true however that any least-squares estimate satisfies the equation

$$A'A\hat{\beta} = A'x.$$

We note, for subsequent use, that this implies that if a is a vector in the range of A, that is, a vector which can be expressed as a linear combination of the columns of A,

then $a'A\hat{\beta} = a'x$.

The second and less obvious complication is the fact that when rank $A < p$,

not every linear parametric function possesses an unbiased linear estimator, for in order that $b'x$ should be an unbiased estimator of $c'\beta$ we require (see Theorem 3.5.1) that $c' = b'A$ or that c' should be expressible as a linear combination of the rows of A. When rank $A = p$, the rows of A span R^p and every p-vector c' can be expressed as a linear combination of these rows. This is not so when rank $A < p$. In considering the general case therefore, we must restrict attention to linear parametric functions $c'\beta$ which have unbiased linear estimators or are *estimable*.

Suppose then that $\phi = c'\beta$ is an estimable parametric function. We wish to show that, if $\hat{\beta}$ is any least-squares estimate of β, $c'\hat{\beta}$ is an unbiased estimator of ϕ and that

$$\text{var}_\beta(c'\hat{\beta}) < \text{var}^\beta(\tilde{\phi})$$

for any other unbiased linear estimator $\tilde{\phi}$ of ϕ; this being subject to the conditions of Theorem 3.5.1 apart from that on rank A. Now since ϕ is estimable there exists a $b \in R^n$ such that $c' = b'A$. Let a be the projection of b on the range of A so that $b - a$ is orthogonal to the range of A, or $(b-a)'A = 0$. Then $a'x$ also is an unbiased estimator of ϕ since

$$E_\beta(a'x) = E_\beta\{(a-b)'x + b'x\} = (a-b)'A\beta + \phi = \phi.$$

Moreover a is the *only* vector in range A such that $a'x$ is an unbiased estimator of ϕ. For suppose there is another such vector a^*. Then we have for every β,

$$E_\beta\{(a-a^*)'x\} = 0$$
i.e. $$(a-a^*)'A\beta = 0$$
and so $$(a-a^*)'A = 0.$$

This means that $a - a^*$ is orthogonal to range A, but since a and a^* are both in range A, so is $a - a^*$. Therefore $a - a^* = 0$, i.e. $a = a^*$.

Now $\text{var}_\beta(a'x) \leqslant \text{var}_\beta(b'x)$, where $b'x$ is any other unbiased estimator of ϕ,

since $\text{var}_\beta(a'x) = \sigma^2 a'a$,
while $\text{var}_\beta(b'x) = \sigma^2 b'b = \sigma^2\{a'a + (b-a)'(b-a)\} \geqslant \sigma^2 a'a$.

We now complete the proof of the fact that

$$\text{var}_\beta(c'\hat{\beta}) \leqslant \text{var}_\beta(\tilde{\phi})$$

by showing that $\text{var}_\beta(c'\hat{\beta}) = \text{var}_\beta(a'x)$.
Since $a'x$ is an unbiased estimator of ϕ, we have $a'A = c'$.

Therefore $c'\hat{\beta} = a'A\hat{\beta} = a'x$,

since a is in range A and $\hat{\beta}$ satisfies the normal equations. It follows that

$$\text{var}_\beta(c'\hat{\beta}) = \sigma^2 a'a.$$

It remains true in the general case, that any least-squares estimate is better than any other unbiased linear estimator in this special sense that it leads to

unbiased estimators of estimable linear functions with smaller variance.

3.5.3 *Remarks*

There are many statements and proofs of this celebrated theorem with various degrees of generality. The proof given in section 3.5.2, which incidentally is valid also when rank A = p, is essentially the same as that given by Scheffé (1960).

It will be seen that in the general proof (contrary to the proof of section 3.5.1) we have never explicitly stated that any least-squares estimator $\hat{\beta}$ of β is unbiased. The reason for this is as follows. We may regard any component of β, say its first component β_1, as a linear parametric function – the function $c'\beta$ where $c' = (1, 0, 0, \ldots, 0)$. It may be that this is not estimable. In the unlikely event that the first column of A consists entirely of zeros, it is immediately obvious that no observation gives any information about β_1 and that β_1 is not estimable. So in this case, as there exists no unbiased estimator of β_1, *a fortiori* there exists none of β. However it is true that if a component β_i of β is estimable, the corresponding component of any least-squares estimate of β is an unbiased estimator of β_i – this follows from the fact established in section 3.5.2 that $c'\hat{\beta} = a'x$, in the notation of that section.

3.6 Weighted least squares

In the last paragraph of section 3.1 we anticipated the possibility that least-squares estimators might not be 'best' when the components of the error vector ε did not all 'have the same chance of being small', and indeed a crucial part is played in the above proofs by the assumption that the variances matrix of the error vector is $\sigma^2 I$, that is, that its components have the same variance and are uncorrelated. The algebra of the Gauss–Markov theorem suggests the appropriate modification to the method of least squares when either the errors have different variances or when they are correlated.

Suppose then that we consider the linear model

$$x = A\beta + \varepsilon$$

with the same assumptions as before except that now, instead of having var $\varepsilon = \sigma^2 I$, we assume that var $\varepsilon = \sigma^2 \Sigma$, where Σ is a *known* positive definite matrix. This allows for the possibility of differing variances among the ε_is and for correlation between them. By a non-singular linear transformation we can transform this model to that previously investigated by the Gauss–Markov theorem. For since Σ is positive definite, it can be expressed in the form PP' where P is non-singular. Now let $\eta = P^{-1}\varepsilon$ and $y = P^{-1}x$. Then we may write the model in the form

$$Py = A\beta + P\eta$$

or $$y = P^{-1}A\beta + \eta = B\beta + \eta, \quad \text{say.}$$

Also var $\eta = P^{-1}(\text{var}\,\varepsilon)P'^{-1} = \sigma^2 P^{-1}PP'P'^{-1} = \sigma^2 I$, so that in terms of y, B and η the model is just that already discussed. We therefore obtain a 'best' estimator of β (that is, an unbiased estimator with minimum dispersion, in the sense of the Gauss–Markov theorem, among the class of linear unbiased estimators) by minimizing the sum of squares

$$(y - B\beta)'(y - B\beta).$$

Now
$$\begin{aligned}(y - B\beta)'(y - B\beta) &= (x - A\beta)(PP')^{-1}(x - A\beta)\\ &= (x - A\beta)\Sigma^{-1}(x - A\beta).\end{aligned}$$

Hence in the case where the error components have variance matrix $\sigma^2\Sigma$, we obtain best estimators by minimizing this quadratic form rather than the straight sum of squares. In particular, suppose that the components of the error vector are uncorrelated, but have unequal variances so that

$$\Sigma = \text{diag}\,\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\}, \quad \text{say.}$$

The expression which we minimize may be written

$$\sum_{i=1}^{n} \frac{(x_i - a_{i1}\beta_1 - a_{i2}\beta_2 - \ldots - a_{ip}\beta_p)^2}{\sigma_i^2}.$$

In other words we weight each square in the sum by the inverse of the variance of the corresponding error component. In this way, as anticipated, we give more weight to the errors which are liable to be small.

It is worth remarking that even when var ε cannot be expressed in the form $\sigma^2 I$, but has the more general form $\sigma^2\Sigma$, an estimator obtained by minimizing the straight sum of squares $(x - A\beta)'(x - A\beta)$ is still unbiased. For instance, in the case where A has rank p, this estimator $\hat{\beta}$ is given by

$$\hat{\beta} = (A'A)^{-1}A'x$$

and since $E_\beta(x) = A\beta$ we still have $E_\beta(\hat{\beta}) = \beta$. However, in this case

$$\text{var}\,\hat{\beta} = \sigma^2(A'A)^{-1}(A'\Sigma A)(A'A)^{-1},$$

whereas if $\tilde{\beta}$ is the estimator minimizing

$$(x - A\beta)'\Sigma^{-1}(x - A\beta) = (y - B\beta)'(y - B\beta),$$

we have

$$\text{var}\,\tilde{\beta} = \sigma^2(B'B)^{-1} = \sigma^2(A'\Sigma^{-1}A)^{-1}.$$

The Gauss–Markov theorem tells us that the matrix

$$(A'A)^{-1}(A'\Sigma A)(A'A)^{-1} - (A'\Sigma^{-1}A)^{-1}$$

is positive semi-definite, so that, in particular, the diagonal elements of this matrix are all non-negative. This means that the variance of any component of $\hat{\beta}$ is at least as large as that of the corresponding component of $\tilde{\beta}$, and it may

well be considerably larger. So while straight least squares yields unbiased estimators of the components of β in this situation, it may be very inefficient in the sense that the variances of these estimators are unnecessarily large, relative to the best we can achieve using linear estimators.

3.7 Estimation of σ^2

In practice when the model

$$x = A\beta + \varepsilon$$

with $\text{var}\,\varepsilon = \sigma^2 I_n$, is appropriate for describing some observational situation, not only will β be unknown, but so also will σ^2. We are then faced with the problem of estimating this unknown quantity as well as β. If $\hat{\beta}$ is any least-squares estimate of β then we might expect that the *residual sum of squares*

$$(x - A\hat{\beta})'(x - A\hat{\beta})$$

will on average increase and decrease with σ^2. And indeed it is not difficult to construct an unbiased estimator of σ^2 from this residual sum of squares. We proceed as follows.

We have $\varepsilon = (x - A\hat{\beta}) + A(\hat{\beta} - \beta)$.

Now $A(\hat{\beta} - \beta)$ is a vector in range A, while $(x - A\hat{\beta})$ is orthogonal to range A (since $A'(x - A\hat{\beta}) = 0$). It follows that if we change the basis in R^n to a new orthonormal basis whose first r elements are in range A (r = rank A) and whose remaining $n - r$ elements are orthogonal to range A and if, under this transformation,

$$(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \to (\eta_1, \eta_2, \ldots, \eta_n),$$

then $A(\hat{\beta} - \beta) \to (\eta_1, \eta_2, \ldots, \eta_r, 0, 0, \ldots, 0)'$

and $(x - A\hat{\beta}) \to (0, 0, \ldots, 0, \eta_{r+1}, \eta_{r+2}, \ldots, \eta_n)'$.

Hence $(x - A\hat{\beta})'(x - A\hat{\beta}) = \eta_{r+1}^2 + \eta_{r+2}^2 + \ldots + \eta_n^2$.

If uncorrelated random variables with zero means and common variance σ^2 are subjected to an orthogonal transformation, the resulting random variables have the same properties. ($\eta = P\varepsilon$, where $PP' = I_n$; $E(\eta) = PE(\varepsilon) = 0$ and $\text{var}\,\eta = P\,\text{var}\,\varepsilon P' = \sigma^2 PP' = \sigma^2 I_n$.)

So $E(\eta_i^2) = \sigma^2$, for all i.

Hence $E_\beta\{(x - A\hat{\beta})'(x - A\hat{\beta})\} = E\left(\sum_{i=r+1}^{n} \eta_i^2\right) = (n - r)\sigma^2$,

so that $\dfrac{1}{n-r}(x - A\hat{\beta})'(x - A\hat{\beta})$

is an unbiased estimator of σ^2.

The modification required when $\text{var}\,\varepsilon = \sigma^2\Sigma$, with Σ known, instead of $\sigma^2 I_n$, is clear. In the notation of section 3.6,

$$\frac{1}{n-r}(y - B\tilde{\beta})'(y - B\tilde{\beta}) = \frac{1}{n-r}(x - A\tilde{\beta})'\,\Sigma^{-1}\,(x - A\tilde{\beta})$$

is then the corresponding unbiased estimator of σ^2.

3.8 Variance of least-squares estimators

Another practical consideration which we must take into account is as follows. There is little point in practice in knowing that an estimator of an unknown parameter is best in some sense without the additional knowledge of how near to the parameter our estimate is liable to be. We shall consider this question in general later and in the meantime we content ourselves with the remark that the variance of an estimator gives some idea of its reliability or accuracy. Thus when an estimate is given in practice it is usual to quote also its standard deviation or an estimate of its standard deviation.

Suppose then that we consider the linear model

$$x = A\beta + \varepsilon,$$

where $\text{var}\,\varepsilon = \sigma^2 I_n$ and A has full rank ($= p$) so that no identifiability problems arise. Then the unique least-squares estimate $\hat{\beta}$ of β is given by

$$\hat{\beta} = (A'A)^{-1}A'x.$$

We can deduce immediately that

$$\begin{aligned}\text{var}\,\hat{\beta} &= (A'A)^{-1}A'\,\text{var}\,x\,A(A'A)^{-1} \\ &= \sigma^2(A'A)^{-1}.\end{aligned}$$

Now we can calculate an unbiased estimator of σ^2 by the previous section, namely,

$$\hat{\sigma}^2 = \frac{1}{n-p}(x - A\hat{\beta})'(x - A\hat{\beta}),$$

and it follows that $\hat{\sigma}^2(A'A)^{-1}$ is an unbiased estimator of the variance matrix of $\hat{\beta}$. If we are interested in estimating a linear parametric function $c'\beta$, say, then $c'\hat{\beta}$ is a minimum variance unbiased linear estimator of this and

$$\text{var}_\beta\, c'\hat{\beta} = \sigma^2 c'(A'A)^{-1}c.$$

Hence an estimate of the standard deviation of our estimator $c'\hat{\beta}$ is

$$\hat{\sigma}\sqrt{\{c'(A'A)^{-1}c\}},$$

a number which can be calculated from the given observations.

In particular, if we wish to estimate a particular component β_i of β, this is estimated by $\hat{\beta}_i$ and

$$\text{var } \hat{\beta}_i = \sigma^2 \times (i,i)\text{th element of } (A'A)^{-1}$$

while estimated S.D. $\hat{\beta}_i = \hat{\sigma}\sqrt{\{(i,i)\text{th element of } (A'A)^{-1}\}}$.

3.9 Normal theory

As a general rule in statistical theory, the more we are prepared to assume about the probabilistic model underlying observations, the stronger the results we can prove regarding estimators. In the preceding sections of this chapter we have made assumptions about the first and second moments of the error vector ε, but no further assumptions about the form of its distribution. Then we were able to demonstrate that least-squares estimators were best in the class of unbiased *linear* estimators. Suppose that we add the assumption that ε has a normal distribution, so that our model now becomes

$$x = A\beta + \varepsilon,$$

where ε is $N(0, \sigma^2 I_n)$; A is known, and β and σ^2 are unknown. Can we now prove something stronger about least-squares estimators? The answer is yes, and we appeal to the Rao–Blackwell theorem to demonstrate this.

With the additional assumption of normality of errors, we have

$$p(x;\beta,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}n}\exp\left[-\frac{1}{2\sigma^2}(x-A\beta)'(x-A\beta)\right]$$

$$= C(\beta,\sigma^2)\exp\left[-\frac{x'x}{2\sigma^2}+\sum_{i=1}^{p}\frac{\beta_i}{\sigma^2}y_i\right],$$

where $y = A'x$.

Now write $t_i(x) = y_i \quad (i = 1, 2, \ldots, p)$,

$t_{p+1}(x) = x'x$,

and $t(x) = \{t_1(x), t_2(x), \ldots, t_{p+1}(x)\}$.

Also, reparametrize in terms of $\theta = (\theta_1, \theta_2, \ldots, \theta_{p+1})$,

where $\theta_i = \dfrac{\beta_i}{\sigma^2} \quad (i = 1, 2, \ldots, p)$,

and $\theta_{p+1} = -\dfrac{1}{2\sigma^2}$.

Then $p(x;\beta,\sigma^2)$ can be expressed in the form

$$C^*(\theta)\exp\left[\sum_{i=1}^{p+1}\theta_i t_i(x)\right].$$

It follows from the factorization theorem that $t(x)$ is sufficient for θ and from Theorem 2.5.4 on exponential families that the family of distributions of t is complete, if there are no prior restrictions on β and σ^2, because then the parameter space contains a $(p+1)$-dimensional rectangle.

Now β_i is a real function of θ, $\beta_i = -\frac{1}{2}\theta_i/\theta_{p+1}$, and $\hat{\beta}_i$ is a function of the sufficient statistic t, since

$$\hat{\beta} = (A'A)^{-1}A'x = (A'A)^{-1}y.$$

Moreover $\hat{\beta}_i$ is an unbiased estimator of β_i. (We assume here that A has full rank so that β_i is estimable.) It follows that $\hat{\beta}_i$ has minimum variance in the class of *all* unbiased estimators of $\beta_i (i = 1, 2, \ldots, p)$, by the argument of section 2.6. Incidentally, if $s^2 = (x-A\hat{\beta})'(x-A\hat{\beta})$, then $s^2/(n-p)$ is a minimum-variance unbiased estimator of σ^2 in this case.

Thus by adding the assumption of normality to the linear model we are able to establish that least-squares estimates are optimal in a stronger sense than they are without this assumption.

3.9.1 *Note*

It is convenient at this stage to prove, for subsequent use, a result concerning the distributions of $\hat{\beta}$ and s^2 when the assumption of normality of the error ε is added to the linear model assumptions previously adopted.

Since $\hat{\beta}$ is linearly related to a normal random vector ($\hat{\beta} = (A'A)^{-1}A'x$, where x is $N(A\beta, \sigma^2 I)$) we can state immediately that $\hat{\beta}$ itself is

$$N\{\beta, \sigma^2(A'A)^{-1}\}.$$

Furthermore we have seen in section 3.4 that there exists an orthogonal matrix P such that if $\eta = P\varepsilon$,

then $PA(\hat{\beta}-\beta) = (\eta_1, \eta_2, \ldots, \eta_p, 0, 0, \ldots, 0)'$

and $P(x-A\hat{\beta}) = (0, 0, \ldots, 0, \eta_{p+1}, \ldots, \eta_n)'$:

Now since the components of ε are independent $N(0, \sigma^2)$ random variables and since P is orthogonal, it follows that $\eta_1, \eta_2, \ldots, \eta_n$ are also independent $N(0, \sigma^2)$. Therefore $\hat{\beta}-\beta$ and $x-A\hat{\beta}$ are independent. Furthermore

$$(x-A\hat{\beta})'(x-A\hat{\beta}) = \sum_{i=p+1}^{n}\eta_i^2$$

and so $s^2 = (x-A\hat{\beta})'(x-A\hat{\beta})$ is distributed as $\sigma^2\chi^2(n-p)$.

In other words, *with the normal assumption the least-squares estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ of $\beta_1, \beta_2, \ldots, \beta_p$ respectively are jointly normally distributed and are independent of the residual sum of squares s^2 which is distributed as $\sigma^2\chi^2(n-p)$.*

3.10 Least squares with side conditions

Until now in this chapter we have considered the linear model with $E(x)$

expressed in the form $A\beta$. Sometimes the natural expression of the model in terms of the parameters of interest does not occur in this way. In particular these parameters may be mathematically related to one another and often the relationships between them are linear. In the latter case we have a model in which $E(x)$ is expressed in the form $A\beta$ with β_is satisfying certain side conditions, say $H\beta = 0$, where H is a $q \times p$ matrix ($q < p$) of known coefficients. It must be emphasized straightway that from a theoretical point of view this new model is not different in essence from that which we have been discussing. In both cases we are stating that $E(x)$ belongs to a subspace of R^n, and indeed by reparametrization we can throw the new model into the form of the previous one. Suppose, for instance, that rank $H = q$. (If rank $H < q$ this simply means that some of the conditions are redundant, being consequences of the rest, and we may simply discard these.) By adjoining $p-q$ suitable chosen rows to the matrix H we can construct a non-singular $p \times p$ matrix K with $K' = (H', H^{*\prime})$, say. Now let $\gamma = K\beta$. The side conditions on β are, in terms of γ,

$$\gamma_1 = \gamma_2 = \ldots = \gamma_q = 0.$$

Now we have

$$E(x) = AK^{-1}\gamma = B\gamma, \quad \text{say,}$$
$$= (B_1\, B_2)\begin{bmatrix}\gamma^{(1)}\\ \gamma^{(2)}\end{bmatrix},$$

where $\gamma^{(1)} = (\gamma_1, \gamma_2, \ldots, \gamma_q)'$ and $\gamma^{(2)} = (\gamma_{q+1}, \ldots, \gamma_p)'$.
Thus $E(x) = B_1\gamma^{(1)} + B_2\gamma^{(2)} = B_2\gamma^{(2)}$, when $\gamma^{(1)} = 0$. In this expression for $E(x)$, the side conditions are incorporated, there are no conditions on $\gamma^{(2)}$ and the model is as previously.

In practice, while it would be possible to treat the problem of least-squares estimation with side conditions in the way just described, to determine $\hat{\gamma}$ and then $\hat{\beta} = K^{-1}\hat{\gamma}$, this would be an unnatural approach. The problem we have is that of minimizing the sum of squares

$$(x - A\beta)'(x - A\beta)$$

subject to the side conditions $H\beta = 0$, since it is natural to require that our least-squares estimates should satisfy the conditions which we know to be satisfied by the parameters being estimated. The obvious way of going about this is to introduce Lagrange multipliers and derive the following equations satisfied by the restricted least-squares estimate $\hat{\beta}$. In these equations, λ is a q-vector of Lagrange multipliers, $\lambda_1, \lambda_2, \ldots, \lambda_q$.

$$A'A\beta + H'\lambda = A'x,$$
$$H\beta = 0,$$

or $$\begin{bmatrix}A'A & H'\\ H & 0\end{bmatrix}\begin{bmatrix}\beta\\ \lambda\end{bmatrix} = \begin{bmatrix}A'x\\ 0\end{bmatrix}.$$

The vector λ here obviously depends in general on x and so we may regard it as a random vector.

We now consider the question of what we can say about the distribution of the restricted least-squares estimate $\hat{\beta}$ when we retain the standard assumptions regarding the distribution of the error vector ε, namely that $E(\varepsilon) = 0$ and $\text{var}\,\varepsilon = \sigma^2 I$, and when $E(x) = A\beta$ where $H\beta = 0$. This question is of interest *per se* and it is also relevant in a problem which will concern us later.

3.10.1 rank $A = p$, rank $H = q$

The first case which we shall discuss is that in which there are no identifiability difficulties regarding β (rank $A = p$) and in which no restrictions are redundant (rank $H = q$). In this case $A'A$ is positive definite, and the matrix

$$\begin{bmatrix}A'A & H'\\ H & 0\end{bmatrix}$$

is non-singular (see Appendix A). Moreover, if its inverse, similarly partitioned, is $\begin{bmatrix}P & Q\\ Q' & R\end{bmatrix}$

then $$\begin{bmatrix}\hat{\beta}\\ \lambda\end{bmatrix} = \begin{bmatrix}PA'x\\ Q'A'x\end{bmatrix}.$$

Also $$E_\beta\begin{bmatrix}\hat{\beta}\\ \lambda\end{bmatrix} = \begin{bmatrix}PA'A\beta\\ Q'A'A\beta\end{bmatrix} = \begin{bmatrix}\beta\\ 0\end{bmatrix},$$

since $PA'A + QH = I$
and $Q'A'A + RH = 0$,

so that $PA'A\beta = \beta - QH\beta = \beta$,
and $Q'A'A\beta = -RH\beta = 0$,

as $H\beta = 0$.

Furthermore $$\text{var}_\beta\begin{bmatrix}\hat{\beta}\\ \lambda\end{bmatrix} = \sigma^2\begin{bmatrix}PA'AP & PA'AQ\\ Q'A'AP & Q'A'AQ\end{bmatrix}$$
$$= \sigma^2\begin{bmatrix}P & 0\\ 0 & -R\end{bmatrix},$$

since in addition to the previous matrix equations we have $HP = 0$ and $HQ = I$, and therefore $PA'AP = QHP = P$, etc. (see Appendix A).

To sum up: With the linear model

$$x = A\beta + \varepsilon,$$

where $H\beta = 0$, $E(\varepsilon) = 0$ and $\text{var}\,\varepsilon = \sigma^2 I$, rank $A = p$ and rank $H = q$, the restricted least-squares estimate $\hat{\beta}$ has mean β and variance matrix P, the

leading $p \times p$ sub-matrix of the inverse of $\begin{bmatrix} A'A & H' \\ H & 0 \end{bmatrix}$.

It is clear from the argument at the beginning of section 3.10 which reparametrizes in terms of γ that

$$\frac{1}{n-p+q}(x-A\hat{\beta})'(x-A\hat{\beta})$$

is in this case an unbiased estimator of σ^2.

3.10.2 rank A $< p$, rank H $= q$

As we have seen, when rank A $< p$ the parameter β is not identifiable and the domain of β is partitioned into equivalence classes of parameters. Any two parameters in the same equivalence class yield the same value for $E(x)$, and we cannot hope to distinguish between these as a result of observation. Indeed in this linear situation it is not difficult to identify these equivalence classes. All parameters β in the null space of A, that is all βs such that $A\beta = 0$ are in the same equivalence class and this is a linear subspace of R^p. Any equivalence class is a 'hyperplane parallel to this subspace'.

One method of proceeding in this case is to introduce restrictions on β in order to focus attention on exactly one member of each equivalence class and to behave as if the true parameter satisfied these restrictions. To take a trivial illustration, suppose that our model specifies that for $i = 1, 2, \ldots, n$, $E(x_i) = \beta_1 + \beta_2$,

i.e., that $E(x) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$.

All parameters β such that $\beta_1 + \beta_2$ has a given value k, say, are equivalent. If now we impose the restriction that $\beta_1 = \beta_2$, this restriction serves to pick out exactly one member of each equivalence class – the member $\begin{bmatrix} \frac{1}{2}k \\ \frac{1}{2}k \end{bmatrix}$ of the equivalence class defined by $\beta_1 + \beta_2 = k$. Then we may proceed to estimate β as if it satisfied the restriction $\beta_1 = \beta_2$, and so estimate the equivalence class to which it belongs.

In general when A has order $n \times p$ and rank $r < p$ it is possible to introduce $p-r$ linear restrictions which serve, as in the illustration, to identify a particular member of each equivalence class. More specifically, there exists a $(p-r) \times p$ matrix L of rank $(p-r)$ such that the equations

$A\beta = k$,
$L\beta = 0$

have exactly one solution. An obvious necessary and sufficient condition for these equations to have a unique solution is that

$$\text{rank}\begin{bmatrix} A \\ L \end{bmatrix} = p,$$

and again obviously we can find many matrices L of order $(p-r) \times p$ which satisfy this condition. Since the conditions $L\beta = 0$ serve to identify β we shall refer to them as identifiability restraints.

It often happens in practice that the natural or symmetric specification of $E(x)$ in a linear model takes the following form:

$E(x) = A\beta$ where $H\beta = 0$; rank A $= r < p$ and rank H $= q$.

Moreover some of the side conditions $H\beta = 0$ serve to identify β and the remainder are 'genuine' restrictions on β. In fact there exists a sub-matrix of H, H_1 say, of order $(p-r) \times p$ such that $\begin{bmatrix} A \\ H_1 \end{bmatrix}$ has rank p. Again in theory this specification of the linear model presents no essentially new difficulty for by a reparametrization we can clearly revert to the original form of the model which we have discussed in detail. However we now consider the practical algebra corresponding to this specification.

We may suppose without any loss of generality that H may be partitioned into $\begin{bmatrix} H_1 \\ H_2 \end{bmatrix}$,

where H_1 has $p-r$ rows and the equations $H_1\beta = 0$ are identifiability constraints, so that $\begin{bmatrix} A \\ H_1 \end{bmatrix}$ has rank p, and $A'A + H_1'H_1$ is a $p \times p$ positive definite matrix. As before, the restricted least-squares estimate $\hat{\beta}$ satisfies the equations

$$\begin{bmatrix} A'A & H' \\ H & 0 \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \lambda \end{bmatrix} = \begin{bmatrix} A'x \\ 0 \end{bmatrix}.$$

But now A'A is singular and some modification of the argument of section 3.8.1 is necessary. This modification is relatively simple. For since $H\hat{\beta} = 0$, so that in particular $H_1\hat{\beta} = 0$, an equivalent set of equations is

$$\begin{bmatrix} A'A + H_1'H_1 & H' \\ H & 0 \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \lambda \end{bmatrix} = \begin{bmatrix} A'x \\ 0 \end{bmatrix},$$

and the matrix $A'A + H_1'H_1$ is positive definite, so that we now have a set of equations similar in structure to those of section 3.8.1; and the matrix on the left hand side is non-singular. If now we set

$$\begin{bmatrix} A'A + H_1'H_1 & H' \\ H & 0 \end{bmatrix}^{-1} = \begin{bmatrix} P & Q \\ Q' & R \end{bmatrix}$$

then we have, as before

$$\hat{\beta} = PA'x$$
and $$\lambda = Q'A'x.$$

The matrix relationships used to establish the results of section 3.8.1 were

$$PA'A+QH = I \qquad 3.1$$
$$Q'A'A+RH = 0 \qquad 3.2$$
$$HP = 0 \qquad 3.3$$
$$HQ = I \qquad 3.4$$

and these are now replaced by the following

$$P(A'A+H_1'H_1)+QH = I \qquad 3.1a$$
$$Q'(A'A+H_1'H_1)+RH = 0 \qquad 3.2a$$
$$HP = 0 \qquad 3.3a$$
$$HQ = I \qquad 3.4a$$

From equation **3.3a** we have in particular the fact that $PH_1' = 0$, so equation **3.1a** is equivalent to equation **3.1** and the only essential difference between the second and first set of equations is the term $Q'H_1'H_1$ in equation **3.2a**. As is easily verified the only difference that this makes to the deductions of section 3.8.1 is that now

$$\text{var}\,\lambda = -R-Q'H_1'H_1Q;$$

everything else remains unaltered. Since $HQ = I$ and $p-r < q$ it follows that

$$(H_1Q)'(H_1Q) = \begin{bmatrix} I_{p-r} & 0 \\ 0 & 0 \end{bmatrix},$$

so the only adjustments required by the non-identifiability of β are that

(a) we replace $A'A$ by $A'A+H_1'H_1$,

(b) var λ becomes $-R-\begin{bmatrix} I_{p-r} & 0 \\ 0 & 0 \end{bmatrix}$ instead of, as previously $-R$.

While we are not particularly interested in the random variable λ in the meantime, these results concerning λ which emerge here in a natural way, will be used in a subsequent problem.

.11 Discussion

There are many variations on the least-squares theme and there are various questions which we have left unanswered. The method is an extremely useful one and it is often applied even when the assumptions of the Gauss–Markov theorem, which justifies it in terms of minimum-variance unbiasedness, are not satisfied. It therefore becomes natural to inquire how the properties of the method are affected by changes in the assumptions regarding the error vector

etc. Much of econometrics is concerned with this kind of question and for a very full discussion of such points the reader is referred to Malinvaud (1966).

Examples

3.1 Assume that observations $x_1, x_2, \ldots, x_n$ can be expressed in the form

$$x_i = \beta_0 + \beta_1 a_i + \varepsilon_i \quad (i = 1, 2, \ldots, n),$$

where $a_1, a_2, \ldots, a_n$ are known values of a concomitant variable and the εs are uncorrelated errors with common variance σ^2. Verify that β_0 and β_1 are both estimable if and only if the a_is are not all equal and confirm the intuitive acceptability of this result by imagining a scatter diagram of the points (a_i, x_i). Show that, when the a_is are not all equal, least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum (a_i-\bar{a})x_i}{\sum (a_i-\bar{a})^2},$$

$$\hat{\beta}_0 = \bar{x}-\bar{a}\hat{\beta}_1.$$

Prove directly from the first expression that $\text{var}\,\hat{\beta}_1 = \sigma^2/\sum (a_i-\bar{a})^2$. Show that $\bar{x}$ and $\hat{\beta}_1$ have zero covariance and deduce that

(a) $\text{cov}\,(\hat{\beta}_0, \hat{\beta}_1) = -\bar{a}\,\text{var}\,\hat{\beta}_1$;

(b) $$\text{var}\,\hat{\beta}_0 = \sigma^2\left[\frac{1}{n}+\frac{\bar{a}^2}{\sum (a_i-\bar{a})^2}\right].$$

Verify these results by writing the model in the matrix notation

$$x = A\beta+\varepsilon$$

and using the general results of chapter 3.

3.2 Observations $x_1, x_2, \ldots, x_n$ can be expressed in the form

$$x_i = \beta_0+\beta_1 a_i+\beta_2 a_i^2+\varepsilon_i \quad (i = 1, 2, \ldots, n),$$

where the a_is are values of a concomitant variable and the εs are uncorrelated errors with common variance. Establish that $\beta = (\beta_0, \beta_1, \beta_2)$ is identifiable if and only if there are at least three different values among $a_1, a_2, \ldots, a_n$.

3.3 The model $x_i = \beta_0+\beta_1 a_i+\varepsilon_i \quad (i = 1, 2, \ldots, n)$,

may be expressed in the form

$$x_i = \beta_0+\beta_1\bar{a}+\beta_1(a_i-\bar{a})+\varepsilon_i$$
$$= \alpha+\beta_1(a_i-\bar{a})+\varepsilon_i, \quad \text{say.}$$

Show that this reparametrization in terms of α and β_1 rather than β_0 and β_1 facilitates calculation of least-squares estimates.

Verify that, in general, the model

$$x = A\beta + \varepsilon$$

can always, by reparametrization, be expressed in the form

$$x = B\gamma + \varepsilon,$$

where B is a matrix whose columns are orthogonal, and that the least-squares estimate of γ is easily calculated.

3.4 Observations $x_{ij} (i = 1, 2, \ldots, r; j = 1, 2, \ldots, n)$, are such that

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

where the εs are uncorrelated errors with a common variance. Verify that $\mu, \tau_1, \tau_2, \ldots, \tau_r$ are not identifiable, but that they are when the restriction $\tau_1 + \tau_2 + \ldots + \tau_r = 0$ is imposed. Show that the least-squares estimates, subject to this restriction are

$$\hat{\mu} = x_{..} = \frac{1}{rn} \sum_{i,j} x_{ij},$$

$$\hat{\tau}_i = x_{i.} - x_{..},$$

where $x_{i.} = \frac{1}{n} \sum_j x_{ij} \quad (i = 1, 2, \ldots, r)$.

3.5 Aerial observations x_1, x_2, x_3, x_4 are made of the angles $\theta_1, \theta_2, \theta_3, \theta_4$ of a quadrilateral on the ground. If these observations are subject to independent errors with zero means and common variance σ^2, determine least-squares estimates of the θs, and obtain an unbiased estimate of σ^2.

Suppose that the quadrilateral is known to be a parallelogram with $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$. What then are the least-squares estimate of its angles, and how would you estimate σ^2?

3.6 A chemical compound can be produced by a certain process without the help of a catalyst, but it is hoped that the yield will be increased if a catalyst is present. To investigate this, five identical containers are used in the following way.

Container	*Treatment*	*Yield*
1	No catalyst	x_1
2	Catalyst A at strength a_1	x_2
3	Catalyst A at strength a_1	x_3
4	Catalyst B at strength a_2	x_4
5	Catalyst B at strength $2a_2$	x_5

Assuming that regression of yield on strength is linear for each catalyst, obtain least-squares estimates of the 'unaided' effect and of the two regression coefficients.

Derive the variance matrix of these estimators (making the usual assumptions about errors) and deduce that, for given a_1, the least-squares estimator of the difference of regression coefficients has minimum variance when $a_2 = a_1$.

3.7 A deterministic process $y_0, y_1, \ldots, y_n$ is governed by the relation

$$y_{i+1} = ay_i \quad (i = 0, 1, \ldots, n-1),$$

where a is a known constant. The y_is cannot be observed without error and observations $x_0, x_1, \ldots, x_n$ are such that

$$x_i = y_i + \varepsilon_i \quad (i = 0, 1, \ldots, n),$$

where the ε_is are uncorrelated errors with common variance. Determine least-squares estimates of $y_0, y_1, \ldots, y_n$.

If a were unknown, how then would you estimate $y_0, y_1, \ldots, y_n$?

In each case obtain an unbiased estimate of the error variance.

3.8 In example 3.4, take $r = 3$ and verify directly the general results of section 3.10.2.

4 The Method of Maximum Likelihood

4.1 The likelihood function

The justification of the method of least squares requires no knowledge of the form of the distribution of the error vector apart from its mean and variance matrix, and the method can be applied without this further knowledge. The method of maximum likelihood, on the other hand, is applicable mainly in situations where the true distribution on the sample space is known apart from the values of a finite number of unknown real parameters. So maximum likelihood is usually applied when the family of possible distributions on the sample space can be labelled by a parameter θ taking values in a finite-dimensional Euclidean space. In addition, its application is generally restricted to the case where this family $\{P_\theta : \theta \in \Theta\}$ (Θ a subset of R^s, say) possesses density functions $\{p_\theta ; \theta \in \Theta\}$ with respect to some 'natural' measure on the sample space, such as counting measure if the sample space is discrete or Lebesgue measure when it is not; in the discrete case $p_\theta(x)$ is 'the probability of the point x when θ is the true parameter'; in the continuous case $p_\theta(x)$ is 'the probability density at x when θ is the true parameter'.

It is convenient now to change our notation and write $p(x, \theta)$ instead of $p_\theta(x)$; and we make a distinction between the function $p(\cdot, \theta)$ which is a density function on the sample space, and the function $p(x, \cdot)$ which is a function on the parameter space. The latter function, $p(x, \cdot)$, is called the *likelihood function* corresponding to the observation x, or simply the likelihood function. It expresses the plausibilities of different parameters after we have observed x, in the absence of any other information we may have about these different values. (This last sentence might well be the subject of some controversy, but we shall return to this point later.)

The method of maximum likelihood has a strong intuitive appeal and according to it, we estimate the true parameter θ by any parameter which maximizes the likelihood function $p(x, \cdot)$; such a parameter belongs to the set most plausible after we have observed x. Often there is a unique maximizing parameter which is *the* most plausible and this is then *the* maximum-likelihood estimate.

4.1.1 *Definition*

A maximum-likelihood estimate $\hat{\theta}(x)$ is any element of Θ such that

$$p\{x, \hat{\theta}(x)\} = \max_{\theta \in \Theta} p(x, \theta).$$

Of course it is possible, if, for instance, Θ is an open set, that no maximum-likelihood estimate exists. However in practice this does not often cause trouble.

Again formally at this stage we make the distinction between the estimate $\hat{\theta}(x)$ and the estimator $\hat{\theta}$, but we shall not maintain this distinction consistently, leaving the context to make it clear whether we are thinking of $\hat{\theta}(x)$ as a function or as a particular value of a function.

4.1.2 *Example*

The results of n independent trials in each of which the probability of success is θ are $x = (x_1, x_2, \ldots, x_n)$, where as usual each x_i is either 0 or 1. Find the maximum-likelihood estimate of θ.

The likelihood function, defined on the interval (0, 1), is given by

$$p(x, \theta) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i},$$

and its maximum occurs at

$$\hat{\theta}(x) = \frac{\sum x_i}{n} = \frac{\text{number of successes}}{\text{number of trials}}.$$

So in this case the maximum-likelihood estimator coincides with the M.V.U.E.

4.1.3 *Example*

Let $x = (x_1, x_2, \ldots, x_n)$ be a random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown. Find maximum likelihood estimates of μ and σ^2 or, equivalently, the maximum likelihood estimate of $\theta = (\mu, \sigma^2)$. In this case the likelihood function, defined for all real μ and all $\sigma^2 > 0$, is

$$p(x, \mu, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{\frac{1}{2}n}} \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right].$$

Maximizing p, which is non-negative, is equivalent to maximizing $\log p$ and

$$\log p(x, \mu, \sigma^2) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2.$$

We find the maximizing values by the standard method for maximizing a function of two variables, namely equating partial derivatives to zero.

This gives $\quad \frac{1}{\sigma^2} \sum (x_i - \mu) = 0$

and $\quad -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (x_i - \mu)^2 = 0,$

equations which have the unique solution

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2.$$

It is not difficult to verify that these values of μ and σ^2 yield an absolute (not only a local) maximum of the log-likelihood, so that they are maximum-likelihood estimates.

4.2 Calculation of maximum-likelihood estimates

In these two examples it was possible to find relatively simple expressions in closed form for maximum-likelihood estimates, but often this is not possible and numerical methods are necessary. It is usually possible to assume that maximum-likelihood estimates emerge as a solution of the 'likelihood equations', namely

$$\frac{\partial}{\partial \theta_i} \log p(x, \theta) = 0, \quad (i = 1, 2, \ldots, s).$$

However, these equations often have to be solved numerically.

A standard method of solving the likelihood equations is Newton's method or an adaptation of it. Symbolically the equations we have to solve may be written

$$D_\theta l(x, \theta) = 0,$$

where $l(x, \theta) = \log p(x, \theta)$ and D_θ is the vector differential operator whose ith component is $\partial/\partial\theta_i$. By exploiting special features of the situation under investigation, as in the example following this section, we can often obtain a good initial approximation $\theta^{(0)}$ to the solution $\hat{\theta}$ of these equations. Then we expand by Taylor's theorem as far as terms of first order in $\hat{\theta} - \theta^{(0)}$ to obtain

$$0 = D_\theta\, l(x, \hat{\theta}) \simeq D_\theta\, l(x, \theta^{(0)}) + \{D_\theta^2\, l(x, \theta^{(0)})\}(\hat{\theta} - \theta^{(0)}),$$

where D_θ^2 is the matrix operator

$$\left[\frac{\partial^2}{\partial\theta_i\, \partial\theta_j}\right].$$

It follows from this that

$$\hat{\theta} \simeq \theta^{(0)} - \{D_\theta^2\, l(x, \theta^{(0)})\}^{-1} D_\theta\, l(x, \theta^{(0)}),$$

and the right hand side of this equation is a new approximation $\theta^{(1)}$ to the maximum-likelihood estimate $\hat{\theta}$.

Now we repeat this process, using $\theta^{(1)}$ instead of $\theta^{(0)}$, to obtain a new approximation $\theta^{(2)}$, and so on. Thus we establish an iterative procedure for obtaining a sequence $(\theta^{(n)})$ which usually converges to $\hat{\theta}$. This is Newton's method.

The laborious aspect of this iterative procedure is the inversion of the matrix $D_\theta^2\, l(x, \theta^{(i)})$ at the ith stage. If our initial approximation $\theta^{(0)}$ is good, then $D_\theta^2\, l(x, \theta^{(0)})$ will be near $D_\theta^2\, l(x, \theta^{(i)})$ in non-pathological conditions, so that we can often use the former matrix *at each stage* of the procedure and so avoid the necessity for a new matrix inversion at every stage. This modified procedure leads to a new sequence of approximations to $\hat{\theta}$, a sequence which usually converges to $\hat{\theta}$, though possibly more slowly than the sequence $(\theta^{(n)})$.

A further modification sometimes reduces the total amount of computation even further. There is sometimes good reason to suppose that the matrix $D_\theta^2\, l(x, \theta^{(0)})$ will be relatively close to its expected value $E_\theta^{(0)}\{D_\theta^2\, l(x, \theta^{(0)})\}$; close enough, that is, to ensure that a sequence of approximations to $\hat{\theta}$, based on the use of this expected value rather than on $D_\theta^2\, l(x, \theta^{(0)})$ itself, will still converge to $\hat{\theta}$. Now it often happens that terms awkward to calculate appear in $D_\theta^2\, l(x, \theta^{(0)})$ but not in its expected value. So again it is sometimes possible to reduce total calculation by using $E_\theta^{(0)}\{D_\theta^2\, l(x, \theta^{(0)})\}$ in place of $D_\theta^2\, l(x, \theta^{(0)})$.

We recall that $E_\theta^{(0)}\{D_\theta^2\, l(x, \theta^{(0)})\}$ is simply, in most instances, $-\mathrm{B}_\theta^{(0)}$ where B_θ is the information matrix (see section 2.13). It follows that the fully modified iterative procedure is defined as follows:

$$\theta^{(n+1)} = \theta^{(n)} + \mathrm{B}_{\theta^{(0)}}^{-1}\{D_\theta\, l(x, \theta^{(n)})\},$$

where $\theta^{(0)}$ is an initial approximation to $\hat{\theta}$, usually obtained by exploiting special statistical features of the problem involved.

4.2.1 *Example*

Suppose that it may be assumed that the probability $\pi(s)$ that an individual responds to the level s of a stimulus can be expressed, at least approximately, in the form

$$\pi(s) = \Phi\left(\frac{s-\mu}{\sigma}\right) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{(s-\mu)/\sigma} e^{-\frac{1}{2}z^2}\, dz,$$

an assumption which may appear somewhat drastic, but which in fact turns out to be valid in many circumstances. The level s_i of the stimulus is applied to n_i individuals $(i = 1, 2, \ldots, r)$ and the numbers m_i $(i = 1, 2, \ldots, r)$ of responses at the different levels are observed. Determine maximum-likelihood estimates of μ and σ.

In this particular case we have, in our general notation,

$$x = (m_1, m_2, \ldots, m_r) \qquad \theta = (\mu, \sigma)$$

and $$p(x, \theta) = \prod_{i=1}^{r} \binom{n_i}{m_i} \{\pi(s_i)\}^{m_i} \{1 - \pi(s_i)\}^{n_i - m_i},$$

assuming, of course, that individuals respond independently of one another.

Hence, writing π_i in place of $\pi(s_i)$ for symmetry of notation, we have

$$l(x, \theta) = \text{constant} + \sum_i \{m_i \log \pi_i + (n_i - m_i) \log (1 - \pi_i)\},$$

and the likelihood equations, $D_\theta l(x, \theta) = 0$, are

$$\frac{\partial}{\partial \mu} l(x, \theta) = \sum_i \frac{m_i - n_i \pi_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \mu} = 0$$

and $$\frac{\partial}{\partial \sigma} l(x, \theta) = \sum_i \frac{m_i - n_i \pi_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \sigma} = 0.$$

It will be appreciated that these equations are not susceptible to methods of solution which are other than numerical, and our first problem is to obtain initial approximations μ_0 and σ_0 to their solution.

Φ is a monotonic increasing function. Let Φ^{-1} denote the inverse function defined on (0, 1). If we knew the π_is, then when we plotted the points $(s_i, \Phi^{-1}(\pi_i))$, according to our assumption regarding π, these points would lie on the straight line

$$\Phi^{-1}(\pi) = \frac{(s - \mu)}{\sigma}.$$

Of course we do not know the π_is but we do have estimates of them since m_i/n_i is an estimate of $\pi_i (i = 1, 2, \ldots, r)$. Consequently if we plot the points $(s_i, \Phi^{-1}(m_i/n_i))$, and if our assumption regarding π is justified, these points should be scattered around a straight line. Plotting these points therefore gives us at the same time a check on the validity of our assumption about π (if they are obviously non-linear the assumption is not justified), and a means of obtaining initial approximations to the solution of the likelihood equations. For if we fit a straight line to this set of points, the parameters of the fitted line yield estimates of the true parameters μ and σ, estimates which approximate to the maximum-likelihood estimates, the solution of the likelihood equations.

We now illustrate the point of replacing $D_\theta^2 l(x, \theta)$ by its expected value. In our example a typical element of the former matrix is

$$\frac{\partial^2}{\partial \mu^2} l(x, \theta),$$

which is rather a complicated expression. Note however that all but one of the terms which arise under the summation sign when we differentiate $\partial l(x, \theta)/\partial \mu$ with respect to μ, contain as a factor $m_i - n_i \pi_i$, whose expectation is zero. It follows that

$$E_\theta\left[\frac{\partial^2}{\partial \mu^2} l(x, \theta)\right] = \sum_i \frac{-n_i}{\pi_i (1 - \pi_i)} \left(\frac{\partial \pi_i}{\partial \mu}\right)^2,$$

and similarly that the information matrix B_θ is given by

$$B_\theta = \begin{bmatrix} \sum_i \frac{n_i}{\pi_i (1 - \pi_i)} \left(\frac{\partial \pi_i}{\partial \mu}\right)^2 & \sum_i \frac{n_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \mu} \frac{\partial \pi_i}{\partial \sigma} \\ \sum_i \frac{n_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \mu} \frac{\partial \pi_i}{\partial \sigma} & \sum_i \frac{n_i}{\pi_i (1 - \pi_i)} \left(\frac{\partial \pi_i}{\partial \sigma}\right)^2 \end{bmatrix}.$$

With our assumption regarding the form of π_i we have

$$\frac{\partial \pi_i}{\partial \mu} = -\frac{1}{\sigma} \phi\left(\frac{s_i - \mu}{\sigma}\right), \quad \frac{\partial \pi_i}{\partial \sigma} = -\frac{s_i - \mu}{\sigma^2} \phi\left(\frac{s_i - \mu}{\sigma}\right),$$

where $\phi(y) = \frac{d}{dy} \Phi(y)$.

The calculations involved in the iterative procedure for evaluating $\hat{\mu}$ and $\hat{\sigma}$ are still not trivial, but they are not prohibitive. For further details of the organization of these calculations, and for numerical examples, the reader may refer to Finney (1947); this is an example of an important practical technique called *probit analysis*.

4.3 Optimal properties of maximum-likelihood estimators

The Gauss–Markov theorem provides a justification for the method of least-squares in terms of the concept of minimum-variance unbiasedness, and it is natural to inquire whether a similar justification for the method of maximum likelihood can be found. Unfortunately it is not generally true that maximum-likelihood estimators are unbiased; for instance if $x_1, x_2, \ldots, x_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown, the maximum-likelihood estimator of $\theta = (\mu, \sigma^2)$ is $(\bar{x}, s^2)$

where $$s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2,$$

and while it is true that $E_\theta(\bar{x}) = \mu$, for all θ, it is *not* true that $E_\theta(s^2) = \sigma^2$. In fact $E_\theta(s^2) = (n-1)n^{-1}\sigma^2$, for all θ. (Of course whether we use this as a criticism of the method of maximum likelihood or as a criticism of the concept of unbiasedness is a moot point.)

We can make one or two fairly obvious statements which provide a very partial justification of the method.

Firstly in a 'regular' situation where there exists an unbiased estimator whose variance attains the Cramér–Rao lower bound, the maximum-likelihood estimator coincides with this. For then (section 2.10.1) $\partial \log p(x, \theta)/\partial \theta$ can be expressed in the form $a(\theta)\{\tilde{\theta}(x) - \theta\}$, and the only solution of the likelihood equation $\partial \log p(x, \theta)/\partial \theta = 0$ is $\theta = \tilde{\theta}(x)$, which gives an absolute maximum of $\log p(x, \theta)$ and therefore $\tilde{\theta}(x)$ is the maximum-likelihood estimate.

Secondly it is often possible to show that a maximum-likelihood estimator has high efficiency (section 2.11) in the Fisherian sense. This of course provides a justification only in particular cases.

Thirdly we can say that the maximum-likelihood estimator is a function of a minimal-sufficient statistic. This follows directly from the factorization theorem (section 2.3.3) and it means that the maximum-likelihood estimator depends only on relevant information contained in an observation. It does not mean necessarily that it makes the 'best use' of this information according to some specified definition of 'best use'. The main justification of the method of maximum likelihood is a 'large-sample' one, which shows that when an observation provides lots of information about an unknown parameter, the method utilizes essentially all of this information. We expand this rather vague statement in the following sections.

4.4 Large-sample properties

When we talk about a large sample we mean that the observation x takes the form $x = (x_1, x_2, \ldots, x_n)$, where n is large, and the x_is are independent and identically distributed.

In this case $p(x, \theta) = \prod_{i=1}^{n} p^*(x_i, \theta)$,

where $p^*(\cdot, \theta)$ is the density function, corresponding to the parameter value θ, on the space of a 'single observation'.

Also $l(x, \theta) = \log p(x, \theta) = \sum_{i=1}^{n} \log p^*(x_i, \theta)$,

regarded as a random variable, is the sum of the independent identically distributed random variables $\log p^*(x_i, \theta) (i = 1, 2, \ldots, n)$.

4.4.1 Now let us fix attention on one particular distribution on the sample space, say that corresponding to the parameter θ_0, which we will think of as the true parameter. For *any* fixed θ, $l(x, \theta)$ is a random variable whose distribution is determined by the 'true' distribution on the sample space.

Let $z(\theta) = E_0\left[\frac{1}{n} l(x, \theta)\right] = E_0\{\log p^*(x_i, \theta)\}$,

where the subscript on the expectation operator is used to emphasize the fact that we are taking expectations relative to the distribution corresponding to θ_0.

This function $z(\theta)$ has a property which is, in a sense, the key to the study of large-sample properties of maximum-likelihood estimators: *$z(\theta)$ attains its maximum value at θ_0, and if distributions on the sample space corresponding to different parameters are essentially different, then for no other θ is $z(\theta)$ equal to $z(\theta_0)$.* This important result is a particular case of the following general result derived from Jensen's inequality.

4.4.2 *Theorem*

Let q and r be density functions of two different probability distributions on the same probability space Y of points y, these distributions being different in the sense that there exists a set of positive q-probability on which $q(y) \neq r(y)$; and let C be any continuous convex function of a non-negative variable.

Then $E_q\left[C\left(\frac{r}{q}\right)\right] \geqslant C(1)$,

with strict inequality if C is strictly convex.

Proof. By Jensen's inequality we have

$$E_q\left[C\left(\frac{r}{q}\right)\right] \geqslant C\left[E_q\left(\frac{r}{q}\right)\right]$$

and the inequality is strict if C is strictly convex, since r/q is not constant with q-probability 1 by assumption.

This simple proof is now completed by the remark that

$$E_q\left(\frac{r}{q}\right) = \int_Y \frac{r(y)}{q(y)} q(y)\, dy, \quad \text{symbolically}$$

$$= \int_Y r(y)\, dy = 1,$$

and so $E_q\left[C\left(\frac{r}{q}\right)\right] \geqslant C(1)$.

4.4.3 In the above theorem, let $C = -\log$, let Y be the space of a 'single observation' and let $q = p^*(\cdot, \theta_0)$, $r = p^*(\cdot, \theta)$.

This gives $E_{\theta_0}\left[-\log \frac{p^*(\cdot, \theta)}{p^*(\cdot, \theta_0)}\right] \geqslant -\log 1 = 0$

i.e. $z(\theta_0) - z(\theta) \geqslant 0$, or $z(\theta_0) \geqslant z(\theta)$,

and the inequality is strict if the distributions corresponding to θ_0 and θ are essentially different.

So far we have assumed no structure on the parameter space Θ. Typically this space will have a mathematical structure; in particular, if it is a Euclidean space, it has a metric, and then it is usually the case that when θ is near θ_0, $z(\theta_0) - z(\theta)$ is small and when θ is far away from θ_0, $z(\theta_0)$ is considerably larger than $z(\theta)$.

Also in the large sample case the *law of large numbers* ensures that when n is large, $n^{-1} l(x, \theta)$ is, for most x and each particular θ, near $z(\theta)$. Suppose that we assume sufficient regularity to enable us to demonstrate that, for large n and most x, $n^{-1} l(x, \theta)$ is *uniformly* (with respect to θ) near $z(\theta)$. Then the

following picture emerges for the case where θ is a real parameter, a case of sufficient generality to illustrate the general case also.

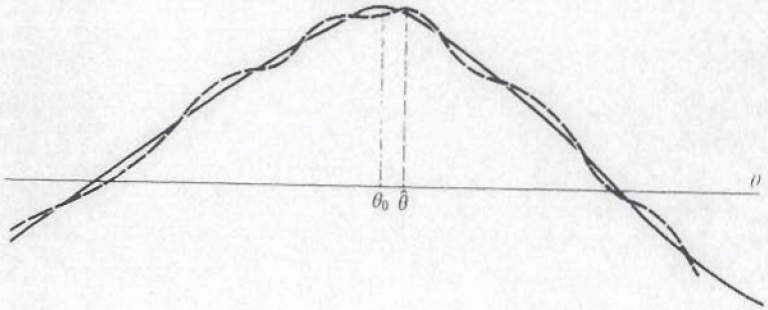

Figure 3 Proximity of the graphs of $z(\theta)$ (unbroken line) and $\frac{1}{n} l(x, \theta)$ (broken line) ensures that $\hat{\theta}$ is near θ_0

In Figure 3, the unbroken curve is the graph of z and the dotted curve is the graph of $n^{-1} l(x, \theta)$ for a typical x. The fact that z assumes its maximum value at θ_0, and that $n^{-1} l(x, \theta)$ is uniformly near $z(\theta)$ ensures that $n^{-1} l(x, \theta)$ assumes its maximum value at a point near θ_0, that is, $\hat{\theta}(x)$ is near θ_0.

4.5 Consistency

In section 4.4 we have outlined the ideas underlying a proof of the fact that the method of maximum likelihood has a property called consistency, defined as follows.

4.5.1 *Definition*

Let $(\tilde{\theta}_n)$ be a sequence of estimators of a parameter θ belonging to a metric space Θ. This sequence is said to be weakly consistent if $\tilde{\theta}_n$ tends in θ-probability to θ; strongly consistent if $\tilde{\theta}_n$ converges with θ-probability one to θ, for all $\theta \in \Theta$.

4.5.2 The reader who is unfamiliar with general metric spaces may think of the parameter space Θ as being the real line, without losing anything essential from the statistical idea here. This idea arises from the following consideration:

Suppose that we continue repeating an experiment which, we feel, is in some sense providing information about an unknown real parameter θ involved in a probabilistic model of the experiment. If the repetitions are independent, then, as their number increases, we feel that we ought to be obtaining more and more information about θ; that, if we are estimating θ, our estimates should get closer and closer to the true value, whatever this may be; and that

finally, when the number of repetitions is very, very large, we ought to be fairly certain about what the true value of the parameter is.

Precise mathematical content is given to this notion by the statement that *consistent* estimation should be possible in the circumstances described. It then becomes important to demonstrate that any method of estimation which we employ does have this property of consistency.

Now if (x_n) is a sequence of random variables whose joint distributions depend on an unknown parameter θ in a metric space Θ, we may define a sequence $(\hat{\theta}_n)$ of maximum-likelihood estimators of θ, $\hat{\theta}_n$ being the maximum-likelihood estimator based on $x_1, x_2, \ldots, x_n$. Section 4.4 outlines the main ideas underlying a proof of the fact that if the x_is are independent and identically distributed, the sequence $(\hat{\theta}_n)$ is consistent: weakly consistent if a weak law of large numbers is employed; strongly consistent if a strong law is used. Of course analytic details are required and regularity conditions must be introduced, for a complete proof, which is quite complicated. The reader is referred to Wald (1949) for such a complete proof.

4.6 Large-sample efficiency

The main justification of the method of maximum likelihood in terms of the criterion of minimum-variance unbiasedness is that it is possible to show that for large samples, subject to regularity conditions, maximum-likelihood estimators are nearly unbiased and have variances nearly equal to the Cramér–Rao lower bound. Again a full proof of this result is hedged around with analytic details and regularity conditions and we content ourselves with a heuristic argument. We consider the case where there is an unknown *real* parameter θ, a case of sufficient generality to illustrate the probabilistic content of the argument.

Suppose then that (x_n) is a sequence of independent identically distributed random variables, the distribution of each being known apart from the value of a single real parameter θ; and let $\hat{\theta}_n$ be the maximum-likelihood estimator (which we assume unique) of θ derived from $x_1, x_2, \ldots, x_n$. We shall now assume that n is large and we shall omit the subscript n for typographical brevity. We assume also that $\hat{\theta}$ emerges as a solution of the likelihood equation

$$D_\theta l(x, \theta) = 0,$$

where $x = (x_1, x_2, \ldots, x_n)$,

$$l(x, \theta) = \log p(x, \theta) = \sum_{i=1}^{n} \log p^*(x_i, \theta)$$

and D_θ is the differential operator $\partial/\partial\theta$, as before.

Section 4.4 tells us that with θ-probability near 1, $\hat{\theta}$ is near θ. We therefore have

$$0 = D_\theta l(x, \hat{\theta}) = D_\theta l(x, \theta) + (\hat{\theta} - \theta) D_\theta^2 l(x, \theta) + R(x, \theta, \hat{\theta}),$$

where $R(x, \theta, \hat{\theta})$ is a remainder term involving $(\hat{\theta}-\theta)^2$, which may be shown to be of smaller order than the first-order term, $(\hat{\theta}-\theta)D_\theta^2 l(x, \theta)$, if regularity conditions are satisfied. We can therefore say that, with θ-probability near 1,

$$\hat{\theta}-\theta \simeq -\frac{D_\theta l(x, \theta)}{D_\theta^2 l(x, \theta)}$$

$$\text{or} \quad \sqrt{n}(\hat{\theta}-\theta) \simeq \frac{n^{-\frac{1}{2}} D_\theta l(x, \theta)}{n^{-1} D_\theta^2 l(x, \theta)}.$$

$$\text{Now} \quad \frac{1}{\sqrt{n}} D_\theta l(x, \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_\theta \log p^*(x_i, \theta),$$

and each random variable in the sum on the right hand side has zero mean and variance i_θ, Fisher's measure of information from a single observation (section 2.9). Consequently by the central limit theorem, $n^{-\frac{1}{2}} D_\theta l(x, \theta)$ is approximately $N(0, i_\theta)$.

$$\text{Moreover} \quad \frac{1}{n} D_\theta^2 l(x, \theta) = \frac{1}{n} \sum_{i=1}^{n} D_\theta^2 \log p^*(x_i, \theta)$$

and, by Lemma 2.11.2,

$$E_\theta\{-D_\theta^2 \log p^*(x_i, \theta)\} = i_\theta.$$

Therefore, by a law of large numbers, $-n^{-1} D_\theta^2 l(x, \theta)$ is approximately equal to i_θ.

It follows that $\sqrt{n}(\hat{\theta}-\theta)$ is approximately $i_\theta^{-1} \times$ (an $N(0, i_\theta)$ random variable), so that $\hat{\theta}$ is approximately $N\{\theta, (ni_\theta)^{-1}\}$, i.e., $N(\theta, I_\theta^{-1})$ where I_θ^{-1} is the inverse of Fisher's measure of information from $x_1, x_2, \ldots, x_n$, or the Cramér–Rao lower bound for the variance of an unbiased estimator of θ based on $x_1, x_2, \ldots, x_n$.

So we have 'proved' that maximum-likelihood estimators are efficient for large samples and in addition that $\hat{\theta}$ is approximately normally distributed, in this case where there is a single unknown real parameter. For a complete proof of this result the reader is referred to Cramér (1946), p. 500.

4.6.1 This property generalizes to the case where θ is a vector-valued parameter. The basic results used in the above proof are:

(a) Taylor's theorem in the expansion of $D_\theta l(x, \hat{\theta})$;

(b) a central limit theorem applied to $n^{-\frac{1}{2}} D\, l(x, \theta)$;

(c) a law of large numbers applied to $n^{-1} D^2\, l(x, \theta)$.

Each of these results has a multivariate version and the vector-parameter argument is simply a straight generalization of that above, yielding the result

that in this case, for large samples, $\hat{\theta}$ is approximately $N(\theta, (n\mathrm{B}_\theta)^{-1})$ where B_θ is the information matrix (section 2.12) for a single observation.

4.7 Restricted maximum-likelihood estimates

4.7.1 On certain occasions, when a family of distributions on a sample space is labelled by a vector-valued parameter θ, we have additional knowledge about the true parameter and we know that it satisfies certain restrictions. Then the parameter space Θ is expressed in the form

$$\Theta = \{\theta : \theta \in \mathrm{R}^s, h(\theta) = 0\},$$

where $h(\theta) = [h_1(\theta), h_2(\theta), \ldots, h_r(\theta)]$ is a vector-valued function mapping R^s into R^r. Of course we wish an estimate of the true parameter to belong to Θ so that, as far as the method of maximum likelihood is concerned, we wish a restricted maximum-likelihood estimate, that is an estimate which maximizes the likelihood function subject to the restriction $h(\theta) = 0$.

4.7.2 As far as the theory of such restricted maximum-likelihood estimates is concerned, the natural mathematical approach is to reduce this case to that studied in section 4.6.1 by an initial re-parametrization. We 'fill out' the restricting functions $h_1, h_2, \ldots, h_r$ to a set $h_1, h_2, \ldots, h_r, h_{r+1}, \ldots, h_s$, in such a way that the function $h^* = (h_1, h_2, \ldots, h_s)$ is a one-to-one function from R^s onto itself.

Then by setting $\phi_i = h_i(\theta_1, \theta_2, \ldots, \theta_s) \quad (i = 1, 2, \ldots, s)$

we obtain a new labelling of the family of possible distributions by the parameter $\phi = (0, 0, \ldots, 0, \phi_{r+1}, \ldots, \phi_s)$, whose first r components we may ignore, since they are all zero. Thus from a theoretical point of view this new problem is essentially the same as that of estimating an unrestricted parameter belonging to R^{s-r} and the properties established for the method of maximum likelihood in the latter case, asymptotic minimum-variance unbiasedness etc. will apply to restricted estimates also. Again only the bones of a rigorous argument are given here and these would require to be supplemented by regularity conditions and more details to complete the discussion.

4.7.3 The natural practical approach to the problem of finding restricted maximum likelihood estimates is a direct attack by the method of Lagrange multipliers which leads to the restricted likelihood equations

$$D_\theta l(x, \theta) - \mathrm{H}_\theta \lambda = 0$$
$$h(\theta) = 0,$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)'$ is a column-vector of Lagrange multipliers and H_θ is the $s \times r$ matrix of partial derivatives $\partial h_j(\theta)/\partial \theta_i$. With sufficient regularity the restricted maximum-likelihood estimate $\hat{\theta}(x)$ emerges as a solution of these equations along with an appropriate Lagrange multiplier $\hat{\lambda}(x)$.

It is not possible to say much in general about this estimate $\hat{\theta}(x)$. However if we know that, with θ-probability near 1, $\hat{\theta}$ is very near θ, then the above restricted likelihood equations are approximately linear and by the same

kind of argument as in section 4.6.1 we can obtain approximations to the distribution of $\hat{\theta}$. In particular, if we are dealing with a large sample then, subject to what is mild regularity from a practical point of view, it is true that $\hat{\theta}$ is very probably very near the true parameter θ – the argument of section 4.4 carries over with little modification, as the reader may verify. Let us suppose then that we are dealing with a sample of n, where n is large, so that $x = (x_1, x_2, \ldots, x_n)$ and the x_is are independent and identically distributed.

We have
$$D_\theta l(x, \hat{\theta}) - H_{\hat{\theta}} \hat{\lambda} = 0$$
$$h(\hat{\theta}) = 0,$$

and using Taylor's theorem to linearize about the true parameter θ (which we recall satisfies $h(\theta) = 0$), we have approximately

$$D_\theta l(x, \theta) + \{D_\theta^2 l(x, \theta)\}\{\hat{\theta} - \theta\} - H_{\hat{\theta}} \hat{\lambda} = 0$$
$$H_\theta'(\hat{\theta} - \theta) = 0.$$

The fact that the term $H_{\hat{\theta}} \hat{\lambda}$ can simply be replaced by $H_\theta \hat{\lambda}$ requires some explanation. This is because when $\hat{\theta}$ is near θ, it is also near the element $\tilde{\theta}$ of R^s at which $l(x, \theta)$ takes its absolute maximum, so that $\hat{\lambda}$ is relatively small; hence when $H_{\hat{\theta}}$ is expanded about θ, the first-order terms in the expansion involve $\hat{\theta} - \theta$ *and* $\hat{\lambda}$ and so are of smaller order than those which we have included. Slight manipulation of these equations yields

$$\left[-\frac{1}{n} D^2 l(x, \theta)\right] \sqrt{n}(\hat{\theta} - \theta) + H_\theta \frac{1}{\sqrt{n}} \hat{\lambda} = \frac{1}{\sqrt{n}} D_\theta l(x, \theta)$$
$$H_\theta' \sqrt{n}(\hat{\theta} - \theta) = 0,$$

or, in matrix notation,

$$\begin{bmatrix} -\frac{1}{n} D^2 l(x, \theta) & H_\theta \\ H_\theta' & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\hat{\theta} - \theta) \\ \frac{1}{\sqrt{n}} \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}} D_\theta l(x, \theta) \\ 0 \end{bmatrix}.$$

We now apply the law of large numbers to $-n^{-1} D^2 l(x, \theta)$ and find, as before, that this is approximately B_θ (the information matrix for a single observation). The central limit theorem applied to $n^{-\frac{1}{2}} D_\theta l(x, \theta)$ shows that it is approximately an $N(0, B_\theta)$ random variable. Carrying our approximation this one stage further shows, therefore, that approximately

$$\begin{bmatrix} B_\theta & H_\theta \\ H_\theta' & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\hat{\theta} - \theta) \\ \frac{1}{\sqrt{n}} \hat{\lambda} \end{bmatrix} = \begin{bmatrix} Z \\ 0 \end{bmatrix}$$

where Z is $N(0, B_\theta)$.

This is virtually the same set of equations as we had when dealing with the linear model subject to restrictions – which is not surprising since we arrived at this point by linearizing a non-linear set of equations. We can therefore carry over the results of section 3.10, simply by replacing $A'A$ by B_θ and H by H_θ and we find that, when B_θ has rank s,

$$\begin{bmatrix} \sqrt{n}(\hat{\theta} - \theta) \\ \frac{1}{\sqrt{n}} \hat{\lambda} \end{bmatrix}$$

is approximately normally distributed with zero mean and variance matrix

$$\begin{bmatrix} P_\theta & 0 \\ 0 & -R_\theta \end{bmatrix}, \quad \text{where} \quad \begin{bmatrix} B_\theta & H_\theta \\ H_\theta' & 0 \end{bmatrix}^{-1} = \begin{bmatrix} P_\theta & Q_\theta \\ Q_\theta & R_\theta \end{bmatrix}.$$

4.7.4 *Non-identifiability and singularity of the information matrix*

There is a connexion between non-identifiability of a vector-valued parameter θ and singularity of the information matrix B_θ, which becomes clearer if we examine the function $z(\theta)$ introduced in section 4.4.

Suppose that a family of distributions is labelled by a parameter θ which ranges over an s-dimensional subset Θ of R^s, that is, Θ contains an s-dimensional rectangle. Let θ_0 be a particular element of Θ and θ a neighbouring point, and as in section 4.4.1,

let $z(\theta) = E_0\{\log p(x, \theta)\}$,

where $p(\cdot, \theta)$ is the density function on the sample space defining the distribution corresponding to the parameter θ. Let us suppose further that there is enough regularity in the family $\{p(\cdot, \theta) : \theta \in \Theta\}$ of density functions to permit the following operations, which we have already encountered.

$$\begin{aligned} z(\theta) &= E_0\{\log p(x, \theta)\} \\ &= E_0\{\log p(x, \theta_0) + [D_\theta \log p(x, \theta_0)]'(\theta - \theta_0) + \\ &\qquad + (\theta - \theta_0)'[D_\theta^2 \log p(x, \theta_0)](\theta - \theta_0)\} + \text{terms of third order} \\ &= z(\theta_0) - (\theta - \theta_0)' M_{\theta_0}(\theta - \theta_0) + \text{small terms,} \end{aligned}$$

where M_θ is the information matrix for x. We know from our previous study of the function z, that if θ is identifiable, (that is, if different θs corresponding to different distributions) then $z(\theta_0)$ is an absolute maximum of z. This usually means in practice that the second-order terms in the expansion of $z(\theta)$ about $z(\theta_0)$ are negative, that is, that M_{θ_0} is positive definite. Of course this is not necessarily so. It is possible that $(\theta - \theta_0)' M_{\theta_0}(\theta - \theta_0)$ is zero and that higher order terms ensure that $z(\theta_0) > z(\theta)$. However this is unusual in practice, and usually identifiability of θ, together with the kind of regularity which permits the expansion of $z(\theta)$ indicated above, ensures that M_{θ_0} is positive definite, at least when θ_0 is an interior point of Θ.

Conversely, if M_{θ_0} is singular and so indefinite, this in practice usually means that there are parameters $\theta \neq \theta_0$ such that $z(\theta) = z(\theta_0)$ and this in turn means that there are different parameters yielding essentially the same distribution on the sample space, so that θ is not identifiable.

It is quite clear that any formal result connecting non-identifiability of θ and singularity of the information matrix which we might try to state would have to be hedged around by so many conditions that its content would be obscured. So we leave the discussion in this informal state, noting that usually lack of identifiability of θ implies singularity of the information matrix and vice versa.

4.7.5 For the linear model we discussed the possibility (section 3.4) of non-identifiability and the adjustment necessary to the technique for finding restricted estimates in this case. A similar adjustment is often possible in the above large-sample theory for restricted maximum-likelihood estimates in the case where the information matrix is singular. For, as indicated in section 4.7.4, this often means that θ is not identifiable without restrictions. However a number of the restrictions

$$h_i(\theta) = 0 \qquad (i = 1, 2, \ldots, r),$$

say the first t of these, are just enough to ensure identifiability (as in the linear case); and this usually ensures that the matrix $B_\theta + H_{1\theta}H'_{1\theta}$ is positive definite, where $H_{1\theta}$ is the leading $s \times t$ submatrix of H_θ. The adjustment is now similar to that in the linear case. We replace B_θ by $B_\theta + H_{1\theta}H'_{1\theta}$ wherever it appears, and now $\sqrt{n}(\hat{\theta} - \theta)$ is approximately normally distributed with zero mean and variance matrix P_θ, the leading $s \times s$ submatrix in

$$\begin{bmatrix} B_\theta + H_{1\theta}H'_{1\theta} & H_\theta \\ H'_\theta & 0 \end{bmatrix}^{-1}.$$

For further details see Silvey (1959).

4.7.6 *Example*

In an experiment for measuring the DNA content of a particular type of cell, there is a chance of mistaking two cells for one, so that the experimental result may be a measurement of the DNA content of a single cell or of the sum of the contents of two cells. From the results $x = (x_1, x_2, \ldots, x_n)$ of a large number n of independent repetitions of this experiment, it is desired to estimate the mean and standard deviation of the DNA content of single cells.

In order that it should be possible to apply the method of maximum likelihood to this problem it is necessary to set up a model which involves only a finite number of unknown parameters. Now in this situation it is fairly realistic to assume that the DNA content of a single cell is normally distributed with unknown mean μ and unknown variance σ^2. There is an unknown probability α of mistaking two cells for one. If we further assume that when two cells are mistaken for one, these two cells may be regarded as independent of one another, then measurements resulting from the observation of two cells are normally distributed with mean 2μ and variance $2\sigma^2$. With these assumptions, the probability density on the line to describe the result of a single replicate of the experiment is

$$p^*(t, \theta) = (1-\alpha)\frac{1}{\sigma\sqrt{(2\pi)}}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] + \alpha\frac{1}{\sqrt{2\sigma}\sqrt{(2\pi)}}\exp\left[-\frac{(t-2\mu)^2}{2\times 2\sigma^2}\right]$$

and the probability density on the sample space for n repetitions, R^n, is

$$p(x, \theta) = \prod_{i=1}^{n} p^*(x_i, \theta).$$

Here $\theta = (\alpha, \mu, \sigma)$ and we have a family of distributions on the sample space parametrized by a 3-vector, so that the method of maximum likelihood may be applied, in the same kind of way as in section 4.2.1.

Another way of setting up a model for this example is less sensible from a computational point of view but yields an illustration of the application of the method of computing restricted estimates. So we consider it for this reason. As before we assume that the DNA content of single cells is normally distributed, with unknown mean and variance which we now denote by μ_1 and σ_1^2 respectively. Again there is an unknown probability α of mistaking two cells for one. However we now assume that a measurement resulting from the observation of two cells is normally distributed with mean μ_2 and variance σ_2^2, so that the probability density on the line to describe the result of a single replicate of the experiment is now

$$p^*(t, \theta) = (1-\alpha)\frac{1}{\sigma_1\sqrt{(2\pi)}}\exp\left[-\frac{(t-\mu_1)^2}{2\sigma_1^2}\right] + \alpha\frac{1}{\sigma_2\sqrt{(2\pi)}}\exp\left[-\frac{(t-\mu_2)^2}{2\sigma_2^2}\right],$$

where $\theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$, a 5-vector.

Correspondingly $p(x, \theta) = \prod_{i=1}^{n} p^*(x_i, \theta)$,

and this density also involves five unknown parameters. If we are prepared to make the assumption that two cells mistaken for one are independent

then $\mu_2 - 2\mu_1 = 0$

and $\sigma_2^2 - 2\sigma_1^2 = 0$,

and we may consider maximizing $p(x, \cdot)$ subject to these restrictions by the Lagrange multiplier technique of section 4.7.3. This of course is equivalent to the previous 'unrestricted maximization', but the reader may find it instructive to follow the theory for each case through in terms of this particular example.

Examples

4.1 Let $x_1, x_2, \ldots, x_n$ be a random sample from a distribution with density $p(x, \theta)$ depending on an unknown real parameter θ. Find the maximum-likelihood estimate of θ in the following cases.

(a) $p(\cdot, \theta)$ is the density function of a Poisson distribution with mean θ;
(b) $p(\cdot, \theta)$ is the density function of an exponential distribution, $p(x, \theta) = \theta e^{-x\theta} \quad (x > 0)$;
(c) $p(\cdot, \theta)$ is the density function of the uniform distribution on $(0, \theta)$.

In each case determine the distribution of the maximum-likelihood estimator. For cases (a) and (b) verify the large sample theory of chapter 4. Show that this theory is not applicable in case (c) and explain why.

4.2 On the Aegean island of Kalythos, the inhabitants suffer from a congenital eye disease whose effects become more marked with increasing age. Samples of fifty people were taken at five different ages and the numbers of blind people counted:

Age	20	35	45	55	70
Number of blind	6	17	26	37	44.

It is conjectured that the probability of blindness at age x, $P(x)$, can be expressed in the form

$$P(x) = \{1 + e^{-(\alpha + \beta x)}\}^{-1}.$$

Comment on whether this hypothesis is reasonable, by constructing a suitable graph. Estimate α and β from the graph and then obtain maximum-likelihood estimates. Estimate also the age at which it is just more likely than not that an islander is blind.

4.3 A certain type of electrical component is manufactured in a large number of factories. The proportion p of defective components varies from factory to factory, and over factories p has approximately a B-distribution with density

$$\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)},$$

where α and β are unknown parameters. Suppose that s factories are chosen at random and that n components produced by each are inspected. Given that m_i of the inspected components of the ith factory are defective $(i = 1, 2, \ldots, s)$, explain in detail how to calculate maximum-likelihood estimates of α and β. Show that if $n = 1$, α and β are not identifiable.

4.4 Suppose that one has n pairs of measurements $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the $2n$ values being distributed normally and independently with variance σ^2. The mean of x_i is ξ_i, that of y_i is η_i and the n pairs (ξ_i, η_i) lie on a circle centre (ξ, η) and radius ρ. It is required to estimate ξ, η and ρ. Obtain a maximum-likelihood solution of this problem, and elaborate the computational details. (*Camb. Dip.*)

4.5 In an experiment to measure the resistance of a crystal, independent pairs of observations $(x_i, y_i)(i = 1, 2, \ldots, n)$ of current x and voltage y are obtained. These are subject to errors (ε_i, η_i), so that

$$x_i = \mu_i + \varepsilon_i, \qquad y_i = v_i + \eta_i,$$

where μ_i and v_i are the true values of current and voltage on the ith occasion and $v_i = \alpha\mu_i$, α being the resistance of the crystal.

On the assumption that the errors are independently and normally distributed with zero means and variances, $\operatorname{var} \varepsilon_i = \sigma_1^2$, $\operatorname{var} \eta_i = \sigma_2^2 = \lambda\sigma_1^2$, where λ is known, show that $\hat{\alpha}$, the maximum-likelihood estimator of α is a solution of the equation

$$\hat{\alpha}^2 S_{xy} + \hat{\alpha}(\lambda S_{xx} - S_{yy}) - \lambda S_{xy} = 0,$$

where $S_{xy} = \frac{1}{n}\sum x_i y_i, \qquad S_{xx} = \frac{1}{n}\sum x_i^2, \qquad S_{yy} = \frac{1}{n}\sum y_i^2.$

Show that, if $\sum \mu_i^2/n$ tends to a limit as $n \to \infty$, then $\hat{\alpha}$ is a consistent estimator of α.

Show that the method of maximum likelihood gives unsatisfactory results when λ is not assumed known. Explain why the standard theorems for maximum-likelihood estimators do not apply to this problem. (*Camb. Dip.*)

4.6 A radioactive sample emits particles randomly at a rate which decays with time, the rate being $\lambda e^{-\kappa t}$ after time t. The first n particles emitted are observed at successive times $t_1, t_2, \ldots, t_n$. Set up equations for maximum-likelihood estimates $\hat{\lambda}$ and $\hat{\kappa}$, and show that $\hat{\kappa}$ satisfies the equation

$$\frac{\hat{\kappa} t_n}{e^{\hat{\kappa} t_n} - 1} = 1 - \hat{\kappa}\bar{t},$$

where $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i.$

Find a simple approximation for $\hat{\kappa}$ when $\hat{\kappa} t_n$ is small. (*Camb. Dip.*)

4.7 A cell contains granules which may be regarded as spheres of equal but unknown radius r, and which may be assumed to be distributed randomly throughout the cell. In order to estimate r, a section of the cell is observed under a microscope and this section contains circular sections of n granules. If the radii of these sections are $x_1, x_2, \ldots, x_n$, determine the maximum-likelihood estimate of r. What is its distribution, for large n?