CHAPTER 3

Models for continuous data with constant variance

3.1 Introduction

Generalized linear models are essentially an extension of classical linear models and this chapter presents these classical models in a way that makes the extension appear natural. There is an enormous literature on classical linear models, not all of it helpful to the reader, and no attempt will be made in this chapter to give a comprehensive account of the subject. Rao (1973), Draper and Smith (1981), Seber (1977) and Atkinson (1985) are excellent reference books covering various aspects of classical linear models.

The subject matter of this chapter is linear models, which we shall write in the following form:

$$\begin{array}{lll} Y_i \sim N(\mu_i, \sigma^2), & \boldsymbol{\mu} = \boldsymbol{\eta}, & \boldsymbol{\eta} = \sum_1^p \mathbf{x}_j \beta_j, \\ \text{observations Normally distributed and independent;} & \text{identity link;} & \text{linear predictor based on covariates } \mathbf{x}_1, \ldots, \mathbf{x}_p. \end{array} \tag{3.1}$$

The data vector $\mathbf{y}$, the mean vector $\boldsymbol{\mu}$, and the linear predictor, $\boldsymbol{\eta}$, all have n components. The leftmost component of (3.1) is a specification of the random part of the model. The other components describe the systematic parts, which include the construction of the linear predictor $\boldsymbol{\eta}$ from the covariates, and the link between $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$. By suppressing the link, and regarding the $\mathbf{x}_j$ as the p columns of a matrix $\mathbf{X}$, we recover the standard matrix formulation

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the set of parameters written in vector form. Note that we have restricted our attention to the sub-class of linear models in which there is only one error component and in which the errors are independent. Models involving components of variance are therefore excluded.

We now consider in more detail the random and systematic parts of the model (3.1).

3.2 Error structure

In classical linear models, the vector of observations, $\mathbf{y}$, is assumed to be a realization of a random variable, $\mathbf{Y}$, which is Normally distributed with moments

$$E(\mathbf{Y}) = \boldsymbol{\mu} \quad \text{and} \quad \text{cov}(\mathbf{Y}) = \sigma^2\mathbf{I}. \tag{3.2}$$

Thus the observations are assumed to have equal variances and to be independent.

The assumption of Normality, although important as the basis for an exact small-sample theory, is not so important in large samples. For there the central-limit theorem offers protection from all but the most extreme distributional deviations from Normality. There may, however, be a modest loss of efficiency, which can be recovered if the true distribution is known and used in place of the Normal. For details, see Cox and Hinkley (1968).

The theory of least squares can be developed using only first- and second-moment assumptions in addition to independence, without requiring the additional assumption of Normality. This is fortunate because in applications we can rarely be entirely confident that the assumed distributional form is correct. It is this second-order aspect of linear models that is emphasized here. From the present viewpoint, therefore, the important assumption in (3.2) is that the variance of an observation is the same for all values of μ. This is an assumption that can and should be checked, either by graphical examination of the residuals or by computing an appropriate test statistic. Checks such as these are described in Chapter 12.

The emphasis on second-moment assumptions over fully specified distributional assumptions extends to all generalized linear models and is discussed more fully in Chapter 9.

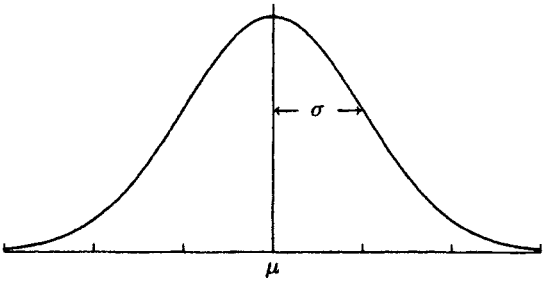

Fig. 3.1. *The Normal (or Gaussian) distribution with mean μ and standard deviation σ.*

The frequency function of the univariate Normal distribution takes the form

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad \text{for} \quad -\infty < y < \infty.$$

The distribution is symmetrical with mode, mean and median all at μ. The standard deviation, σ, is the horizontal distance between the mean and the point of inflection of the density. About 68%, 95% and 99.8% of the distribution lies in the ranges $\mu \pm \sigma$, $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$ respectively. The log-likelihood function for a single observation with known variance is a parabola whose maximum is at y and whose second derivative is $-1/\sigma^2$.

The Normal distribution is useful primarily as a model for measurements of continuous quantities, though it can also be used as an approximation for discrete measurements. It is frequently used to model data, such as weights, lengths and time, which, though continuous, are essentially positive, although the distribution itself covers the entire real line. Such usage is acceptable in practice provided that the data values are sufficiently far removed from zero. If, for example, data have a mean of 100 and a standard deviation of 10, the part of the Normal distribution covering the negative half of the real line is negligible for most practical purposes. If data y that are essentially positive approach the origin, then it will often be found that the data themselves contradict the assumption of constant variance independent of μ. When this occurs, a Normal distribution for $\log Y$ will often be found to be a better approximation than a Normal distribution for Y. Alternatively, the gamma distribution (Chapter 8) may be used.

3.3 Systematic component (linear predictor)

We aim in this section to study various aspects of the linear predictor

$$\eta = \sum_{1}^{p} \mathbf{x}_j \beta_j,$$

which occurs in all generalized linear models. The covariates, $\mathbf{x}_1, \ldots, \mathbf{x}_p$, may be continuous measurements, incidence vectors for qualitative factors of various types, or incidence vectors for interactions among these. Concise description and automatic construction of such vectors is an important aspect of the specification and fitting of generalized linear models.

3.3.1 *Continuous covariates*

These comprise covariates such as mass, temperature, time, amount of fertilizer or drug, concentration of a solute and so on, which can take values on a continuous scale. Models containing only terms with continuous covariates are often called *regression models*, to be contrasted with *analysis-of-variance* models, which have only terms involving qualitative factors. Provided that there is only one component of error variance, we shall not make this distinction. Indeed, by introducing mixed terms in section 3.3.4, we shall deliberately seek to blur the distinction because many interesting models involve terms of both types.

Linearity in the present context means linearity of η in the parameters. Consequently a continuous covariate x in a model term may be replaced by an arbitrary function $g(x)$, such as log(dose) in a dose-response model, without destroying the linearity of the model. In particular we may use $x^2, x^3, \ldots$ in addition to x to build up a polynomial in x, without destroying the linearity. Similarly, the linear model $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ may be expanded to include the product term $\beta_{12} x_1 x_2$, producing a bilinear relationship. If the terms are rearranged in the form

$$(\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_{12} x_2) x_1,$$

they show a linear relationship in x_1 in which both slope and intercept are linear functions of x_2. The alternative rearrangement

$$(\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_{12} x_1) x_2$$

expresses the bilinearity in complementary form.

A function such as $\exp(\gamma x)$, however, produces a non-linear model unless γ is known a priori. If γ is unknown, the model is not linear, and some non-linear optimization technique is required to minimize the discrepancy function. It may, however, be helpful to fit the model for a few suitably chosen values of γ. Such models, which are partly linear and partly non-linear, are discussed in Chapter 11.

3.3.2 *Qualitative covariates*

Sets of observations are frequently indexed by one or more classifying factors, or factors for short. Each factor has an associated index, whose values partition the data into disjoint groups or classes. Thus, in a field experiment, one such factor might define the block into which each unit (plot) falls, while another might define the crop variety to be planted in that plot.

A factor can take only a limited set of possible values, to be called *levels*. The k levels can always be coded using the integers $1, 2, \ldots, k$, although the coding $0, 1, \ldots, k-1$ is sometimes more convenient. Such a coding defines the *formal levels* of a factor. In practice the levels usually have names or numerical values and these we call *actual levels*. Actual levels may be

1. ordered with numerical formal levels, such as the amount of fertilizer in an agricultural experiment; or
2. ordered but without relative magnitudes for the levels, such as socioeconomic status; or
3. unordered, such as the names of crop varieties in a variety trial.

Factors occurring in a model may be of primary interest, meaning that a principal purpose of the study is to measure their effect. Treatment factors in a designed experiment are obviously of this kind. In surveys, classification factors such as educational status, marital status, religious affiliation and so on are of this type. Factors of secondary interest are those producing effects that must be accommodated in the model, but which are not of primary interest. Examples are blocking factors in a randomized blocks design and, usually, census enumeration district in a survey. The distinction between primary and secondary factors is not absolute, but depends on the aims of the study concerned.

The simplest term in a linear predictor generated by a factor is a component of the intercept. Consider a model with one covariate x and linear predictor

$$\eta = \alpha + \beta x.$$

If A is a factor with index i, then the extended linear predictor might become

$$\eta_i = \alpha_i + \beta x,$$

implying a separate intercept for each level of A, but a common slope β, assumed constant over the levels of the factor. Note that if a factor has numerical levels, we could also treat it as a quantitative covariate having only a few distinct values. If we treat it as a factor, we fit a separate effect for each level in an unstructured way, whereas if we treat it as a quantitative variate, we impose a linear form on the response. Alternatively, and perhaps preferably, we may use polynomials in the actual levels to detect deviations from linearity.

Frequently data are cross-classified by many factors simultaneously. If A, B and C are three such factors with indices i, j, k respectively, the simplest model ordinarily considered has the form

$$\alpha_i + \beta_j + \gamma_k.$$

This is the so-called main-effects model, which implies that if we arrange the data in a rectangular block and then look at cross-sections of the data for each level of A, we shall find that they can be modelled by effects of B and C that are additive and equal in each cross-section. Similarly for the other factors. In order to achieve a satisfactory fit, however, it may be necessary to include terms analogous to $\beta_{12}x_1x_2$ with continuous covariates. Such terms, of the algebraic form $(\alpha\beta)_{ij}$, imply a separate effect for each combination of the indices i and j and are called *interactions*. We shall refer to $(\alpha\beta)_{ij}$ as a two-factor interaction, but the term 'first-order interaction' is also used, the order being one less than the number of factors involved.

The relationships between interactions and main effects have been the subject of much confusion in the literature. We consider them in more detail in section 3.5.

3.3.3 *Dummy variates*

If i is the index for the levels of factor A with k levels, the term α_i may be written in vector notation as

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \ldots + \alpha_k\mathbf{u}_k,$$

where the $\mathbf{u}_j$ are *dummy variates* whose components take the value 1 if the unit has factor A at level j, and zero otherwise. The terms *incidence vector* and *indicator vector* are also used. Thus if $k = 3$ and the formal levels for five observations are 1, 2, 2, 3, 3, the dummy variates $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ take values as follows:

Unit	A	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$
1	1	1	0	0
2	2	0	1	0
3	2	0	1	0
4	3	0	0	1
5	3	0	0	1

Note that,

$$\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 \equiv \mathbf{1}$$

irrespective of the allocation of levels to units. The constant vector, $\mathbf{1}$, is the dummy variate corresponding to the intercept term, often written as μ, in the linear predictor. The relation between the terms μ and α_j is a simple instance of intrinsic aliasing, to be discussed in section 3.5.

A compound term such as $(\alpha\beta)_{ij}$ has dummy variates, $(\mathbf{uv})_{ij}$, whose values are products of corresponding components of $\mathbf{u}_i$ and $\mathbf{v}_j$, the dummy variates for A and B as single-factor terms. It follows then that

$$\sum_i (\mathbf{uv})_{ij} = \mathbf{v}_j \quad \text{and} \quad \sum_j (\mathbf{uv})_{ij} = \mathbf{u}_i,$$

again irrespective of the allocation of factor levels to units. Thus main effects are intrinsically aliased with interactions in which they are included.

3.3.4 *Mixed terms*

In section 3.3.2 we considered a model

$$\eta_i = \alpha_i + \beta x,$$

in which the intercept varies with the factor level, but where the slope is constant over levels. Sometimes, however, the slope may also change with the factor level, requiring the term βx to be replaced by $\beta_i x$. Terms in the linear predictor in which a slope or regression coefficient changes with the level of one or more factors are called *mixed*, because they include aspects of both continuous and qualitative covariates. It is important that any computer program for fitting linear models should allow mixed terms to be specified as easily as continuous and qualitative terms, because the assumption frequently made, that a slope is the same for all levels of a factor, ought to be easily testable. The simplest test is to compare the fit of the model having constant slope with the fit when the slope is allowed to vary from level to level.

Dummy variates for mixed terms take the same form as those for factors except that the 1s are replaced by the corresponding x-values. Using the same factor allocation as in the previous section, and with the covariate x as shown, the dummy variates for the mixed term $\beta_i x$, again written as $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$, take values as follows:

Unit	*A*	x	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$
1	1	1	1	0	0
2	2	3	0	3	0
3	2	5	0	5	0
4	3	7	0	0	7
5	3	9	0	0	9

Here

$$\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 = \mathbf{x},$$

again irrespective of the allocation of levels to units.

3.4 Model formulae for linear predictors

3.4.1 *Individual terms*

We now describe a notation that is helpful for the specification of linear predictors in generalized linear models. The notation, due to Wilkinson and Rogers (1973), is compact and is easily adapted for use in computer programs. The convention is continued that names beginning with letters from the first half of the alphabet refer to factors, and those from the second half to continuous covariates. The indices associated with the levels of factors $A, B, C, \ldots$ are $i, j, k, \ldots$. The Table below lists some kinds of terms that occur in simple model formulae. Algebraic expressions are presented together with the corresponding model formula term. Note the use of λ instead of β for the coefficient of a continuous covariate to avoid confusion with the parameters for factor B.

Type of term	*Algebraic*	*Model formula term*
Continuous covariate	λx	X
Factor	α_i	A
Mixed	$\lambda_i x$	$A.X$
Compound	$(\alpha\beta)_{ij}$	$A.B$
Compound mixed	$\lambda_{ij} x$	$A.B.X$

In the model-formula version X stands for itself, a single vector. By contrast, A stands for a *set* of dummy variates, one variate as indicator for each level of the factor. The remaining types of term also stand for the appropriate set of dummy variates. Thus terms in a model formula represent vector subspaces and do not involve the parameters explicitly. Parameters occur only implicitly, one per basis vector in each subspace.

3.4.2 *The dot operator*

This operator, already exemplified in the formation of compound terms, implies the formation of all elementwise products of the constituent vectors. For example, if A is the three-level factor and X the covariate vector with values shown at the end of section 3.3.4, then $A.X$ denotes the three vectors $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ shown there. By extension, if B is a two-level factor with dummy variates $\mathbf{v}_1, \mathbf{v}_2$, then $A.B$ denotes six dummy vectors corresponding to all

elementwise products of $\mathbf{u}_i$ with $\mathbf{v}_j$. Note that if A has k levels, $A.A$ comprises k vectors equal to the k dummy vectors for A, and $k(k-1)$ null vectors with all components zero. Such null vectors may be omitted, effectively making

$$A.A = A.$$

However, in general,

$$X.X \neq X,$$

the left-hand side being a vector with components x_i^2. (Note that in both the computer programs GLIM and Genstat, where this notation has been adopted, compound terms involving more than one continuous covariate are not permitted in model formulae. They must be computed explicitly, preferably after subtracting column means.)

The dot operator is commutative so that

$$A.B \equiv B.A,$$

and associative, so that

$$(A.B).C \equiv A.(B.C).$$

Thus we may write $A.B.C$ without ambiguity, the order in which the factors are included being unimportant.

3.4.3 *The + operator*

Terms in a model formula may be joined using the operator +, with exactly the same usage as in the algebraic expression for the model formula. Repetitions of terms are ignored, so that

$$A + A \equiv A,$$

it being pointless to specify the same vector subspace twice. In vector-space terminology $A + B$ defines a subspace in R^n spanned by linear combinations of vectors in A and B.

It is convenient to assign lower priority to + than to the dot, so that

$$A.B + C \equiv (A.B) + C.$$

The dot is distributive with respect to +, so that

$$A.(B + C) \equiv A.B + A.C.$$

These are the fundamental operators required in the specification of model formulae and the other useful operators that follow are defined in terms of them.

3.4.4 *The crossing* (∗) *and nesting* (/) *operators*

The crossing operator, denoted by ∗, is used mainly to simplify the specification of factorial models. Thus

$$A * B \equiv A + B + A.B$$
$$A * B * C \equiv A + B + C + A.B + A.C + B.C + A.B.C,$$

and so on. In these expansions A and B may themselves be replaced by model formulae. The operator ∗ has higher priority than +, but lower priority than dot. Thus

$$A * B + C \equiv A + B + C + A.B$$
$$A * B.C \equiv A + B.C + A.B.C.$$

Note the convention followed in expanding expressions, that all simple terms come first, followed by two-component terms and so on. This convention, though not essential, is helpful when it comes to understanding intrinsic aliasing in models (Section 3.5).

The crossing operator is associative, and distributive with respect to +, for

$$\begin{aligned} A * (B + C) &\equiv A + (B + C) + A.(B + C) \\ &\equiv A + B + C + A.B + A.C \\ &\equiv A + B + A.B + A + C + A.C \\ &\equiv A * B + A * C. \end{aligned}$$

When a compound term such as $A.B$ is preceded in an expanded model formula by both constituent terms A and B, it is called the *interaction* of A and B. The nature of the interaction term will be discussed further in section 3.5.

The nesting operator / relates to an indexing system, which, in its simplest form, has two indices i and j, but no connection between observations (i, j) and (i', j), though there is a connection between observations (i, j) and (i, j'). Typically, i defines the levels of a blocking factor and j identifies an element within a block. There is no necessary connection between the first observation in one block and the first observation in another, but two observations in the same block have their block in common and may tend to be similar on that account. For a nested treatment structure, consider

a set of plant varieties categorized as early ($i = 1$), mid-season ($i = 2$), or late ($i = 3$) in cropping. Within each group there is a number of distinct varieties, no variety belonging to more than one group. Two varieties may be connected by being in the same group (i the same), but there is no connection between the first variety in two different cropping groups (j the same, i different). The appropriate linear predictor for nesting is written as

$$A/B \equiv A + A.B,$$

In the expanded formula the compound term $A.B$ is preceded by only one constituent term. The interpretation of $A.B$ is now that of B within A.

As before, A and B may themselves be model formulae, with the rule that if $\mathrm{pt}(A)$ denotes the product term (using dots) of all elements in A, then A/B is defined by

$$A/B \equiv A + \mathrm{pt}(A).B.$$

Thus, for example

$$(A * B)/C \equiv A * B + A.B.C.$$

The nesting operator is associative, so that

$$A/(B/C) \equiv (A/B)/C,$$

and distributive with respect to $+$, since

$$A/(B+C) = A + A.(B+C) = A + A.B + A + A.C = A/B + A/C.$$

Like the crossing operator, the nesting operator is given a priority between . and $+$. By convention we give it higher priority than $*$.

3.4.5 *Operators for the removal of terms*

The operator $-$ has the obvious meaning as the inverse or opposite of $+$. It is used for the removal of terms in a model formula. Thus

$$A * B - A.B \equiv A + B.$$

Similarly,

$$A * B * C - A.B.C \equiv A + B + C + A.B + A.C + B.C$$

is a concise notation for a model with all main effects and two-factor interactions.

It is sometimes required to remove from a model all those compound terms that include a given factor or factors. Two operators $-/$ and $-*$ cater for this; $-/A$ means 'remove all compound terms that include A, but excluding A itself', while $-*A$ means 'remove all terms that include A'. Thus

$$A * B * C -/ A \equiv A + B + C + B.C$$

and

$$A * B * C -* A \equiv B + C + B.C.$$

3.4.6 *Exponential operator*

If M is a model formula and I is an integer, then

$$M ** I \equiv M * M * \ldots * M,$$

the right side containing I Ms. This operator is useful for specifying factorial models that include all terms up to a given level of interaction. For example

$$(A + B + C) ** 2 \equiv A + B + C + A.B + A.C + B.C.$$

This operator has highest priority.

We shall use this notation for the specification of linear predictors wherever possible. Readers should bear in mind that this model-formula notation (strictly speaking, a subset of it) can be used directly for this purpose in the computer systems Genstat and GLIM.

3.5 Aliasing

Each term in a model formula describes a set of covariates to be included in a linear predictor. If such a set is denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_p$, the $\mathbf{x}$s being n-vectors, then the covariates can be thought of as defining p directions in n-dimensional Euclidean space. These p vectors define a subspace of up to p dimensions. The maximum dimension is achieved if the $\mathbf{x}$s are linearly independent, i.e. if there does not exist a set of coefficients ξ_j, not all zero, such that

$$\sum_1^p \xi_j \mathbf{x}_j = \mathbf{0}.$$

If k independent linear relations exist, then the set of covariates spans a space of dimension $p - k$. Ordinarily the individual terms in an expanded model formula will form subspaces of maximum dimension. Loss of dimension may occur, however, when we consider joint subspaces covered by more than one term.

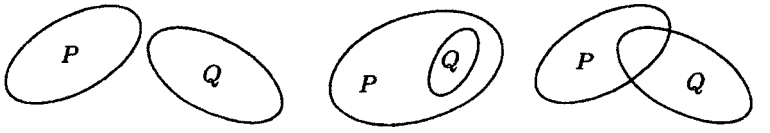

Fig. 3.2. *Venn diagrams for relationships between subspaces of terms in a linear model:* (i) *P and Q linearly independent;* (ii) *Q entirely aliased with P;* (iii) *Q partially aliased with P.*

We now consider the possible relationships between the subspaces defined by two terms in a model formula. The terms are denoted by P and Q, their dimensions by p and q, with $p \geq q$. There are three possible relationships between P and Q.

1. All $p + q$ vectors defining P and Q are linearly independent, so that the dimension of the space $P + Q$ is $p + q$.
2. All the vectors of Q are expressible as linear combinations of the p vectors in P, so that the dimension of $P + Q$ is p.
3. k of the q vectors in Q are expressible as linear combinations of those in P.

The corresponding Venn diagrams are shown in Fig. 3.2. Clearly (i) and (ii) are extreme cases of (iii) for which $k = 0$ and $k = q$ respectively. Note the special case of (ii) when $p = q$, so that P and Q span identical subspaces.

The effect on the terms in a generalized linear model of overlapping subspaces is to produce what is called aliasing. Certain combinations of covariates are then identical to other combinations, so that the corresponding combinations of parameters cannot be distinguished. Consider, for example, measurements made on leaves having the property that area = constant×length×breadth, with length and breadth being measured as well as area. Suppose that the covariates in the model are

$$\begin{aligned} x_1 &= \log \text{length}, \\ x_2 &= \log \text{breadth}, \\ x_3 &= \log \text{area}, \end{aligned}$$

and that the linear predictor is to be formed as

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Now, since area = constant×length×breadth, we have

$$x_3 = c + x_1 + x_2, \tag{3.3}$$

where c is the logarithm of the constant in the formula for area. Hence η may be expressed in terms of x_1 and x_2 as

$$\begin{aligned} \eta &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (c + x_1 + x_2) \\ &= \beta_0 + \beta_3 c + (\beta_1 + \beta_3) x_1 + (\beta_2 + \beta_3) x_2. \end{aligned}$$

Thus we can distinguish the three combinations of the βs

$$\beta_0 + \beta_3 c, \qquad \beta_1 + \beta_3, \quad \text{and} \quad \beta_2 + \beta_3,$$

but not the four parameters $\beta_0, \beta_1, \beta_2, \beta_3$ separately. If we write x_0 for the constant vector, i.e. the dummy vector for the term β_0, we see from (3.3) that x_3 is a linear combination of x_0, x_1 and x_2. In other words, the subspace for x_3, with one dimension, is contained in the sum or span of the subspaces for x_0, x_1 and x_2.

If in addition the leaves are all of the same shape, in the sense that the ratio of length to breadth is constant, we have

$$x_2 = c' + x_1,$$

where breadth/length = $\exp(c')$. The linear predictor now reduces further to

$$\begin{aligned}\eta &= \beta_0 + \beta_1 x_1 + \beta_2(c' + x_1) + \beta_3(c + x_1 + c' + x_1) \\ &= \beta_0 + \beta_2 c' + \beta_3(c + c') + (\beta_1 + \beta_2 + 2\beta_3)x_1.\end{aligned}$$

Now, only two parameter combinations, namely

$$\beta_0 + \beta_2 c' + \beta_3(c + c') \qquad \text{and} \qquad \beta_1 + \beta_2 + 2\beta_3,$$

are distinguishable, and the dimension of the space spanned by x_0, x_1, x_2 and x_3 is reduced from four to two.

An important aspect of this example is that the aliasing is intrinsic to the problem. Given that all leaves are the same shape and that all measurements are made without error, aliasing will occur whatever the sizes of the leaves. Such *intrinsic aliasing* is found most commonly, however, where terms involving factors occur in a model.

3.5.1 *Intrinsic aliasing with factors*

Consider a model formula containing the intercept together with the single factor A, which we write as

$$1 + A,$$

where 1 stands for the dummy vector with all elements 1. An equivalent algebraic expression for the components of the linear predictor is

$$\eta_{ij} = \mu + \alpha_i,$$

where i indexes the groups defined by A and j indexes the units or observations within the groups. The dummy vectors for A add up to the constant vector, or dummy vector for μ, because each observation has factor A at exactly one level. Thus μ is aliased with $\sum \alpha_i$, and further, it is intrinsically aliased because

the relation holds whatever the allocation of units to the groups. The relationship between μ and α_i is not symmetric because the dummy vector for μ lies wholly in the space of the dummy vectors for α_i, but not vice versa. We say that μ is *marginal* to the αs. As a consequence, the terms in the model $\mu + \alpha_i$ are ordered because of the marginality relationship. One effect of this ordering is that it does not make sense to consider the hypothesis that $\mu = 0$ when the α_i are not assumed known.

The linear predictor is clearly unchanged if we add a constant to μ and subtract the same constant from each α_i. This operation leaves unchanged the quantities $\mu + \alpha_i$ and also any contrast $\sum \lambda_i \alpha_i$ with $\sum \lambda_i = 0$. Combinations that are unaffected by this operation are said to be *estimable*. The parameters μ and α_i separately are not estimable because the aliasing pattern makes them indistinguishable from $\mu + c$ and $\alpha_i - c$. When we come to estimate the parameters, this ambiguity can be resolved by imposing a constraint on the estimates to give a unique solution to the least-squares equations. It must be stressed, however, that any such constraint on the estimates $\hat{\mu}, \hat{\alpha}_i$ of μ, α_i is a convention only, and is of no significance in judging the adequacy of the model. Constraints are not to be thought of as part of the model specification: they are merely a convenient way of resolving an ambiguity and they do not affect the meaning or interpretation of the model. In particular, there is no implication that a similar constraint should be imposed on the parameters μ and α_i; in fact, where intrinsic aliasing occurs, the imposition of constraints on parameters as well as on their estimates is a common source of confusion.

For the above model, three possible constraints, chosen from an infinity of possibilities, are as follows:

1. $\hat{\mu} = 0$, so that the $\hat{\alpha}_i$ give the group means directly;
2. $\hat{\alpha}_1 = 0$, so that the first group mean is $\hat{\mu}$, and $\hat{\alpha}_2, \hat{\alpha}_3, \ldots$ measure differences between other group means and the first;
3. $\sum \hat{\alpha}_i = 0$, so that $\hat{\mu}$ is the average of the group means and $\hat{\alpha}_i$ is the deviation of the ith group mean from $\hat{\mu}$.

As an example, consider four groups with means 6, 9, 12 and 13. Then the three constraints produce parameter estimates in the linear predictor with the following values:

	Estimate with constraint		
Parameter	(1.)	(2.)	(3.)
μ	0	6	10
α_1	6	0	−4
α_2	9	3	−1
α_3	12	6	2
α_4	13	7	3

Another constraint that is sometimes used if the group sizes are unequal is

$$\sum w_i \hat{\alpha}_i = 0,$$

where w_i is the ith group size. With this constraint $\hat{\mu}$ is a weighted average of the group means and $\hat{\alpha}_i$ is the deviation of the ith group mean from the weighted average.

3.5.2 *Aliasing in a two-way cross-classification*

Failure to recognize the aliasing pattern and the arbitrariness of imposed constraints has led to much confusion in the literature, especially in the analysis of models for two-way cross-classifications. The discussion here follows the lines of Nelder (1977).

We are concerned with the linear model

$$1 + A + B + A.B,$$

expressed algebraically by the linear predictor

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

The dummy vectors for the four terms show the following relationships, here written in terms of the parameters rather than the dummy vectors. (The equivalence sign may be read as 'is indistinguishable from'.)

$$\sum \alpha_i \equiv \mu, \qquad \sum \beta_j \equiv \mu,$$

$$\sum_j \gamma_{ij} \equiv \alpha_i, \qquad \sum_i \gamma_{ij} \equiv \beta_j.$$

These identities imply that the sum of all the dummy vectors for $A.B$ is the constant vector, or, in terms of the parameters,

$$\sum_{ij} \gamma_{ij} \equiv \mu.$$

Thus the relationships among the terms are as follows:

$$\begin{aligned} &\mu \text{ is marginal to } \alpha_i, \beta_j \text{ and } \gamma_{ij}, \\ &\alpha_i \text{ is marginal to } \gamma_{ij} \\ \text{and } &\beta_j \text{ is marginal to } \gamma_{ij}. \end{aligned}$$

The terms are thus partially ordered as first μ, then α_i and β_j together, and finally γ_{ij}. The estimable parameter combinations are the linear predictor itself,

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

and also the contrasts

$$\begin{aligned} &\textstyle\sum \lambda_i(\alpha_i + \bar{\gamma}_{i\cdot}); \quad \text{with} \quad \sum \lambda_i = 0, \\ &\textstyle\sum \lambda_j(\beta_j + \bar{\gamma}_{\cdot j}); \quad \text{with} \quad \sum \lambda_j = 0, \\ \text{and} \quad &\textstyle\sum \lambda_{ij}\gamma_{ij}; \quad \text{with} \quad \sum_i \lambda_{ij} = \sum_j \lambda_{ij} = 0, \end{aligned}$$

where $\bar{\gamma}_{i\cdot}, \bar{\gamma}_{\cdot j}$ denote averages over the indicated indices.

The ambiguities about the values of the estimates of individual parameters can again be resolved by suitable constraints. Two such constraints are now discussed for a 2×2 array.

The full parameterization has nine parameters, namely $(\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22})$, but only four linearly independent estimable combinations. Thus, five suitably chosen constraints are required to produce unique estimates. The conventional system of symmetric constraints is given by

$$\begin{aligned} \hat{\alpha}_1 + \hat{\alpha}_2 &= 0, & \hat{\beta}_1 + \hat{\beta}_2 &= 0, \\ \hat{\gamma}_{11} + \hat{\gamma}_{12} &= 0, & \hat{\gamma}_{11} + \hat{\gamma}_{21} &= 0, \\ \hat{\gamma}_{21} + \hat{\gamma}_{22} &= 0, & \hat{\gamma}_{12} + \hat{\gamma}_{22} &= 0. \end{aligned} \tag{3.4}$$

Note that only three of the last four constraints are linearly independent, so that only five independent constraints are being applied. With these constraints we can solve the four equations

$$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = y_{ij}$$

to give

$$\begin{aligned} &\hat{\mu} = \bar{y}_{..}, \\ &\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}, \qquad \hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}, \\ &\hat{\gamma}_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}. \end{aligned} \tag{3.5}$$

Thus $\hat{\mu}$ is the average of the four observations, $\hat{\alpha}_i$ is the deviation of the ith row mean from the grand mean, and $\hat{\beta}_j$ is a similar deviation for column means. The interaction parameter $\hat{\gamma}_{ij}$ is the deviation of y_{ij}, the linear predictor for that cell, from one based on the addition of main effects, $\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$.

A second set of constraints that lacks symmetry, but is in some ways simpler, is that used by the computer program GLIM, namely

$$\hat{\alpha}_1 = \hat{\beta}_1 = \hat{\gamma}_{11} = \hat{\gamma}_{12} = \hat{\gamma}_{21} = 0. \tag{3.6}$$

More generally, the parameter estimates for the first level of each factor, the first row and column of each two-factor interaction term, and so on, are set equal to zero. With this choice of constraints, the top left-hand corner cell is taken as the baseline, and the estimated parameters are

$$\begin{aligned} &\hat{\mu} = y_{11}, \\ &\hat{\alpha}_2 = y_{21} - y_{11}, \qquad \hat{\beta}_2 = y_{12} - y_{11}, \\ &\hat{\gamma}_{22} = y_{22} - y_{12} - y_{21} + y_{11}. \end{aligned} \tag{3.7}$$

The remaining estimates are zero on account of the constraints. Note that if $\hat{\gamma}_{ij} = 0$, but not otherwise, the $\hat{\alpha}$-contrasts and the $\hat{\beta}$-contrasts given by formulae (3.5) and (3.7) are identical. If further, $\hat{\alpha}_i = \hat{\beta}_j = 0$, then the $\hat{\mu}$s also become identical. These properties are consequences of the marginality relations among the terms in the two-factor model.

We stress again that constraints such as (3.4) and (3.6) are not a part of the model, but merely a convention whereby unique values for estimates of the intrinsically aliased parameters can be produced. For fitting, testing and so on, only estimable combinations are relevant, and those combinations are independent of the constraint system imposed.

3.5.3 *Extrinsic aliasing*

The aliasing patterns considered so far have resulted from intrinsic characteristics of the model formula rather than from particular idiosyncrasies of the data observed. However, aliasing can also occur because the particular covariate vectors observed happen to contain linear dependencies. Suppose we have two factors with three levels each, but data are observed for only 5 of the nine possible combinations as shown below.

Factor			*B*	
	level	1	2	3
	1	×	×	
A	2	×	×	
	3			×

Because of the configuration of the observed factor levels, the dummy vector for α_3 is identical to the dummy vector for β_3. In a complete design, the main-effect subspaces for factors A and B have only a single dimension in common, but here they have a two-dimensional space in common.

The additional aliasing observed is a consequence of the fact that the table of observed factor levels can be split into two disconnected portions, of sizes 2×2 and 1×1. If we move one of the occupied cells to produce the following configuration,

Factor			*B*	
	level	1	2	3
	1	×	×	
A	2	×		
	3		×	×

the aliasing disappears along with the disconnectedness. This example shows how extrinsic aliasing depends on the particular values of covariates in the observed data, in contrast to intrinsic aliasing, which is a property of the model formula alone.

3.5.4 *Functional relations among covariates*

Covariates may be functionally related without being linearly related. The most familiar example is polynomial regression, in which a linear predictor such as

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

contains the power terms x, x^2 and x^3. Provided that more than three distinct x-values are observed, the covariates x, x^2 and x^3 are linearly independent. Thus there is no aliasing of parameters. Nonetheless, there is usually an implied ordering of terms that must be respected in fitting polynomial regression models.

Looking first at the terms β_0 and $\beta_1 x$, we must ask when it makes sense not to use the sequence β_0, $\beta_0 + \beta_1 x$ in model fitting, but to use instead the reverse sequence in which $\beta_1 x$ is fitted first without the intercept. For the latter procedure to make sense, $x = 0$ must correspond to a special point on the scale at which η must be zero. Though this may sometimes happen, there is usually no strong reason for paying special attention to a particular value of x. In agricultural field experiments with fertilizers, for example, there is invariably some small amount of the relevant nutrient already present in the soil, so that zero fertilizer applied does not mean that no nutrient is available to the plant. Thus, zero is not a special point in this example.

Consider next the relationship between the terms $\beta_1 x$ and $\beta_2 x^2$. To fit the terms β_0 and $\beta_0 x^2$ without including $\beta_1 x$ implies that the maximum (or minimum) of the response occurs at $x = 0$, i.e. exclusion of the linear term implies that $x = 0$ is a special point on the scale. For if the x-scale might equally well be measured by $x + c$ as by x, the response $\beta_0 + \beta_2 x^2$ becomes

$$\beta_0 + \beta_2 (x + c)^2 = (\beta_0 + \beta_2 c^2) + 2\beta_2 cx + \beta_2 x^2,$$

and a linear term appears with coefficient $2\beta_2 c$. Ordinarily there is no reason to suppose that the turning point of the response is at a specified point on the x-scale, so that the fitting of $\beta_0 x^2$ without the linear term is usually unhelpful.

A further example, involving more than one covariate, concerns the relation between a cross-term such as $\beta_{12} x_1 x_2$ and the corresponding linear terms $\beta_1 x_1$ and $\beta_2 x_2$. To include the former in

a model formula without the latter two is equivalent to assuming that the point $(0, 0)$ is a col or saddle-point of the response surface. Again there is usually no reason to postulate such a special property for the origin, so that the linear terms must be included with the cross-term. Likewise, the inclusion of quadratic terms $\beta_{11}x_1^2, \beta_{22}x_2^2$ without the cross-term implies that the elliptical contours of constant response are oriented parallel to the axes. Again, there is usually no reason to expect such behaviour in practice, and all second-degree terms should normally be entered into the model simultaneously, assuming, of course, that the linear terms are already present. Thus the relationships among polynomial terms are very similar to those among factors and interactions. This functional marginality is not a true marginality in the sense of section 3.5.1 because no linear dependencies among covariates are involved. Nevertheless, in a similar way, it does impose constraints on the order in which terms should be introduced into a model.

3.6 Estimation

3.6.1 *The maximum-likelihood equations*

Maximum likelihood is the principal method of estimation used for all generalized linear models. For Normal errors, the log likelihood, l, based on n observations is given by

$$-2l = n\log(2\pi\sigma^2) + \sum_{i=1}^{n}(y_i - \mu_i)^2/\sigma^2.$$

For fixed σ^2, known or unknown, maximization of l is equivalent to minimization of the sum of squares

$$\sum(y - \mu)^2$$

for variation in μ. If, in addition, the model is assumed to be linear, we have

$$\eta_i = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

Differentiating with respect to β_j and equating the derivative to zero gives estimating equations in the form

$$\sum_i x_{ij}(y_i - \hat{\mu}_i) = 0 \quad \text{for} \quad j = 1, \ldots, p, \tag{3.8}$$

where the fitted means are given by

$$\hat{\mu}_i = \hat{\eta}_i = \sum x_{ij}\hat{\beta}_j.$$

A useful way of looking at the equations (3.8) is that the p linear combinations of the observations $\sum_i x_{ij}y_i$, $j = 1, \ldots, p$ are set equal to the corresponding linear combinations of the fitted values, namely $\sum_i x_{ij}\hat{\mu}_i$. To state the same thing in an equivalent way, the vector of residuals with components $y_i - \hat{\mu}_i$ is orthogonal to the columns of the model matrix $\mathbf{X}$, so that

$$\mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}.$$

In particular, if $\mathbf{X}$ is the incidence matrix for the main-effects model in a two-way classification, $\mathbf{X}^T\mathbf{y}$ is the set of observed marginal totals. Maximum-likelihood estimation for this Normal-theory model then corresponds to finding fitted values satisfying the model that have the same marginal totals as those observed.

3.6.2 *Geometrical interpretation*

Fitting by ordinary least squares has a simple geometrical interpretation. The data vector $\mathbf{y}$ may be regarded as a point in n-dimensional Euclidean space. For any given value of the parameter vector $\boldsymbol{\beta}$, the vector of fitted values $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ is a point in the same space. As $\boldsymbol{\beta}$ varies over all possible values it might take, $\boldsymbol{\mu}$ traces out a linear subspace or hyperplane called the *solution locus*. If $\mathbf{y}$ falls on the solution locus, the observed values can be reproduced exactly by the model. Ordinarily, however, the observed data point $\mathbf{y}$ does not lie on the solution locus and no value of $\boldsymbol{\beta}$ reproduces the data exactly. If $\boldsymbol{\mu}$ represents a point on the solution locus then $\sum(y_i - \mu_i)^2$ is just the squared Euclidean distance between the observed vector $\mathbf{y}$ and $\boldsymbol{\mu}$. Maximizing the likelihood is then equivalent to choosing the point $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ that is nearest to the observed $\mathbf{y}$ in the sense of minimum Euclidean distance.

To illustrate this geometrical construction, consider the model whose components satisfy $\eta_i = x_i\beta$, with only one covariate and one parameter. The solution locus is the set of all vectors $\mathbf{x}\beta$ for $-\infty < \beta < \infty$, i.e. all points on the line through the origin in R^n in the direction $\mathbf{x}$ (Fig. 3.3). The point on the solution locus that

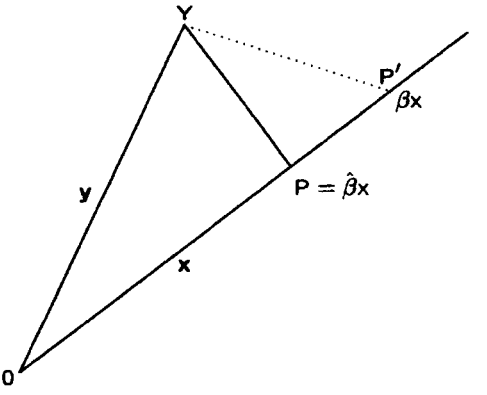

Fig. 3.3. *Least squares: the geometry for one parameter.*

is nearest to $\mathbf{y}$ is found by dropping a perpendicular YP onto $\mathbf{x}\beta$. The coordinates of P are $\mathbf{x}\hat{\beta}$ where $\hat{\beta}$ is the maximum-likelihood estimate of β. The vector $YP = \mathbf{y} - \mathbf{x}\hat{\beta}$ is called the *residual vector*. The condition that OP and PY should be orthogonal, expressed algebraically, is

$$\mathbf{x}^T(\mathbf{y} - \mathbf{x}\hat{\beta}) = 0,$$

i.e.

$$\hat{\beta} = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}}.$$

The *fitted vector*, or vector of *fitted values*, OP, is the *orthogonal projection* of $\mathbf{y}$ on the space $\mathbf{x}$.

3.6.3 *Information*

Further insight into the fit can be obtained by considering how the goodness-of-fit statistic, considered as a function of $\boldsymbol{\beta}$, varies with $\boldsymbol{\beta}$. Let P' be an arbitrary point $\mathbf{x}\beta$ on the solution locus as shown in Fig. 3.3. Then in the triangle YPP' we have

$$(\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta) = (\mathbf{y} - \mathbf{x}\hat{\beta})^T(\mathbf{y} - \mathbf{x}\hat{\beta}) + (\hat{\beta} - \beta)\mathbf{x}^T\mathbf{x}(\hat{\beta} - \beta),$$

expressing the Pythagorean relationship among the sides of the triangle YPP'. If we plot the squared length of YP', which

measures the discrepancy of the data from an arbitrary point of the solution locus, as a function of the parameter β, we obtain a parabola with its minimum at $\beta = \hat{\beta}$, the maximum-likelihood estimate. The minimum discrepancy is $D_{\min} = (\mathbf{y} - \mathbf{x}\hat{\beta})^T(\mathbf{y} - \mathbf{x}\hat{\beta})$, as shown in Fig. 3.4.

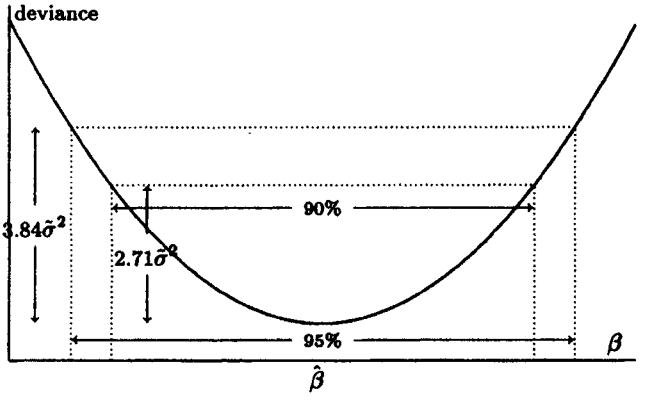

Fig. 3.4. *Information curve for one parameter, together with approximate 90% and 95% confidence intervals.*

The second derivative at the minimum, as indeed elsewhere for a parabola, is given by $2\mathbf{x}^T\mathbf{x}$. If we now restore the dispersion parameter σ^2, which divided the sum of squares, the second derivative becomes $2\mathbf{x}^T\mathbf{x}/\sigma^2$. Apart from the factor 2, this is known as the *Fisher information* for β. If the Fisher information, or curvature, is large, the parabola is steep-sided, so that small changes in β away from $\hat{\beta}$ produce large changes in the discrepancy or deviance. In other words, β is well determined by the data. By contrast, if the Fisher information for β is small, the parabola is rather flat and β is not well-determined by the data.

The Fisher information for β is the ratio of two quantities. The numerator depends only on the model matrix, i.e. on the values of the covariates in the model, and not at all on the response values. The denominator depends only on the error variance of the response. The inverse information gives the theoretical sampling variance of the estimate $\hat{\beta}$, i.e. $\operatorname{var}\hat{\beta} = \sigma^2/(\mathbf{x}^T\mathbf{x})$. Ordinarily, σ^2 is unknown and an estimate is required, either from replicate observations for the same x, or from the residual sum of squares

after fitting an adequate model. The usual unbiased estimate is

$$\tilde{\sigma}^2 = s^2 = D_{\min}/(n-p)$$

where p is the number of covariates in the model and $D_{\min}$ is the minimized discrepancy or deviance.

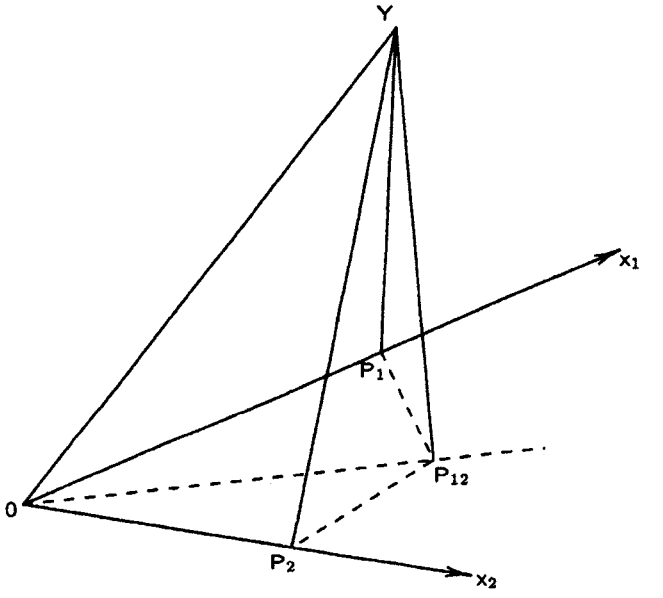

Fig. 3.5. *Least squares: the geometry for two positively correlated covariates.*

3.6.4 *A model with two covariates*

If there are two covariates $\mathbf{x}_1$ and $\mathbf{x}_2$, say, then the solution locus is the plane in R^n defined by the points $\mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$ for varying values of β_1 and β_2. The process of obtaining fitted values for the model

$$\boldsymbol{\eta} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$$

is represented geometrically by the dropping of a perpendicular from the data point $\mathbf{y}$ onto the $(\mathbf{x}_1, \mathbf{x}_2)$ plane. Figures 3.5, 3.6 and

3.7 show the geometry connecting the fits of the single-term models $\mathbf{x}_1\beta_1$ and $\mathbf{x}_2\beta_2$ with the model containing both covariates. In these diagrams, $\mathbf{x}_1$ and $\mathbf{x}_2$ are respectively positively correlated (i.e. make an acute angle), negatively correlated (i.e. make an obtuse angle), and uncorrelated (i.e. make a right angle).

The points P_1, P_2 and P_{12} are respectively the feet of the perpendiculars from $\mathbf{y}$ onto the $\mathbf{x}_1$-line, the $\mathbf{x}_2$-line and the $(\mathbf{x}_1, \mathbf{x}_2)$-plane. The angles $\widehat{OP_1}P_{12}$ and $\widehat{OP_2}P_{12}$, are both right angles because $\widehat{OP_2}Y$, $\widehat{OP_{12}}Y$ and $P_2\widehat{P_{12}}Y$ are all right angles by definition. Consequently P_1 is also the projection of P_{12} onto the $\mathbf{x}_1$-line. Similarly, P_2 is the projection of P_{12} onto the $\mathbf{x}_2$-line. We can thus express the projection of $\mathbf{y}$ on the $(\mathbf{x}_1, \mathbf{x}_2)$-plane in the two forms

$$(OP_{12})^2 = (OP_1)^2 + (P_1P_{12})^2 = (OP_2)^2 + (P_2P_{12})^2.$$

The interpretation of these squared lengths is as follows:

$$\begin{aligned}(OP_1)^2 &= \text{sum of squares for } x_1 \text{ before } x_2,\\ (OP_2)^2 &= \text{sum of squares for } x_2 \text{ before } x_1,\\ (OP_{12})^2 &= \text{sum of squares for } x_1 \text{ and } x_2.\end{aligned}$$

In addition,

$$\begin{aligned}(P_1P_{12})^2 &= \text{sum of squares for } x_2 \text{ adjusted for } x_1,\\ (P_2P_{12})^2 &= \text{sum of squares for } x_1 \text{ adjusted for } x_2,\\ (OY)^2 &= \text{total sum of squares},\\ (YP_{12})^2 &= \text{residual sum of squares after fitting } x_1 \text{ and } x_2.\end{aligned}$$

The words 'before' and 'after' are often replaced by 'ignoring' and 'eliminating' respectively.

Corresponding to the two sequences of fitting we have analyses of variance whose geometrical interpretations are

$$\begin{aligned}(OY)^2 &= (OP_1)^2 + (P_1P_{12})^2 + (P_{12}Y)^2\\ \text{total} &= (x_1 \text{ before } x_2) + (x_2 \text{ after } x_1) + \text{residual},\end{aligned}$$

and

$$\begin{aligned}(OY)^2 &= (OP_2)^2 + (P_2P_{12})^2 + (P_{12}Y)^2\\ \text{total} &= (x_2 \text{ before } x_1) + (x_1 \text{ after } x_2) + \text{residual},\end{aligned}$$

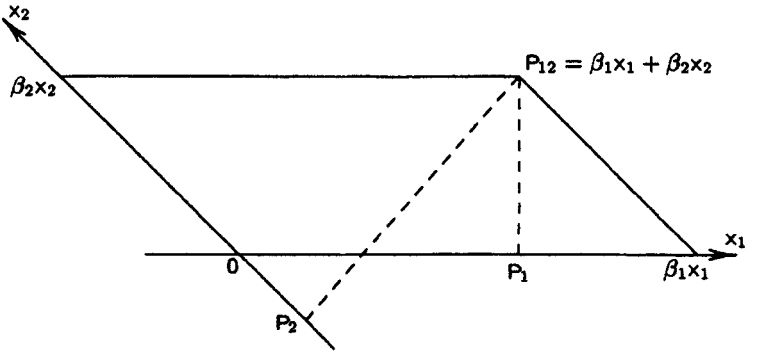

Fig. 3.6. *Least squares projections for two negatively correlated covariates:* P_1 *is the projection on* x_1 *alone,* P_2 *is the projection on* x_2 *alone, and* P_{12} *is the projection on the joint space.*

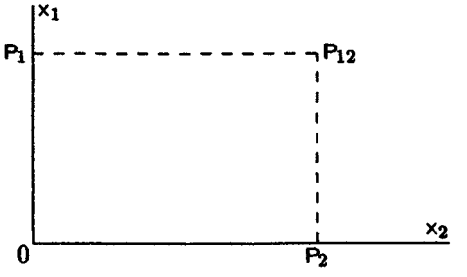

Fig. 3.7. *Least squares projections for two orthogonal covariates.*

In terms of the parameter estimates we have

$$(OP_1) = \mathbf{x}_1 b_1,$$
$$(OP_2) = \mathbf{x}_2 b_2,$$
$$(OP_{12}) = \mathbf{x}_1 \hat{\beta}_1 + \mathbf{x}_2 \hat{\beta}_2,$$

where b_1 and b_2 are the estimates for the single-term models and $\hat{\beta}_1$ and $\hat{\beta}_2$ for the joint model.

There are several important special cases:

1. $\mathbf{y}$ is coplanar with $(\mathbf{x}_1, \mathbf{x}_2)$, so that Y and P_{12} coincide. The residual vector is then null and the joint model gives a perfect fit.

2. $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal (Fig. 3.7). The three feet of the perpendiculars, P_1, P_2 and P_{12} form a rectangle with O, so that $(OP_1) = (P_2P_{12})$ and $(OP_2) = (P_1P_{12})$. The order of fitting the terms in the joint model is then irrelevant, and there is just one analysis of variance. Sums of squares and parameter estimates are unaffected by the order in which terms are entered into the model.
3. $\mathbf{y}$ is orthogonal to $\mathbf{x}_1$. Then b_1 is zero for a single-term model, but the estimate $\hat{\beta}_1$ in the joint model is not zero unless $\mathbf{x}_2$ is orthogonal to $\mathbf{x}_1$ or to $\mathbf{y}$.

3.6.5 *The information surface*

If P is an arbitrary point $\mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$ on the solution locus, then from the relation among the total sum of squares, the residual sum of squares and the regression sum of squares,

$$(YP)^2 = (YP_{12})^2 + (P_{12}P)^2,$$

we obtain

$$\begin{aligned}&(\mathbf{y} - \mathbf{x}_1\beta_1 - \mathbf{x}_2\beta_2)^T(\mathbf{y} - \mathbf{x}_1\beta_1 - \mathbf{x}_2\beta_2) = \\ &(\mathbf{y} - \mathbf{x}_1\hat{\beta}_1 - \mathbf{x}_2\hat{\beta}_2)^T(\mathbf{y} - \mathbf{x}_1\hat{\beta}_1 - \mathbf{x}_2\hat{\beta}_2) \\ &+ (\hat{\beta}_1 - \beta_1)^2\mathbf{x}_1^T\mathbf{x}_1 + 2(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)\mathbf{x}_1^T\mathbf{x}_2 + (\hat{\beta}_2 - \beta_2)^2\mathbf{x}_2^T\mathbf{x}_2.\end{aligned}$$

Note that $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimates in the joint fit of $\mathbf{x}_1$ and $\mathbf{x}_2$ simultaneously. The first term on the right of the above equation is the residual sum of squares from the least squares fit to both covariates. This term does not depend on (β_1, β_2), but only on $\mathbf{y}$. The second term measures the squared distance of the arbitrary point P, determined by (β_1, β_2), from the point of best fit, $(\hat{\beta}_1, \hat{\beta}_2)$. The contours of this latter term, considered as a function of (β_1, β_2), are similar, similarly situated ellipses centered at $(\hat{\beta}_1, \hat{\beta}_2)$ as shown in Fig. 3.8.

The second derivative matrix of the function with respect to (β_1, β_2) is

$$2\begin{pmatrix}\mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2\end{pmatrix},$$

which, apart from the factor $\sigma^2/2$, is the Fisher information matrix for (β_1, β_2).

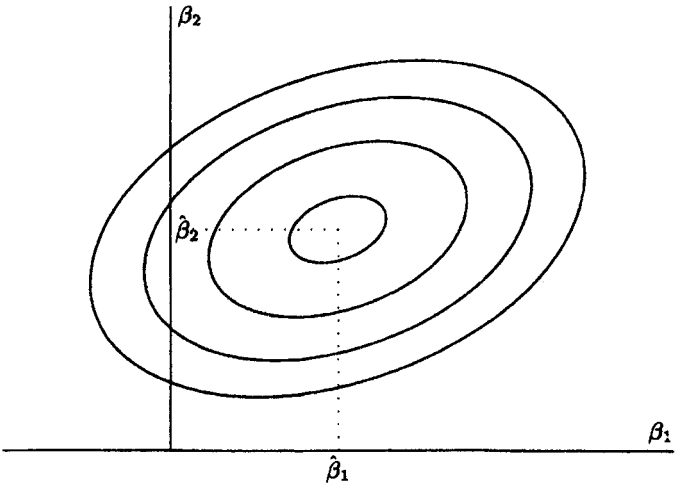

Fig. 3.8. *Least squares: contours of the information surface for two parameters.*

3.6.6 *Stability*

The point P_{12} depends only on the data vector $\mathbf{y}$ and its relation to the space spanned by $\mathbf{x}_1, \mathbf{x}_2$. Any pair of vectors that spans the same space, for example $\mathbf{x}_1 + \mathbf{x}_2$ and $\mathbf{x}_1 - \mathbf{x}_2$, gives rise to the same projection of $\mathbf{y}$. However the identification of the point P_{12} by the coefficients $(\hat{\beta}_1, \hat{\beta}_2)$, namely $P_{12} = \mathbf{x}_1\hat{\beta}_1 + \mathbf{x}_2\hat{\beta}_2$, depends heavily on the particular pair of vectors chosen as a basis for the subspace. If θ, the angle between the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in R^n, is small, then the coefficients $\hat{\beta}_1, \hat{\beta}_2$ are more sensitive to small perturbations of the data than either b_1 or b_2, the coefficients in the single-term models. That is to say P_{12} itself is unstable in the sense that while small perturbations of $\mathbf{x}_1$ and $\mathbf{x}_2$ may produce correspondingly small perturbations of P_1 and P_2, together they may produce a large perturbation of P_{12}. Thus, a small perturbation of one or both covariates may have a big effect on the space spanned by the two vectors. Consequently, the allocation of variation in Y to $\mathbf{x}_1$ and $\mathbf{x}_2$ in the joint regression is sensitive to perturbations of either covariate.

When θ, the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$, is small the information matrix for β_1, β_2 is nearly singular because its determinant is equal

to

$$||\mathbf{x}_1||^2 \, ||\mathbf{x}_2||^2 \sin^2 \theta.$$

The log-likelihood contours in the (β_1, β_2) plane are now ellipses having one principal axis very long compared to the other. In fact the ratio of the lengths of the principal axes behaves like $1/\sin^2 \theta$. As a consequence, large changes in β_1, β_2 in the direction of the longer axis produce small changes in the likelihood, while similar changes in the perpendicular direction have a large effect.

3.7 Tables as data

It is common to find generalized linear models being fitted not to the original data expressed in data-matrix form, but to data that have already been summarized in the form of a multi-way table. In the process of tabulation, the y-values for units having the same levels of the classifying factors are added together to form a table of totals: parallel tabulation of a vector of 1s gives the associated table of counts showing how many units contribute to each cell total. Division of totals by their associated counts gives a table of means. For continuous data, it is usually this table of means that will be analysed, with the associated counts acting as prior weights. In surveys, however, it is often the counts themselves that are of interest. Suitable methods of analysis for such counts are described in Chapters 4 to 6. Section 5.2.3 emphasizes the duality between models that treat cell averages or scores as the response with the counts acting as weights, and models that treat the observed count as the response, with the cell averages used for generating contrasts.

Broadly speaking, the process of fitting generalized linear models to data in the form of tables is similar to that described previously for data in the form of a data matrix. The following points are not peculiar to tabular data, but they are most often encountered in that context.

3.7.1 *Empty cells*

When any variate, be it a continuous measurement or an integer-valued variable, is discretized and tabulated as described above, a table of averages and an associated table of counts is formed. The table of averages is different in one important respect from the table

of counts, namely in the significance of zeros. Table 8.1, giving the average value of insurance claims together with the number of claims, is a case in point. For some purposes the value of the claims is of interest, while for others the number of claims might be the most interesting response. It so happens that no teenagers who owned ten-year-old cars of types C or D made claims against the company. These are genuine zeros indicating either that such drivers are unusually careful or few in number or both. As far as the average claim is concerned, however, these are not to be treated as zeros, but rather as 'empty cells' contributing no information whatever about averages. We cannot infer from the absence of claims in these categories that, if and when a claim occurs, its value will be small. Consequently, we use the term 'empty cell' rather than 'structural zero' because there is no suggestion of any value, let alone zero.

It is important to distinguish two varieties of empty cell, namely *necessarily empty cells* and *accidentally empty cells*. Necessarily empty cells occur when some combination of levels of the factors is a priori impossible. Simple examples are the class of pregnant males or a self-fertilized cross from a self-sterile variety of plant. When all possible crosses are made between varieties of a self-sterile species, the diagonal cells, corresponding to the selfs, are all necessarily empty. If some varieties are cross-incompatible there may be off-diagonal necessarily empty cells as well. When a model is fitted to a table of associated counts, the necessarily empty cells must not be included as data. For other tables such as Table 8.1, they cannot be included in any analysis because there is no value for that cell. Ordinarily it makes no sense to compute fitted values for necessarily empty cells by extrapolation from the non-empty cells.

An accidentally empty cell is one for which the combination of factor levels is possible, but the combination happens not to occur in the observed data. The empty cells in Table 8.1 are of this type. For this type of empty cell it does usually make sense to compute fitted values by extrapolation from the non-empty cells.

Table 6.2 contains both accidentally empty and necessarily empty cells.

It has been proposed (see, for example, Urquhart and Weeks, 1978) that models fitted to tables should not involve the population means of accidentally empty cells, on the grounds that the data give no information about such means. This proposal would imply that

an additive model of the form

$$A + B$$

for a two-way table should not be fitted if any cells are accidentally empty. The point has been discussed by Nelder (1982), who argues that such a rule is unnecessarily restrictive.

3.7.2 *Fused cells*

It sometimes happens that the position of a unit in a table is not known uniquely, though it is known to belong to one of a subset of cells. Examples of this phenomenon occur in tables classified by genetic factors when several distinct genotypes produce the same phenotype, which is what is observed. For the analysis we have just the total for the set of cells, fused into a single observable cell. Fused cells may also occur when the individual cells were potentially observable, but for some reason the level of one or more factors was recorded with less precision than intended. The occurrence of fused cells results in an obvious loss of information, and utilization of the data they contain may require prior knowledge of the relative frequency of occurrence in the unobserved component cells. Such knowledge is often available for genetic data.

3.8 Algorithms for least squares

For the linear models discussed in this chapter the estimation procedure requires us to minimize the quadratic form

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

with respect to the components of $\boldsymbol{\beta}$. Equating the derivative to zero produces the *normal equations*

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}. \tag{3.9}$$

If $\mathbf{X}$ has full rank these equations have a unique solution, namely $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. If $\mathbf{X}$ is rank-deficient, either because of intrinsic aliasing among factors or for some other reason, we may

replace the inverse of $\mathbf{X}^T\mathbf{X}$ by any generalized inverse. The solution is then not unique, but all estimable contrasts among the βs are independent of the choice of inverse (Pringle and Rayner, 1971). When multiplied by the dispersion parameter σ^2, the generalized inverse also produces correct variances and covariances for these contrasts.

There are two classes of numerical methods for solving equations (3.9). In the first method $\mathbf{X}^T\mathbf{X}$ is formed explicitly and subsequent computations are performed on this matrix. The second class of methods focuses on the matrix $\mathbf{X}$ and attempts to simplify equations (3.9) by suitably factoring $\mathbf{X}$. In both cases it is usual to express both $\mathbf{y}$ and the columns of $\mathbf{X}$ about their means. Within each class there are further sub-divisions, which are described below.

In the interests of efficient bookkeeping, both algebraic and numerical, it is convenient in much of the discussion that follows to imagine the observation vector $\mathbf{y}$ appended as an additional column to $\mathbf{X}$. Thus any row operations applied to $\mathbf{X}$ are considered also to be applied to $\mathbf{y}$. In addition, the extended information matrix $\mathbf{X}^T\mathbf{X}$ now consists of the sums of squares and products of $\mathbf{y}$ and the p covariates.

3.8.1 *Methods based on the information matrix*

The two most common methods that operate on the matrix $\mathbf{X}^T\mathbf{X}$ are Gaussian elimination and Choleski decomposition. We discuss these in turn.

A modern form of Gaussian elimination, due to Beaton (1964), uses a symmetric sweep operator. This operator, when applied to the kth row and column of a positive-definite symmetric matrix $\mathbf{A}$, will be denoted by $\mathcal{S}_k$. The effect of $\mathcal{S}_k$ is to transform the components of $\mathbf{A}$ from a_{ij} to

$$
\begin{aligned}
a_{ij} &\to a_{ij} - \frac{a_{ik}a_{jk}}{a_{kk}}; && i \neq k,\ j \neq k,\\
a_{ik} &\to \frac{a_{ik}}{|a_{kk}|}; && i \neq k,\\
a_{kj} &\to \frac{a_{kj}}{|a_{kk}|}; && j \neq k,\\
a_{kk} &\to -\frac{1}{a_{kk}}.
\end{aligned}
$$

With this definition it is then easily shown that $\mathcal{S}_k\mathcal{S}_k\mathbf{A} = \mathbf{A}$. In other words, a second application of the symmetric sweep restores the original matrix. The statistical interpretation of the symmetric sweep is as follows. Let $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ be a $p \times p$ matrix of sums of squares and products of the variates $\mathbf{x}_1, \ldots, \mathbf{x}_p$. Suppose that the sweeps $\mathcal{S}_1, \ldots, \mathcal{S}_k$ have been applied to the first $k < p$ rows and columns of $\mathbf{A}$. Following this series of sweeps, $\mathbf{A}$ has been reduced to the form shown in Fig. 3.9, in which only the lower triangle is displayed. The component matrix $\mathbf{R}$ now holds the residual sum-of-squares-and-products matrix for the unswept variates $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_p$ after regressing them on $\mathbf{x}_1, \ldots, \mathbf{x}_k$. The rows of the matrix $\mathbf{B}$ are the regression coefficients of these formal linear regression equations, while $\mathbf{V}$ is the unscaled covariance matrix for these regressions.

Note that if the final row and column of $\mathbf{A}$ contain the sums of squares and products of the response, then sweeping all but the final row and column gives minus the inverse information matrix $-\mathbf{V}$, bordered by the vector of regression coefficients $\mathbf{B} = \hat{\boldsymbol{\beta}}$, and the residual sum of squares, $\mathbf{R}$, now a scalar.

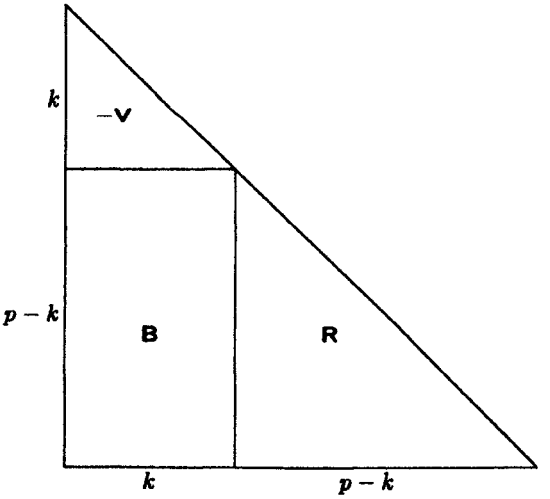

Fig. 3.9. *The matrix of sums and squares and products after the symmetric sweep has been applied to the first k rows and columns.*

If the original $\mathbf{X}^T\mathbf{X}$ is exactly or nearly singular, this will usually

show up during the sweeping process by the appearance of a pivot or diagonal element that is small compared to its original value (Clarke, 1982). For if $\mathbf{x}_{k+1}$ is expressible as a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, then the residual sum of squares for $\mathbf{x}_{k+1}$ after linear regression on $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is zero. If there is a near singularity then the residual sum of squares for $\mathbf{x}_{k+1}$ is small compared to its original total sum of squares. Statistically, this exact collinearity or near singularity means that there is either no information or little information about the corresponding parameter, given that the first k terms are included in the model. If there is an exact singularity, the term may be omitted from the model without affecting the span of $\mathbf{X}$. Algebraically, this is equivalent to setting the estimate to zero with variance zero. In the algorithm such rows/columns are not swept, but are marked to show their special status. If such a term is subsequently to be removed from the model, then again no sweep is done; however, if other terms involved in the collinearity are subsequently removed, the pivot for the first term may again become substantial. Should this occur, the term can again be included in the model and a reliable estimate of the parameter obtained.

The second method that operates on the information matrix is the Choleski decomposition, which aims to find a lower-triangular $p \times p$ matrix $\mathbf{L}$ that satisfies

$$\mathbf{X}^T\mathbf{X} = \mathbf{L}\mathbf{L}^T.$$

$\mathbf{L}$ is thus a square-root matrix of $\mathbf{X}^T\mathbf{X}$. Details of algorithms for computing $\mathbf{L}$ can be found in the books by Chambers (1977) and Healy (1986). Having computed $\mathbf{L}$, the inversion of $\mathbf{X}^T\mathbf{X}$ is accomplished via the formula

$$(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{L}^{-1})^T\mathbf{L}^{-1}.$$

There is a simple inversion algorithm for triangular matrices, and the inversion can be combined with subsequent multiplication by the transpose. Again, generalized inverses can be obtained by setting any row of $\mathbf{L}$ with a small pivot to zero.

The *condition number* of a matrix is a measure of closeness to singularity, large values indicating near-singularity; algorithms that use the matrix $\mathbf{X}^T\mathbf{X}$ directly suffer from the disadvantage

that the condition number of $\mathbf{X}^T\mathbf{X}$ is the square of the condition number of $\mathbf{X}$. Large values of the condition number can give rise to numerical instability from rounding errors in the calculations. For this reason the second class of methods is designed to avoid the formation of $\mathbf{X}^T\mathbf{X}$ altogether.

3.8.2 *Direct decomposition methods*

Direct decomposition methods operate on the model matrix $\mathbf{X}$ directly. The aim is to decompose $\mathbf{X}$ into the product of an $n\times n$ orthogonal matrix $\mathbf{Q}$ and an $n\times p$ matrix $\mathbf{R}$ of the form

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{R}_1$ is $p\times p$ upper triangular.

The statistical interpretation of the decomposition is as follows. If $\mathbf{y}$ is the observation vector with mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}$, we may make an orthogonal transformation to new variables $\mathbf{u}$ defined by $\mathbf{u} = \mathbf{Q}^T\mathbf{y}$, where $\mathbf{Q}$ is the $n\times n$ orthogonal matrix described above. The mean and variance of the new variables are

$$\begin{aligned} E(\mathbf{U}) = \mathbf{Q}^T E(\mathbf{Y}) &= \mathbf{Q}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{Q}^T\mathbf{Q}\mathbf{R}\boldsymbol{\beta} \\ &= \mathbf{R}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{R}_1\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \\ \operatorname{cov}(\mathbf{U}) &= \mathbf{Q}^T\mathbf{I}\mathbf{Q}\sigma^2 = \mathbf{I}\sigma^2. \end{aligned}$$

Thus the last $(n-p)$ components of $\mathbf{U}$ have zero expectation, and so give no information about $\boldsymbol{\beta}$. Hence the least-squares solution reduces to equating the first p components of $\mathbf{u}$, here denoted by $\mathbf{u}_1$, to their expectation as a function of $\hat{\boldsymbol{\beta}}$. Thus we arrive at

$$\mathbf{R}_1\hat{\boldsymbol{\beta}} = \mathbf{u}_1$$

which is easily solved because $\mathbf{R}_1$ is upper triangular.

It is not necessary to compute $\mathbf{Q}$ explicitly because, if $\mathbf{y}$ is appended to $\mathbf{X}$, the sequence of operations that takes $\mathbf{X}$ to $\mathbf{R}$ also transforms $\{\mathbf{X} : \mathbf{y}\}$ to $\{\mathbf{R} : \mathbf{u}\}$. The first p rows of this augmented matrix give the coefficients in the above equation for $\hat{\boldsymbol{\beta}}$. The sum of squares of the last $n-p$ components of $\mathbf{u}$ gives the residual sum of squares.

Note that

$$\mathbf{R}_1^T\mathbf{R}_1 = \mathbf{R}^T\mathbf{R} = \mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R} = \mathbf{X}^T\mathbf{X},$$

so that $\mathbf{R}_1$ is the upper triangular Choleski square-root matrix of $\mathbf{X}^T\mathbf{X}$. In fact $\mathbf{R}_1$ is the transpose of $\mathbf{L}$ as described in the previous section.

Three methods for finding $\mathbf{Q}$ and $\mathbf{R}$ are associated with the names of

Householder:	$\mathbf{Q}$ is a product of reflections,
Givens:	$\mathbf{Q}$ is a product of rotations,
and Gram-Schmidt:	successive orthogonalization.

A Householder reflection takes the matrix form

$$\mathbf{I} - 2\mathbf{v}\mathbf{v}^T,$$

where $\mathbf{v}$ is an n-vector of unit length ($\mathbf{v}^T\mathbf{v} = 1$). It is possible, given a vector $\mathbf{x}$, to choose $\mathbf{v}$ so that, after reflection, all components of $\mathbf{x}$ except the first are zero. In the Householder decomposition, $\mathbf{v}_1$ is chosen to reduce the first column of $\mathbf{X}$ to this form. A second vector $\mathbf{v}_2$ is chosen to reduce components 3 to n of the second column to zero. Since elements 2 to n of the first column are already zero, this reflection leaves them unaffected. The process continues on components $j+1$ to n of column j for $j = 1, \ldots, p-1$. If $\mathbf{Q}_j = \mathbf{I} - 2\mathbf{v}_j\mathbf{v}_j^T$, then the matrix

$$\mathbf{Q} = \mathbf{Q}_{p-1} \cdots \mathbf{Q}_2\mathbf{Q}_1$$

is the product of $p-1$ reflections. It has the property that

$$\mathbf{Q}^T\mathbf{X} = \mathbf{R},$$

where $\mathbf{R}$ has the form described at the beginning of this section.

Givens rotations are planar rotations through an angle θ. A single rotation is applied to two components of a vector, corresponding to a rotation through an angle θ in the plane of these two components. The angle is chosen to make one of the components equal to zero. We denote by G_{ijk} the rotation that replaces the

ith and jth rows of $\mathbf{X}$ by linear combinations that make the kth element in row j equal to zero. Then the sequence

$$\left((G_{kjk},\ j = k+1 \text{ to } n),\ k = 1 \text{ to } p\right)$$

annihilates the elements by columns, like Householder, but setting one element to zero at a time. The sequence

$$\left((G_{kjk},\ k = 1 \text{ to } \min(j-1,p)),\ j = 2 \text{ to } n\right)$$

annihilates by rows. The latter sequence is more useful if $\mathbf{X}$ is more easily processed by rows than by columns. The idea of rotations goes back to Jacobi, but the Givens sequence of rotations ensures that previously formed zeros remain zero after subsequent rotations.

The Gram-Schmidt method relies on successive orthogonalization of the columns of $\mathbf{X}$. The preferred algorithm is due to Björck (1967) and begins by forming

$$\begin{aligned}\mathbf{q}_1 &= \mathbf{x}_1/||\mathbf{x}_1|| \\ \mathbf{q}_j &= \mathbf{x}_j - (\mathbf{q}_1^T\mathbf{x}_j)\mathbf{q}_1; \qquad j = 2,\ldots,p.\end{aligned}$$

The first row of $\mathbf{R}$ is given by

$$r_{1i} = \mathbf{q}_1^T\mathbf{x}_i.$$

This process is then repeated, using at the second stage the vectors $\mathbf{q}_2,\ldots,\mathbf{q}_p$ in place of $\mathbf{x}_1,\ldots,\mathbf{x}_p$, and so on.

Statistically, we regress columns 2 to p of $\mathbf{x}$ on column 1 and replace them by the vectors of residuals. At the second stage, columns 3 to p are regressed on column 2, and so on for successive stages. The matrix $\mathbf{Q}$ thus formed is $n\times p$ with orthonormal columns and $\mathbf{R}$ is $p\times p$ upper triangular. The first j columns of $\mathbf{Q}$ span the same space as the first j columns of $\mathbf{X}$ for $j = 1,\ldots,p$. Calculating the regression of $\mathbf{y}$ on the orthogonalized covariates is easy because of the orthogonality.

The direct decomposition methods are somewhat less convenient for updating models than the symmetric sweep method. For example, if a column of $\mathbf{X}$ is deleted, all the columns of $\mathbf{Q}$ and $\mathbf{R}$ to the right of the deleted column must be recalculated. Details of updating with the Givens algorithm are given by Clarke (1981).

3.8.3 *Extension to generalized linear models*

In Chapter 2 it was shown that estimation in generalized linear models can be accomplished by iteratively weighted least squares. In order to adapt the algorithms discussed above we must (a) allow iteration and (b) introduce weights and an adjusted dependent variate, both of which ordinarily vary from one iteration to the next. Introduction of weights is straightforward in principle. In the algorithms that use the information matrix we replace $\mathbf{X}^T\mathbf{X}$ by $\mathbf{X}^T\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is the diagonal matrix of weights, while in the QR algorithms we replace $\mathbf{X}$ by $\mathbf{W}^{1/2}\mathbf{X}$. The attractiveness of QR methods and the Choleski decomposition is then greatly reduced because a new decomposition must be computed at each cycle of the iteration.

There are two new features of the algorithms, the first related to unbounded parameter estimates and the second to *pseudo-aliasing*. Infinite parameter estimates arise most commonly in the fitting of a log-linear model if one or more fitted values are zero. The link function $\hat{\eta} = \log \hat{\mu}$ implies that one or more of the $\hat{\beta}$s contributing to $\hat{\eta}$ must be negatively infinite. Cells in the table for which $\hat{\mu} = 0$ must also have $y = 0$, so that there is no contribution to the deviance. Such cells are best omitted from the fit. Similar effects arise in models for proportions where the fitted proportion is either 0 or 1.

The second feature, pseudo-aliasing, can arise when the changing weights in an iterative fit produce so little information on a parameter that the pivot in the information matrix falls below the tolerance set by the algorithm. However, removal of the covariate will now be found to increase the deviance sharply, showing that it is really required in the model. In this respect, pseudo-aliasing is quite different from true aliasing. The fit obtained immediately before the algorithm detected the apparent aliasing is often a reasonable assessment of the fit of the model.

Numerical iteration requires a definition of effective convergence of the process. If all parameter estimates are finite, straightforward monitoring of the progress of the deviance is sufficient. Convergence is usually rapid, though divergence may occasionally occur for ill-fitting models using non-canonical links. If one or more components are infinite, convergence as measured by the deviance may be slow. It is usually best in these circumstances to halt the

iteration after about 10 cycles and to inspect the estimates at that stage. Usually it is clear which components are tending to $\pm\infty$. If necessary, action can then be taken to omit a subset of the data or to modify the model or both.

3.9 Selection of covariates

Apart from the choice of link function and error distribution, the problem of modelling reduces to finding one or more appropriate parsimonious sets of covariates corresponding to a model matrix $\mathbf{X}$ of order $n \times p$. As elsewhere it is important that the final model or models should make sense physically: at a minimum, this usually means that interactions should not be included without main effects nor higher-degree polynomial terms without their lower-degree relatives. Furthermore, if the model is to be used as a summary of the findings of one out of several studies bearing on the same phenomenon, main effects should usually be included whether significant or not. Strict adherence to this policy makes it easier to compare the results of various studies and helps to avoid the apparent conflicts that occur when different fitted models with different sets of terms are used in each study. The danger is that a term with a coefficient of $+1$, say, might be rejected in one study because it was insignificant, while in a second study the same term might have a numerically similar coefficient that was highly significant. The fitted models are then different and apparently in conflict, while in reality the two studies are highly concordant.

We discussed in Chapter 1 the justification for seeking a parsimonious model to represent a set of data. Parsimony implies, among other things, that covariates having no detectable effect on the response should ordinarily be excluded from the linear predictor. In a survey concerned with the incidence of a particular disease, large numbers of covariates may be available, describing perhaps age structures of the populations involved, their dietary and smoking habits, aspects of the environment, and so on. The selection of a useful set of covariates from such a large set of possible covariates to form a parsimonious model is then a non-trivial exercise. There are both statistical and computing problems, the latter arising from the 'combinatorial explosion' that occurs when all possible subsets of covariates are to be tested for inclusion in the model.

On the statistical side, the problem is that of defining the balance to be struck between two opposing effects of including a new term in the model. The good effect may be a reduction in the discrepancy between the data and the fitted values. The bad effect is that, unless there is good prior knowledge that the covariate has a non-negligible influence on the response, inclusion of the covariate usually complicates the model and statements of conclusions derived from it. At one extreme, if the addition of a single covariate reduces the residual mean square to, say, one third of its original value we have no hesitation in including it in the model, particularly if the number of residual degrees of freedom is large. At the other extreme, if such an addition causes no reduction, by the principle of Occam's razor, parsimony wins and we exclude it. It is the intermediate cases that cause problems. For example, if there is a large number of irrelevant covariates, then statistical accidents will produce a few false positives that appear to influence the response.

The usual F-statistic for the reduction in deviance or sum of squares is the basis of most criteria for selection of covariates. In order to exclude irrelevant terms the significance level for acceptance is set at a low level, but it must not be set so low that important terms are thereby excluded. Another approach is based on the idea of providing the best prediction of response values over a set of covariate values, and yet another uses a criterion based on a measure of information. Atkinson (1981b) points out that all of these procedures can be represented (in our notation) as special cases of minimizing the expression

$$Q = D + \alpha q \phi, \tag{3.10}$$

where D is the deviance function, q is the number of estimable parameters in the linear predictor, ϕ is the dispersion parameter, and α is either constant or a function of n. The idea behind the second term is to penalize the inclusion of unnecessary covariates in the model. Use of Q presumes a knowledge of ϕ. For Poisson and binomial models without over-dispersion, $\phi = 1$, but otherwise ϕ is usually unknown. Even with counted data it is often wise to assume that over-dispersion is present unless the data or prior information indicate otherwise. For details see Chapter 4. When comparing a sequence of models we have the option of replacing ϕ

in (3.10) either by a common estimate for all models in the sequence or by separate estimates $\hat{\phi}_i$ derived from the fit of each model in turn. To make the comparison fair, it seems best in practice to use a single estimate, usually derived from the most complex model in the sequence.

If two models in a nested sequence differ only by the inclusion of one covariate, then the use of the 5% point of the F- or t-distribution as the criterion for model selection is equivalent to setting $\alpha \simeq 4$ in (3.10), assuming adequate residual degrees of freedom for estimating ϕ. The most common criteria based on errors of prediction (Akaike, 1969; Mallows, 1973) lead to $\alpha = 2$. For Normal-theory linear models an argument based on maximum posterior probabilities leads to $\alpha(n) = \log(n)$. Atkinson (1981b) suggests that the range $\alpha = 2$ to 6 may provide 'a set of plausible initial models for further analysis'.

The computing problem, which ignores any relationships that may exist among the covariates, may be specified as follows: 'find the best s subsets of size r among the covariates'. If k, the total number of covariates available, is small, say $k \leq 12$, the best subsets for each r from 1 to $k-1$ can be found by complete enumeration. For larger k, say up to 35, tree-search methods, using short-cuts, are feasible (Furnival and Wilson, 1974). Approximate methods for generating a single 'optimum' subset include:

1. *forward selection*, whereby at each stage the best unselected covariate satisfying the selection criterion is added until no further candidates remain;
2. *backward elimination*, which begins with the full set and eliminates the worst covariates one by one until all remaining covariates are necessary; and
3. *stepwise regression* (Efroymson, 1960), which combines the two previous procedures, following backward elimination by forward selection until both fail to change the model.

In GLIMPSE (Wolstenholme, O'Brien and Nelder, 1988), a knowledge-based front-end for GLIM (Payne 1986), a model selection strategy is used that results in a tree of candidate models, with the extreme node of each branch forming a possible parsimonious model. The basic step in the algorithm has as input a *kernel*, which contains terms already accepted as necessary, and a set of *free terms*, whose status is currently uncertain. The maximal model

contains the kernel and all the free terms. For each free term two F-statistics are calculated, a forward F-statistic formed by adding it to the kernel, and a backwards F-statistic formed by removing it from the maximal model. The two F-statistics are classified by a decision rule as being either large or small, and action is then taken as shown in the table below.

Forward F	*Backward F*	*Action*
large	*large*	*add term to kernel*
large	*small*	*leave as free term*
small	*large*	*leave as free term*
small	*small*	*discard term*

The process is begun with a kernel of terms considered necessary a priori and continues until the set of free terms is either null or unchanging. If it is null we have a unique preferred model; if not we add each remaining free term in turn to the kernel, producing a branching in the tree and repeat the basic step. Further branching may then occur, but eventually the final node on each branch will contain a null set of free terms.

Unthinking use of automatic selection procedures has frequently, and rightly, been criticized. Clearly the notion that a particular subset is optimum is hard to sustain when many other subsets of similar size produce almost equally good fits. It may also happen that some covariates are much more expensive to measure than others, and this is not allowed for in a criterion based on purely statistical considerations. Expense may be an important consideration if the goal is to produce good forecasts at reasonable cost. However, if the goal is to understand the mechanism by which the process is generated, cost is largely irrelevant. A further criticism of automatic selection procedures is that they do not take into account the marginality constraints among factors nor functional marginality among polynomial terms (Section 3.5). Further, certain factors, e.g. treatment and block effects, would often be kept in the model whether statistically significant or not.

Further modification of these selection procedures is required for models that require iterative solution, because of the presence of weights and adjusted dependent variates, both of which are functions of the fitted values, and so change as the fitted model changes. The amount of computing is reduced by using an approximate

procedure, which appears to work well in practice; this involves doing the full iterative fit for a large but well-fitting model, and afterwards following the same algorithms as for the non-iterative case. In other words, the weights and adjusted dependent variate are kept fixed throughout. The fully iterated fit may then be recalculated at intervals as a check on the approximation.

3.10 Bibliographic notes

For a history of least squares, see a series of papers by Harter, summarized in Harter (1976).

Among the many texts on linear models, see Atkinson, (1985), Draper and Smith (1981), Mosteller and Tukey (1977), Plackett (1960), Searle (1971), Seber (1977), Sprent (1969) and Williams (1959).

Model formulae for linear predictors were introduced by Nelder (1965a,b) and developed by Wilkinson and Rogers (1973).

Aliasing, marginality and the role of constraints are discussed by Nelder (1977).

For numerical methods for least squares, see Lawson and Hanson (1974), Gentleman (1974a,b), Healy (1986), Chambers (1977), Thisted (1988) and Wampler (1979).

The statistical problems of covariate selection are discussed by Akaike (1973), Mallows (1973), Stone (1977), and summarized by Atkinson (1981b). For a discussion of the computing aspects of covariate selection, see Efroymson (1960), Beale (1970), Stewart (1973), Furnival and Wilson (1974) and Jennrich (1977). Lawless and Singhal (1978) deal explicitly with generalized linear models.

3.11 Further results and exercises 3

3.1 Use the following data to familiarize yourself with a suitable linear regression program (S, GLIM or Minitab should be fine).

1. Plot y against x_1. Comment on any strong relationships or unusual features of the plot.
2. Plot y against x_2. Comment on any strong relationships or unusual features of the plot.

x_1	x_2	y	x_1	x_2	y
2.23	9.66	12.37	3.04	7.71	12.86
2.57	8.94	12.66	3.26	5.11	10.84
3.87	4.40	12.00	3.39	5.05	11.20
3.10	6.64	11.93	2.35	8.51	11.56
3.39	4.91	11.06	2.76	6.59	10.83
2.83	8.52	13.03	3.90	4.90	12.63
3.02	8.04	13.13	3.15	6.96	12.46
2.14	9.05	11.44			

3. Plot x_1 against x_2. Comment on any strong relationships or unusual features of the plot.
4. Regress y on x_1. Plot the residuals against x_2.
5. Regress y on x_2. Plot the residuals against x_1.
6. Regress y on x_1 and x_2 simultaneously. Compare the coefficients obtained in the joint regression with those in the marginal regressions. Compare the (multiple) correlation coefficients.

[Hamilton, 1987].

3.2 Write out explicitly the model matrix corresponding to a randomized blocks design with three treatments in each of four blocks.

3.3 Suppose that the three treatments mentioned in the previous exercise actually denote increasing concentrations of a chemical used for weed control. Re-parameterize the model using linear and quadratic treatment contrasts. Write out the corresponding model matrix.

3.4 Suppose that two factors A and B with levels i and j respectively have the property that observations are possible only when $i \geq j$. Now define a new factor C with levels $k = i - j + 1$. Assuming that A and B have five levels each, and that each possible combination is observed once, answer the following:

1. Which, if any, of the following models are equivalent:

$$A + B, \qquad A + C, \qquad B + C \quad \text{and} \quad A + B + C?$$

2. What are the ranks of the model matrices in part 1?
3. Answer part 1 assuming instead that A, B and C are quantitative covariates taking *values* i, j and k respectively.

It may be helpful to construct the required factors and variates on the computer and to fit the models to computer-generated data.

For an example of such a triangular arrangement of factors, see Cox and Snell (1981, p. 58).

3.5 Suppose that $\mathbf{x}_1$ and $\mathbf{x}_2$ are positively correlated variates in a two-variable linear regression model that includes the intercept. Show that the regression coefficients $\hat{\beta}_1, \hat{\beta}_2$ are negatively correlated. Express the statistical correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of the angle between the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in R^p.

3.6 In section 3.6.6 an expression is given for the determinant of the Fisher information matrix in a two-variable regression model with no intercept. Derive the corresponding expression when the intercept is included.

3.7 Show that the sweep operator, as defined in section 3.8.1, is self-inverse. What advantages accrue from this property?

3.8 The sweep operator has the property that a new variable can be added to an existing regression equation with a single sweep. This produces automatically the updated parameter estimates, the new residual sum of squares and the Fisher information matrix with minimal computational effort. Yet few statistical programs make full use of this property. Discuss briefly the organizational difficulties involved in making full use of the sweep algorithm on an interactive computer system.

3.9 Suppose that A is a factor with 4 levels whose effects are denoted by α_r. Write out explicitly the model matrix $\mathbf{X}$, of order 6×4, corresponding to the algebraic expression

$$\eta_{rs} = \alpha_r - \alpha_s, \quad \text{for} \quad r < s.$$

You may assume that all 6 combinations $1 \leq r < s \leq 4$ are observed. What is the rank of $\mathbf{X}$? Does the constant vector lie in the column space of $\mathbf{X}$? Under what circumstances might such a model formula arise?

3.10 Let A, B, C, D be four factors each with four levels, having the exclusion property that no two factors can simultaneously have the same level. There are thus only 4! possible factor combinations

instead of the more usual $4^4 = 256$ unrestricted combinations. What are the ranks of the models

$$A + B + C + D, \qquad A + B + C, \quad \text{and} \quad (A + B + C + D) ** 2 ?$$

You may assume that all 4! permutations are observed.

For what purposes might such models be used?

Table 3.1 *Ascorbic acid concentrations of samples of snap-beans after a period of cold storage.*

	Weeks of storage				
Temp. °F	2	4	6	8	*Total*
0	45	47	46	46	184
10	45	43	41	37	166
20	34	28	21	16	99
Total	124	118	108	99	449

3.11 The data in Table 3.1, taken from Snedecor and Cochran (1967, p.354), were obtained as part of an experiment to determine the effects of temperature and storage time on the loss of ascorbic acid in snap-beans. The beans were all harvested under uniform conditions at the Iowa Agricultural Experiment Station before eight o'clock one morning. They were prepared and quick-frozen before noon the same day. Three packages were assigned at random to each temperature and storage-time combination. The sum of the three ascorbic acid determinations is shown in the Table.

Suppose for the purpose of model construction that the ascorbic acid concentration decays exponentially fast, with a decay rate that is temperature-dependent. In other words, for a given storage temperature T, the expected concentration after time t (measured in weeks) is $\mu = E(Y) = \exp\{\alpha - \beta_T t\}$. The initial concentration, $\exp(\alpha)$ is assumed in this model to be independent of the storage temperature. Express the above theory as a generalized linear model, treating temperature as a factor and storage time as a variate.

[The above model is unusual in that it contains an interaction between time and temperature, but no main effect of temperature. By design, the concentrations are equal at time zero.]

Estimate the times taken at each of the three temperatures for the ascorbic acid concentration to be reduced to 50% of its original value. Consider carefully how you might construct confidence intervals for this half-life.

Compare your analysis with the factorial decomposition model, using orthogonal polynomial contrasts, as described by Snedecor and Cochran (1967, pp. 354–8).

The mean squared error for individual packets, obtained from the replicates, was 0.706 on 24 degrees of freedom. Is this value consistent with the above analyses?