7
ANALYSIS OF VARIANCE AND COVARIANCE

7.1 Introduction

This chapter concerns linear models of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad \text{with} \quad \mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

where $\mathbf{y}$ and $\mathbf{e}$ are random vectors of length N, $\mathbf{X}$ is an $N \times p$ matrix of constants, $\boldsymbol{\beta}$ is a vector of p parameters and $\mathbf{I}$ is the unit matrix. These models differ from the regression models of the previous chapter in that $\mathbf{X}$, called the *design matrix*, consists entirely of dummy variables for analysis of variance (ANOVA) or dummy variables and measured covariates for analysis of covariance (ANCOVA). Since the choice of dummy variables is to some extent arbitrary, a major consideration is the optimal choice of $\mathbf{X}$. The main questions addressed by analysis of variance and covariance involve comparisons of means. Traditionally the emphasis is on hypothesis testing rather than estimation or prediction.

In this book we only consider *fixed effects models* in which the levels of factors are regarded as fixed so that $\boldsymbol{\beta}$ is a vector of constants. We do not consider *random effects models* where the factor levels are regarded as a random selection from a population of possible levels and $\boldsymbol{\beta}$ is treated as a vector of random variables. The problem of estimating variances for the elements of $\boldsymbol{\beta}$ in random effects models, also called *variance components* models, is discussed by McCullagh and Nelder (1983) in the framework of generalized linear models. Also the elements of the response vector $\mathbf{y}$ are assumed to be independent and therefore we do not consider situations involving *repeated measures* on the same experimental units so that the observations are likely to be correlated.

Wider coverage of analysis of variance and covariance is provided by any of the conventional books on the subject, for example Graybill (1976), Searle (1971), Scheffé (1959) or Winer (1971).

A. J. Dobson, *Introduction to Statistical Modelling*

7.2 Basic results

Since the random components **e** in ANOVA and ANCOVA models are assumed to be Normally distributed many of the results obtained in the previous chapter apply here too. For instance the log-likelihood function is

$$l = -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{N}{2}\log(2\pi\sigma^2)$$

so the maximum likelihood (or least squares) estimator **b** is the solution of the normal equations

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\,\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{7.1}$$

In ANOVA models there are usually more parameters than there are independent equations in $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ therefore $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is singular and there is no unique solution of (7.1). In this case $\boldsymbol{\beta}$ is said to not be *estimable* or *identifiable*. To obtain a particular solution extra equations are used so that **b** is the solution of

$$\left.\begin{aligned}\mathbf{X}^{\mathrm{T}}\mathbf{X}\,\mathbf{b} &= \mathbf{X}^{\mathrm{T}}\mathbf{y}\\ \mathbf{C}\,\mathbf{b} &= \mathbf{0}\end{aligned}\right\}. \tag{7.2}$$

and

In anticipation of the need for the extra equations $\mathbf{Cb} = \mathbf{0}$, the model $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ often includes the *constraint equations* $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. However, the minimum of $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ is unique and it is given by any solution of (7.1) (see Exercise 7.4), so the value of $(\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb})$ does not depend on the choice of the constraint equations. Other properties of **b** do depend on the choice of **C** as illustrated in the numerical examples in Sections 7.3 and 7.4.

For a maximal model $\boldsymbol{\beta}_{\max}$ has N elements $[\beta_1, \beta_2, \ldots, \beta_N]^{\mathrm{T}}$. Therefore, without loss of generality, we can take **X** to be the $N \times N$ unit matrix **I** so that $\mathbf{b}_{\max} = \mathbf{y}$ and hence

$$l(\mathbf{b}_{\max}; \mathbf{y}) = -\frac{N}{2}\log(2\pi\sigma^2).$$

For any other model with p parameters and estimator **b**, the log-likelihood ratio statistic is

$$D = 2[l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})] = \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb}) = \frac{1}{\sigma^2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}). \tag{7.3}$$

If the model is correct $D \sim \chi^2_{N-p}$, otherwise D has a non-central chi-squared distribution. As with regression models D is not completely determined when σ^2 is unknown so that hypotheses are tested by comparing appropriate ratios of log-likelihood ratio statistics and using the F-distribution.

7.3 One factor ANOVA

The data in Table 7.1 are an extension of the plant weight example of Chapter 2. An experiment is conducted to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions. Thus the response, plant weight, depends on one factor, growing condition, with three levels – control, treatment A and treatment B. We are interested in whether response means differ between the three groups.

Table 7.1 Plant weights from three different growing conditions.

Control	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Treatment A	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
Treatment B	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26

More generally, if experimental units are randomly allocated to groups corresponding to J levels of a factor, this is called a *completely randomized experimental design* and the data can be set out as in Table 7.2.

Table 7.2 Data for one factor ANOVA with J levels of the factor and unequal sample sizes.

Factor level	Responses	Totals
A_1	Y_{11} Y_{12} ... Y_{1n_1}	$Y_{1\cdot}$
A_2	Y_{21} Y_{22} ... Y_{2n_2}	$Y_{2\cdot}$
$\vdots$		
A_J	Y_{J1} Y_{J2} ... Y_{Jn_J}	$Y_{J\cdot}$

The responses can be written as the vector

$$\mathbf{y} = [Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, \ldots, Y_{Jn_J}]^{\mathrm{T}}$$

of length $N = \sum_{j=1}^{J} n_j$. For simplicity we only consider the case when all the samples are of the same size, i.e. $n_j = K$ for all j so $N = JK$.

We consider three different formulations of the model corresponding to the hypothesis that the response means differ for different levels of the factor. The simplest version of the model is

$$E(Y_{jk}) = \mu_j, \qquad j = 1, \ldots, J. \tag{7.4}$$

In terms of the vector $\mathbf{y}$ this can be written as

$$E(Y_i) = \sum_{j=1}^{J} x_{ij}\mu_j, \qquad i = 1, \ldots, N$$

where $x_{ij} = 1$ if response Y_i corresponds to level A_j and $x_{ij} = 0$ otherwise.

Thus $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$

with $\boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix} 1 & 0 \ldots 0 \\ 0 & 1 \\ \vdots & \mathbf{O} \\ 0 & \end{bmatrix}$

where $\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

are vectors of length K. Then $\mathbf{X}^T\mathbf{X}$ is the $J \times J$ diagonal matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} K & 0 \ldots 0 \\ 0 & K \\ \vdots & \mathbf{O} \\ 0 & \end{bmatrix} \quad \text{and} \quad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} Y_{1.} \\ Y_{2.} \\ \vdots \\ Y_{J.} \end{bmatrix} \quad \text{so that } \mathbf{b} = \frac{1}{K}\begin{bmatrix} Y_{1.} \\ Y_{2.} \\ \vdots \\ Y_{J.} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_J \end{bmatrix}$$

and

$$\mathbf{b}^T\mathbf{X}^T\mathbf{y} = \frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2.$$

In addition, the fitted values are $\hat{\mathbf{y}} = [\bar{y}_1, \bar{y}_1, \ldots, \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_J]^T$. The disadvantage of this simple formulation of the model is that it cannot be extended to more than one factor. For generalizability, we need to specify the model so that parameters for levels and combinations of levels of factors reflect differential effects beyond some average response.

One such formulation is

$$E(Y_{jk}) = \mu + \alpha_j, \qquad j = 1, \ldots, J,$$

where μ is the average effect and α_j is an additional effect due to A_j. For this parameterization there are $J+1$ parameters.

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_J \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \ldots 0 \\ 1 & 0 & 1 \\ \vdots & & \mathbf{O} \\ 1 & & \end{bmatrix}$$

where $\mathbf{0}$ and $\mathbf{1}$ are vectors of length K. Thus

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} Y_{..} \\ Y_{1.} \\ \vdots \\ Y_{J.} \end{bmatrix} \quad \text{and} \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} N & K \ldots K \\ K & K \\ \vdots & \mathbf{O} \\ K & \end{bmatrix}.$$

The first row of the $(J+1) \times (J+1)$ matrix $\mathbf{X}^T\mathbf{X}$ is the sum of the remaining

rows so $\mathbf{X}^T\mathbf{X}$ is singular and there is no unique solution of the normal equations $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$. The general solution can be written as

$$\mathbf{b} = \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_J \end{bmatrix} = \frac{1}{K}\begin{bmatrix} 0 \\ Y_{1.} \\ \vdots \\ Y_{J.} \end{bmatrix} - \lambda \begin{bmatrix} -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

where λ is an arbitrary constant. It is traditional to impose the additional '*sum to zero*' constraint

$$\sum_{j=1}^{J} \hat{\alpha}_j = 0 \qquad \text{so that } \frac{Y_{..}}{K} = J\lambda, \quad \text{i.e. } \lambda = \frac{Y_{..}}{N} \text{ since } N = JK,$$

giving the solution

$$\hat{\mu} = \frac{Y_{..}}{N} \quad \text{and} \quad \hat{\alpha}_j = \frac{Y_{j.}}{K} - \frac{Y_{..}}{N} \qquad \text{for } j = 1, \ldots, J.$$

Hence

$$\mathbf{b}^T\mathbf{X}^T\mathbf{y} = \frac{Y_{..}^2}{N} - \sum_{j=1}^{J} Y_{j.}\left(\frac{Y_{j.}}{K} - \frac{Y_{..}}{N}\right) = \frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2$$

and the fitted values are $\hat{\mathbf{y}} = [\bar{y}_1, \bar{y}_1, \ldots, \bar{y}_J]^T$, as for the first version of the model.

A third version of the model is $E(Y_{jk}) = \mu + \alpha_j$ with the constraint that $\alpha_1 = 0$ so that α_j measures the difference between the first and jth levels of the factor and μ represents the effect of the first level. This is called a *corner-point parameterization*. For this version there are J parameters.

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_J \end{bmatrix}, \qquad \text{also} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & & \\ \vdots & & \mathbf{O} & \\ 1 & \mathbf{O} & \ddots & 1 \end{bmatrix}$$

so $$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} Y_{..} \\ Y_{2.} \\ \vdots \\ Y_{J.} \end{bmatrix} \qquad \text{and} \qquad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} N & K & \ldots & K \\ K & K & & \\ \vdots & & \mathbf{O} & \\ K & \mathbf{O} & \ddots & K \end{bmatrix}.$$

The $J \times J$ matrix $\mathbf{X}^T\mathbf{X}$ is non-singular so there is a unique solution

$$\mathbf{b} = \frac{1}{K}\begin{bmatrix} Y_{1.} \\ Y_{2.} - Y_{1.} \\ \vdots \\ Y_{J.} - Y_{1.} \end{bmatrix}$$

for the normal (7.1). Hence

$$\mathbf{b}^T\mathbf{X}^T\mathbf{y} = \frac{1}{K}[Y_{..}\, Y_{1.} + \sum_{j=2}^{J} Y_{j.}(Y_{j.} - Y_{1.})] = \frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2$$

and the fitted values $\hat{\mathbf{y}} = [\bar{y}_1, \bar{y}_1, \ldots, \bar{y}_J]^T$ as before.

Table 7.3 ANOVA table for one factor with J levels and equal sample size K per level.

Source of variation	Degrees of freedom	Sum of squares	Mean square	f
Mean	1	$\frac{1}{N}Y_{\cdot\cdot}^2$		
Between levels of factor A	$J-1$	$\frac{1}{K}\sum_{j=1}^{J} Y_{j\cdot}^2 - \frac{1}{N}Y_{\cdot\cdot}^2$ $= \sigma^2(D_0-D_1)$	$\frac{\sigma^2(D_0-D_1)}{J-1}$	$\frac{D_0-D_1}{J-1}\Big/\frac{D_1}{N-J}$
Residual	$N-J$	$\sum_{j=1}^{J}\sum_{k=1}^{J} Y_{jk}^2 - \frac{1}{K}\sum_{j=1}^{J} Y_{j\cdot}^2$ $= \sigma^2 D_1$	$\sigma^2 D_1/N-J$	
Total	N	$\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}^2$		

Thus although the three specifications of the model differ, the values of $\hat{\mathbf{y}} = \mathbf{Xb}$, $\mathbf{b}^T\mathbf{X}^T\mathbf{y}$ and

$$D_1 = \frac{1}{\sigma^2}(\mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}) = \frac{1}{\sigma^2}\left[\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}^2 - \frac{1}{K}\sum_{j=1}^{J} Y_{j\cdot}^2\right]$$

are the same in each case.

These three versions of the model $E(Y_{jk}) = \mu_j$ all correspond to the hypothesis H_1: the response means for each level may differ. To compare this with the null hypothesis H_0: the means are all equal, we consider the model $E(Y_{jk}) = \mu$ so that $\boldsymbol{\beta} = [\mu]$ and $\mathbf{X} = [\mathbf{1}\,\mathbf{1}\ldots\mathbf{1}]^T$. Then $\mathbf{X}^T\mathbf{X} = N$, $\mathbf{X}^T\mathbf{y} = Y_{\cdot\cdot}$ and hence $\mathbf{b} = \hat{\mu} = Y_{\cdot\cdot}/N$ with $\mathbf{b}^T\mathbf{X}^T\mathbf{y} = Y_{\cdot\cdot}^2/N$ and

$$D_0 = \frac{1}{\sigma^2}\left[\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}^2 - \frac{Y_{\cdot\cdot}^2}{N}\right].$$

To test H_0 against H_1 we assume that H_1 is correct so that $D_1 \sim \chi^2_{N-J}$. If, in addition, H_0 is correct then $D_0 \sim \chi^2_{N-1}$, otherwise D_0 has a non-central chi-squared distribution. Thus if H_0 is correct

$$D_0 - D_1 = \frac{1}{\sigma^2}\left[\frac{1}{K}\sum_{j=1}^{J} Y_{j\cdot}^2 - \frac{1}{N}Y_{\cdot\cdot}^2\right] \sim \chi^2_{J-1}$$

and so

$$f = \frac{D_0-D_1}{J-1}\Big/\frac{D_1}{N-J} \sim F_{J-1,N-J};$$

if H_0 is not correct f is likely to be larger than predicted from the $F_{J-1,N-J}$ distribution. Conventionally this hypothesis test is set out as in Table 7.3.

For the plant weight data the results are summarized in Table 7.4.

Table 7.4 ANOVA table for plant weight data in Table 7.1.

Source of variation	Degrees of freedom	Sum of squares	Mean square	f
Mean	1	772.060		
Between treatments	2	3.766	1.833	4.85
Residual	27	10.492	0.389	
Total	30	786.318		

Since $f = 4.85$ is significant at the 5% level when compared with the $F_{2,27}$ distribution, we conclude that the mean responses differ. From the model $E(Y_{jk}) = \mu_j$ the means are

$$\mathbf{b} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{bmatrix} = \begin{bmatrix} 5.032 \\ 4.661 \\ 5.526 \end{bmatrix}.$$

We assume that H_1 is correct so that $D_1 \sim \chi^2_{27}$. Using the observed value of D_1, $10.492/\sigma^2$, and the fact that $E(\chi^2_n) = n$ we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{10.492}{E(D_1)} = \frac{10.492}{27} = 0.389$$

(i.e. the residual mean square in Table 7.4). Thus the standard error of each mean is $(0.389/10)^{\frac{1}{2}} = 0.197$ because the variance–covariance matrix of $\mathbf{b}$ is $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ where

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix}.$$

Now it can be seen that the significant effect is due to the mean for treatment B being significantly larger than the others.

7.4 Two factor ANOVA with replication

Consider the fictitious data in Table 7.5 in which factor A (with $J = 3$ levels) and factor B (with $K = 2$ levels) are *crossed* so that there are JK subclasses formed by all combinations of A and B levels. In each subclass there are $L = 2$ observations or *replications*.

Table 7.5 Fictitious data for two factor ANOVA with equal numbers of observations in each subclass.

Levels of factor A	Levels of factor B		Total
	B_1	B_2	
A_1	6.8, 6.6	5.3, 6.1	24.8
A_2	7.5, 7.4	7.2, 6.5	28.6
A_3	7.8, 9.1	8.8, 9.1	34.8
Total	45.2	43.0	88.2

The main hypotheses are:

H_I: there are no interaction effects, i.e. the effects of A and B are additive;
H_A: there are no differences in response associated with different levels of factor A;
H_B: there are no differences in response associated with different levels of factor B.

Thus we need to consider a *full model* and three *reduced models* formed by omitting various terms from the full model.

(i) The full model is

$$E(Y_{jkl}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} \tag{7.5}$$

where the terms $(\alpha\beta)_{jk}$ correspond to *interaction effects* and α_j and β_k to *main effects* of the factors.

(ii) The *additive model* is

$$E(Y_{jkl}) = \mu + \alpha_j + \beta_k. \tag{7.6}$$

This is compared to the full model to test hypothesis H_I.

(iii) The model formed by omitting effects due to B is

$$E(Y_{jkl}) = \mu + \alpha_j. \tag{7.7}$$

This is compared to the additive model to test hypothesis H_B.

(iv) The model formed by omitting effects due to A is

$$E(Y_{jkl}) = \mu + \beta_j. \tag{7.8}$$

This is compared to the additive model to test hypothesis H_A.

The models (7.5)–(7.8) have too many parameters; for instance replicates in the same subclass have the same expected value so there can be at most JK independent expected values but the full model has

$1+J+K+JK=(J+1)(K+1)$ parameters. To overcome this difficulty (which leads to the singularity of $\mathbf{X}^T\mathbf{X}$) we can impose the extra constraints

$$\alpha_1+\alpha_2+\alpha_3=0, \quad \beta_1+\beta_2=0,$$
$$(\alpha\beta)_{11}+(\alpha\beta)_{12}=0, \quad (\alpha\beta)_{21}+(\alpha\beta)_{22}=0, \quad (\alpha\beta)_{31}+(\alpha\beta)_{32}=0,$$
$$(\alpha\beta)_{11}+(\alpha\beta)_{21}+(\alpha\beta)_{31}=0$$

(the remaining condition $(\alpha\beta)_{12}+(\alpha\beta)_{22}+(\alpha\beta)_{32}=0$ follows from the last four equations). These are the conventional constraint equations for ANOVA. Alternatively we can take $\alpha_1=\beta_1=(\alpha\beta)_{11}=(\alpha\beta)_{12}=(\alpha\beta)_{21}=(\alpha\beta)_{31}=0$ as the corner-point constraints. In either case the numbers of (linearly) independent parameters are: 1 for μ, $J-1$ for the α_js, $K-1$ for the β_ks, and $(J-1)(K-1)$ for the $(\alpha\beta)_{jk}$s.

Details of fitting all four models using either the sum-to-zero constraints or the corner-point constraints are given in Appendix 3.

In general, hypothesis tests may not be statistically independent so that the order in which the models are fitted affects the results. For the data in Table 7.5, however, it is possible to specify the design matrix $\mathbf{X}$ so that it has orthogonal components corresponding to the mean, H_I, H_A and H_B and therefore the hypothesis tests are independent. Details of this orthogonal parameterization are also given in Appendix 3.

For models (7.5)–(7.8) the estimates $\mathbf{b}$ depend on the choice of constraints and dummy variables. However, the fitted values $\hat{\mathbf{y}}=\mathbf{Xb}$ are the same for all specifications of the models and so the values of $\mathbf{b}^T\mathbf{X}^T\mathbf{y}$ and $\sigma^2D=\mathbf{y}^T\mathbf{y}-\mathbf{b}^T\mathbf{X}^T\mathbf{y}$ are the same. For these data $\mathbf{y}^T\mathbf{y}=664.1$ and the other results are summarized in Table 7.6 (the subscripts F, I, A and B refer to the full model and the models corresponding to H_I, H_A and H_B, respectively).

Table 7.6 Summary of calculations for data in Table 7.5.

Terms in model	Hypothesis	Number of parameters	$\mathbf{b}^T\mathbf{X}^T\mathbf{y}$	$\sigma^2D=\mathbf{y}^T\mathbf{y}-\mathbf{b}^T\mathbf{X}^T\mathbf{y}$
$\mu+\alpha_j+\beta_k+(\alpha\beta)_{jk}$		6	662.6200	$\sigma^2D_F=$ 1.4800
$\mu+\alpha_j+\beta_k$	H_I	4	661.4133	$\sigma^2D_I=$ 2.6867
$\mu+\alpha_j$	H_B	3	661.0100	$\sigma^2D_B=$ 3.0900
$\mu+\beta_k$	H_A	2	648.6733	$\sigma^2D_A=$ 15.4267
μ		1	648.2700	

To test H_I we assume that the full model is correct so that $D_F\sim\chi^2_6$ because there are $N=12$ observations and the model has $JK=6$ independent parameters. If H_I is correct also then $D_I\sim\chi^2_8$ so that $D_I-D_F\sim\chi^2_2$ and

$$f=\frac{D_I-D_F}{2}\Big/\frac{D_F}{6}\sim F_{2,6}.$$

The value of

$$f = \frac{2.6867 - 1.48}{2\sigma^2} \Big/ \frac{1.48}{6\sigma^2} = 2.45$$

is not significant so the data provide no evidence against H_I. Since H_I is not rejected we proceed to test H_A and H_B. For H_B we consider the difference in fit between the models $E(Y_{jkl}) = \mu + \alpha_j$ and $E(Y_{jkl}) = \mu + \alpha_j + \beta_k$ i.e. $D_B - D_I$ and compare this with D_F using

$$f = \frac{D_B - D_I}{1} \Big/ \frac{D_F}{6} = \frac{3.09 - 2.6867}{\sigma^2} \Big/ \frac{1.48}{6\sigma^2} = 1.63$$

which is not significant compared to the $F_{1,6}$ distribution, suggesting that there are no differences due to levels of factor B. The corresponding test for H_A gives $f = 25.82$ which is significant compared with the $F_{2,6}$ distribution. Thus we conclude that the response means are only affected by differences in the levels of factor A. These results are usually summarized as shown in Table 7.7.

Table 7.7 ANOVA table for data in Table 7.5.

Source of variation	Degrees of freedom	Sum of squares	Mean square	f
Mean	1	648.2700		
Levels of A	2	12.7400	6.3700	25.82
Levels of B	1	0.4033	0.4033	1.63
Interactions	2	1.2067	0.6033	2.45
Residual	6	1.4800	0.2467	
Total	12	664.1		

7.5 Crossed and nested factors and more complicated models

In the example in Section 7.4 the factors A and B are said to be *crossed* because there is a subclass corresponding to each combination of levels A_j and B_k and all the comparisons represented by the terms α_j, β_k and $(\alpha\beta)_{jk}$ in the full model $E(Y_{jkl}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ are of potential interest.

This contrasts with the two factor *nested* design shown in Table 7.8 which represents an experiment to compare two drugs (A_1 and A_2) one of which is tested in three hospitals (B_1, B_2 and B_3) and the other in two hospitals (B_4 and B_5).

We want to compare the effects of the two drugs and possible differences in response between hospitals using the same drug. It is not sensible to make

Table 7.8 Nested two factor experiment.

	Drug A_1			Drug A_2	
Hospitals	B_1	B_2	B_3	B_4	B_5
Responses	Y_{111}	Y_{121}	Y_{131}	Y_{241}	Y_{251}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Y_{11n_1}	Y_{12n_2}	Y_{13n_3}	Y_{24n_4}	Y_{25n_5}

comparisons between hospitals using different drugs. A suitable full model is $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ where

$$\boldsymbol{\beta} = [\mu, \alpha_1, \alpha_2, (\alpha\beta)_{11}, (\alpha\beta)_{12}, (\alpha\beta)_{13}, (\alpha\beta)_{24}, (\alpha\beta)_{25}]^{\mathrm{T}}$$

and the response vector $\mathbf{y}$ has length $N = \sum_{k=1}^{5} n_k$. For the conventional ANOVA constraints we let $\alpha_1 + \alpha_2 = 0$, $(\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{13} = 0$ and $(\alpha\beta)_{24} + (\alpha\beta)_{25} = 0$, or for the corner-point constraints we take $\alpha_1 = (\alpha\beta)_{11} = (\alpha\beta)_{24} = 0$. Reduced models to compare hospitals using the same drug are formed by omitting the terms $(\alpha\beta)_{1k}$, $k = 1, 2, 3$ and, separately, $(\alpha\beta)_{2k}$, $k = 4, 5$. The reduced model for the hypothesis of no difference between the drugs (but allowing for differences between hospitals) is $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta} = [\mu, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]^{\mathrm{T}}$ where the β_ks correspond to hospitals and $\beta_1 + \beta_2 + \beta_3 = 0$ and $\beta_4 + \beta_5 = 0$ or $\beta_1 = \beta_4 = 0$.

ANOVA models can readily be defined for more than two factors. The factors may be crossed or nested or some mixture of these forms. The models can include higher order interaction terms such as $(\alpha\beta\gamma)_{jkl}$ as well as the first order interactions like $(\alpha\beta)_{jk}$ and the main effects. These extensions do not involve any fundamental differences from the examples already considered so they are not examined further in this book.

7.6 More complicated hypotheses

In all the above examples we only considered hypotheses in which certain parameters in the full model are omitted in the reduced models. For instance, in the plant weight example $E(Y_{jk}) = \mu + \alpha_j$ in the full model and $E(Y_{jk}) = \mu$ in the reduced model corresponding to the hypothesis that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Sometimes we are interested in testing more complicated hypotheses such as treatments A and B in the plant weight experiment being equally effective but different from the control, i.e. $\alpha_2 = \alpha_3$ but α_1 not necessarily the same. Such hypotheses can be readily accommodated in the model fitting approach by the appropriate choice of parameters and dummy variables, for example the hypothesis $\alpha_2 = \alpha_3$ is equivalent to fitting $E(Y_{1k}) = \beta_1$ and $E(Y_{2k}) = E(Y_{3k}) = \beta_2$.

7.7 Independence of hypothesis tests

In the two factor ANOVA example in Section 7.4 the tests of the three hypotheses H_I, H_A, H_B are statistically independent because there is an orthogonal form of the design matrix $\mathbf{X}$ for the full model so that $\mathbf{X}^T\mathbf{X}$ is block diagonal with blocks corresponding to the mean and the three hypotheses. Hence the total sum of squares can be partitioned into disjoint components corresponding to the mean, H_I, H_A, H_B and the residual. For two factor ANOVA such a partition is only possible if the numbers n_{jk} of observations in each subclass satisfy $n_{jk} = n_{j.}n_{.k}/n_{..}$ (see Winer, 1971, Section 5.23–8).

In general, multiple hypothesis tests are only independent if there is a design matrix with orthogonal components so that the total sum of squares can be partitioned into disjoint terms corresponding to the hypotheses. Usually this is only possible if the hypotheses are particularly simple (e.g. interaction and main effects are zero) and if the experimental design is *balanced* (i.e. there are equal numbers of observations in each subclass). If the hypotheses are not independent then care is needed in interpreting simultaneous significance tests.

7.8 Choice of constraint equations and dummy variables

The numerical examples also illustrate several major issues relating to the choice of constraint equations and dummy variables for ANOVA models.

ANOVA models are usually specified in terms of parameters which are readily interpretable as marginal effects due to factor levels and interactions. However, the models contain more parameters than there are independent normal equations. Therefore extra equations, traditionally in the form of sum-to-zero constraints are added. (If the design is unbalanced there is some controversy about the most appropriate choice of constraint equations.) In the framework of generalized linear models this means that the equations (7.2) to be solved are not the normal equations obtained by the methods of maximum likelihood or least squares. Therefore the standard computational procedures cannot be used. Also the terms of $\boldsymbol{\beta}$ are generally not identifiable, and unique unbiased point estimates and confidence intervals can only be obtained for certain linear combinations of parameters, called *estimable functions*. Nevertheless, if the main purpose of analysing the data is to test hypotheses, the use of sum-to-zero constraints is entirely appropriate and convenient provided that special purpose computer programs are used.

If the corner-point constraints are used the elements of $\boldsymbol{\beta}$ and the corresponding columns of $\mathbf{X}$ are arranged as $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]^T$ and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ so that $\mathbf{X}_1^T\mathbf{X}_1$ is non-singular and $\boldsymbol{\beta}_2$ is set to $\mathbf{0}$. Thus

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1.$$

Then the normal equations

$$\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\mathbf{b}_1 = \mathbf{X}_1^{\mathrm{T}}\mathbf{y}$$

can be solved using standard multiple regression or generalized linear modelling programs and the estimators have various desirable properties (e.g. $\mathbf{b}_1$ is unbiased and has variance–covariance matrix $\sigma^2(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}$). However, the interpretation of parameters subject to corner-point constraints is perhaps less straightforward than with sum-to-zero constraints. Also all the calculations usually have to be repeated for each new model fitted. In practice, estimation using corner-point constraints is performed so that parameters are estimated sequentially in such a way that the redundant corner-point parameters (which are said to be *aliased*) are systematically identified and set equal to zero (for example, this is the procedure used in GLIM).

In the two factor ANOVA example in Section 7.4, the most elegant analysis was obtained by choosing the dummy variables so that the design matrix $\mathbf{X}$ had orthogonal components corresponding to each of the hypotheses to be tested. For simple well-planned experiments where this form of analysis is possible there are computational benefits (e.g. parameter estimates are the same for all models) and advantages in interpretation (e.g. independence of the hypothesis tests). However, for unbalanced experimental designs or hypotheses involving more complicated contrasts, it is unlikely that orthogonal forms exist.

In summary, for any particular sequence of models the choice of constraints and dummy variables affects the computational procedures and the parameter estimates. However, it does not influence the results for hypothesis testing. The reason is that any solution $\mathbf{b}$ of the normal equations (7.1) corresponds to the unique minimum of $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$. Hence the statistics $\sigma^2 D = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ are the same regardless of the way the models are specified.

7.9 Analysis of covariance

This is the term used for mixed models in which some of the explanatory variables are dummy variables representing factor levels and others are continuous measurements, called covariates. As with ANOVA we are interested in comparing means for subclasses defined by factor levels but, recognizing that the covariates may also affect the responses, we compare the means after 'adjustment' for covariate effects.

A typical example is provided by the data in Table 7.9. The responses Y_{jk} are achievement scores, the levels of the factor represent three different training methods and the covariates x_{jk} are aptitude scores measured before training commenced. We want to compare the training methods, taking into account differences in initial aptitude between the three groups of subjects.

Table 7.9 Achievement scores (data from Winer, 1971, p. 766).

Training method	A_1		A_2		A_3	
	y	x	y	x	y	x
	6	3	8	4	6	3
	4	1	9	5	7	2
	5	3	7	5	7	2
	3	1	9	4	7	3
	4	2	8	3	8	4
	3	1	5	1	5	1
	6	4	7	2	7	4
Total	31	15	53	24	47	19
Sum of squares	147	41	413	96	321	59
Σxy	75		191		132	

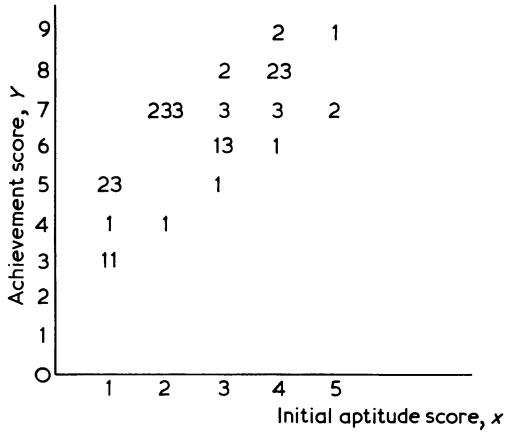

Figure 7.1 Plot of data in Table 7.9, 1, 2 and 3 indicate the corresponding training methods.

The data are shown in Fig. 7.1. There is evidence that the achievement scores Y increase linearly with aptitude x and that the Y values are generally higher for treatment groups A_2 and A_3 than for A_1.

We compare the models

$$E(Y_{jk}) = \mu_j + \gamma x_{jk} \tag{7.9}$$

and

$$E(Y_{jk}) = \mu + \gamma x_{jk} \tag{7.10}$$

for $j = 1, 2, 3$ and $k = 1, \ldots, 7$. Model (7.10) corresponds to the null hypothesis that there are no differences in mean achievement scores between the three

training methods. Let $\mathbf{y}_j = [Y_{j1}, \ldots, Y_{j7}]^T$ and $\mathbf{x}_j = [x_{j1}, \ldots, x_{j7}]^T$ so that in matrix notation model (7.9) is $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \gamma \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{x}_1 \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{x}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{x}_3 \end{bmatrix}$$

where $\mathbf{0}$ and $\mathbf{1}$ are vectors of length 7. Then

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 7 & 0 & 0 & 15 \\ 0 & 7 & 0 & 24 \\ 0 & 0 & 7 & 19 \\ 15 & 24 & 19 & 196 \end{bmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 31 \\ 53 \\ 47 \\ 398 \end{bmatrix} \quad \text{and so} \quad \mathbf{b} = \begin{bmatrix} 2.837 \\ 5.024 \\ 4.698 \\ 0.743 \end{bmatrix}.$$

Also $\mathbf{y}^T\mathbf{y} = 881$ and $\mathbf{b}^T\mathbf{X}^T\mathbf{y} = 870.698$ so for model (7.9)

$$\sigma^2 D_1 = \mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y} = 10.302.$$

For the reduced model (7.10)

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \gamma \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 \\ \mathbf{1} & \mathbf{x}_2 \\ \mathbf{1} & \mathbf{x}_3 \end{bmatrix} \quad \text{so} \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 21 & 58 \\ 58 & 196 \end{bmatrix}$$

and

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 131 \\ 398 \end{bmatrix}.$$

Hence

$$\mathbf{b} = \begin{bmatrix} 3.447 \\ 1.011 \end{bmatrix}, \quad \mathbf{b}^T\mathbf{X}^T\mathbf{y} = 853.766$$

so $\sigma^2 D_0 = 27.234$.

If we assume that model (7.9) is correct, then $D_1 \sim \chi^2_{17}$. If the null hypothesis corresponding to model (7.10) is true then $D_0 \sim \chi^2_{19}$ so

$$f = \frac{D_0 - D_1}{2} \Big/ \frac{D_1}{17} \sim F_{2,17}.$$

For these data

$$f = \frac{16.932}{2} \Big/ \frac{10.302}{17} = 13.97$$

indicating a significant difference in achievement scores for the training methods, after adjustment for initial differences in aptitude. The usual presentation of this analysis is given in Table 7.10.

Table 7.10 ANCOVA table for data in Table 7.9.

Source of variation	Degrees of freedom	Sum of squares	Mean square	f
Mean and covariate	2	853.766		
Factor levels	2	16.932	8.466	13.97
Residual	17	10.302	0.606	
Total	21	881.000		

7.10 Exercises

7.1 Total solids (%) were determined in each of the six batches of cream ($B_1, \ldots, B_6$) by each of three analysts (A_1, A_2 and A_3) with the results shown in the Table 7.11.

(i) Test the hypothesis H_A that there are no differences due to analysts.

(ii) Estimate the solid content of each batch. Test the hypothesis H_B that there are no differences between batches.

(iii) Examine the residuals for the most appropriate model and comment on the results of your analysis.

Table 7.11 Total solids measured in batches of cream.

	Batches					
Analysts	B_1	B_2	B_3	B_4	B_5	B_6
A_1	35.3	32.3	38.7	30.1	32.4	35.1
A_2	35.7	34.5	36.1	29.8	32.1	34.2
A_3	34.8	31.9	40.2	31.2	33.0	34.6

7.2 Perform a complete analysis of variance for the two factor experiment shown in Table 7.12. Verify that the null hypotheses of no differences due to interactions or main effects are not all independent.

Table 7.12 Example of two factor experiment.

	Factor B	
Factor A	B_1	B_2
A_1	5	3, 4
A_2	6, 4	4, 3
A_3	7	6, 8

7.3 For the achievement score data in Table 7.9:

(i) test the hypothesis that the treatment effects are equal, ignoring the covariate, i.e. compare $E(Y_{jk}) = \mu_j$ with $E(Y_{jk}) = \mu$;

(ii) test the assumption that initial aptitude has the same effect for all training methods, i.e. compare $E(Y_{jk}) = \mu_j + \gamma_j x_{jk}$ with $E(Y_{jk}) = \mu_j + \gamma x_{jk}$.

7.4 Show that $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \geqslant (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b})$ where $\mathbf{b}$ is any solution of the normal equations $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$. Hence the minimum of $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is attained when $\boldsymbol{\beta} = \mathbf{b}$ and is the same for all solutions of the normal equations.