Introduction to Mixed Modelling

Beyond Regression and Analysis of Variance

N. W. Galwey
Genetic Analysis, GlaxoSmithKline, UK

John Wiley & Sons, Ltd

Introduction to Mixed Modelling

Introduction to Mixed Modelling

Beyond Regression and Analysis of Variance

N. W. Galwey
Genetic Analysis, GlaxoSmithKline, UK

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, ONT, L5R 4J3, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data:

Galwey, Nick.
Introduction to mixed modelling : beyond regression and analysis of variance / Nicholas W. Galwey.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-470-01496-7 (acid-free paper)
ISBN-10: 0-470-01496-2 (acid-free paper)
1. Multilevel models (Statistics) 2. Experimental design. 3. Regression analysis. 4. Analysis of variance. I. Title.
QA276.G33 2006
591.5–dc22

2006023991

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470-01496-7 (HB)
ISBN-10: 0-470-01496-2 (HB)

Typeset in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface **ix**

1 The need for more than one random-effect term when fitting a regression line **1**

1.1 A data set with several observations of variable Y at each value of variable X 1
1.2 Simple regression analysis. Use of the software GenStat to perform the analysis 2
1.3 Regression analysis on the group means 10
1.4 A regression model with a term for the groups 12
1.5 Construction of the appropriate F test for the significance of the explanatory variable when groups are present 15
1.6 The decision to regard a model term as random: a mixed model 16
1.7 Comparison of the tests in a mixed model with a test of lack of fit 17
1.8 The use of residual maximum likelihood to fit the mixed model 18
1.9 Equivalence of the different analyses when the number of observations per group is constant 21
1.10 Testing the assumptions of the analyses: inspection of the residual values 27
1.11 Use of the software R to perform the analyses 29
1.12 Fitting a mixed model using GenStat's GUI 33
1.13 Summary 39
1.14 Exercises 40

2 The need for more than one random-effect term in a designed experiment **45**

2.1 The split plot design: a design with more than one random-effect term 45
2.2 The analysis of variance of the split plot design: a random-effect term for the main plots 47
2.3 Consequences of failure to recognise the main plots when analysing the split plot design 55
2.4 The use of mixed modelling to analyse the split plot design 57

	2.5	A more conservative alternative to the Wald statistic	60
	2.6	Justification for regarding block effects as random	61
	2.7	Testing the assumptions of the analyses: inspection of the residual values	62
	2.8	Use of R to perform the analyses	63
	2.9	Summary	67
	2.10	Exercises	68
3	**Estimation of the variances of random-effect terms**		**73**
	3.1	The need to estimate variance components	73
	3.2	A hierarchical random-effect model for a three-stage assay process	73
	3.3	The relationship between variance components and stratum mean squares	78
	3.4	Estimation of the variance components in the hierarchical random-effect model	80
	3.5	Design of an optimum strategy for future sampling	82
	3.6	Use of R to analyse the hierarchical three-stage assay process	85
	3.7	Genetic variation: a crop field trial with an unbalanced design	87
	3.8	Production of a balanced experimental design by 'padding' with missing values	92
	3.9	Regarding a treatment term as a random-effect term. The use of mixed-model analysis to analyse an unbalanced data set	96
	3.10	Comparison of a variance-component estimate with its standard error	99
	3.11	An alternative significance test for variance components	101
	3.12	Comparison among significance tests for variance components	103
	3.13	Inspection of the residual values	104
	3.14	Heritability. The prediction of genetic advance under selection	105
	3.15	Use of R to analyse the unbalanced field trial	109
	3.16	Estimation of variance components in the regression analysis on grouped data	113
	3.17	Estimation of variance components for block effects in the split plot experimental design	115
	3.18	Summary	117
	3.19	Exercises	118
4	**Interval estimates for fixed-effect terms in mixed models**		**125**
	4.1	The concept of an interval estimate	125
	4.2	SEs for regression coefficients in a mixed-model analysis	126
	4.3	SEs for differences between treatment means in the split plot design	130
	4.4	A significance test for the difference between treatment means	133
	4.5	The least significant difference between treatment means	137
	4.6	SEs for treatment means in designed experiments: a difference in approach between analysis of variance and mixed-model analysis	141

4.7 Use of R to obtain SEs of means in a designed experiment 147
4.8 Summary 148
4.9 Exercises 150

5 Estimation of random effects in mixed models: best linear unbiased predictors 151
5.1 The difference between the estimates of fixed and random effects 151
5.2 The method for estimation of random effects. The best linear unbiased predictor or 'shrunk estimate' 154
5.3 The relationship between the shrinkage of BLUPs and regression towards the mean 156
5.4 Use of R for the estimation of random effects 162
5.5 Summary 164
5.6 Exercises 165

6 More advanced mixed models for more elaborate data sets 169
6.1 Features of the models introduced so far: a review 169
6.2 Further combinations of model features 170
6.3 The choice of model terms to be regarded as random 172
6.4 Disagreement concerning the appropriate significance test when fixed- and random-effect terms interact 174
6.5 Arguments for regarding block effects as random 181
6.6 Examples of the choice of fixed- and random-effect terms 186
6.7 Summary 190
6.8 Exercises 193

7 Two case studies 193
7.1 Further development of mixed-modelling concepts through the analysis of specific data sets 193
7.2 A fixed-effect model with several variates and factors 194
7.3 Use of R to fit the fixed-effect model with several variates and factors 209
7.4 A random-effect model with several factors 214
7.5 Use of R to fit the random-effect model with several factors 229
7.6 Summary 238
7.7 Exercises 238

8 The use of mixed models for the analysis of unbalanced experimental designs 251
8.1 A balanced incomplete block design 251
8.2 Imbalance due to a missing block. Mixed-model analysis of the incomplete block design 255
8.3 Use of R to analyse the incomplete block design 259
8.4 Relaxation of the requirement for balance: alpha designs 261

	8.5	Use of R to analyse the alphalpha design	269
	8.6	Summary	271
	8.7	Exercises	272
9	**Beyond mixed modelling**		**275**
	9.1	Review of the uses of mixed models	275
	9.2	The generalised linear mixed model. Fitting a logistic (sigmoidal) curve to proportions of observations	276
	9.3	Fitting a GLMM to a contingency table. Trouble-shooting when the mixed-modelling process fails	284
	9.4	The hierarchical generalised linear model	298
	9.5	The role of the covariance matrix in the specification of a mixed model	303
	9.6	A more general pattern in the covariance matrix. Analysis of pedigree data	307
	9.7	Estimation of parameters in the covariance matrix. Analysis of temporal and spatial variation	317
	9.8	Summary	327
	9.9	Exercises	327
10	**Why is the criterion for fitting mixed models called residual maximum likelihood?**		**333**
	10.1	Maximum likelihood and residual maximum likelihood	333
	10.2	Estimation of the variance σ^2 from a single observation using the maximum likelihood criterion	334
	10.3	Estimation of σ^2 from more than one observation	334
	10.4	The μ-effects axis as a dimension within the sample space	338
	10.5	Simultaneous estimation of μ and σ^2 using the maximum likelihood criterion	339
	10.6	An alternative estimate of σ^2 using the REML criterion	342
	10.7	Extension to the general linear model. The fixed-effect axes as a subspace of the sample space	345
	10.8	Application of the REML criterion to the general linear model	349
	10.9	Extension to models with more than one random-effect term	351
	10.10	Summary	352
	10.11	Exercises	353
References			**357**
Index			**361**

Preface

This book is intended for research workers and students who have made some use of the statistical techniques of regression analysis and analysis of variance (anova), but who are unfamiliar with *mixed models* and the criterion for fitting them called *residual maximum likelihood* (REML, also known as *restricted maximum likelihood*). Such readers will know that, broadly speaking, regression analysis seeks to account for the variation in a response variable by relating it to one or more explanatory variables, whereas anova seeks to detect variation among the mean values of groups of observations. In regression analysis, the statistical significance of each explanatory variable is tested using the same estimate of residual variance, namely the residual mean square, and this estimate is also used to calculate the standard error of the effect of each explanatory variable. However, this choice is not always appropriate. Sometimes, one or more of the terms in the regression model (in addition to the residual term) represents *random variation*, and such a term will contribute to the observed variation in other terms. It should therefore contribute to the significance tests and standard errors of these terms, but in an ordinary regression analysis, it does not do so. Anova, on the other hand, does allow the construction of models with additional random-effect terms, known as block terms. However, it does so only in the limited context of balanced experimental designs.

The capabilities of regression analysis can be combined with those of anova by fitting to the data a mixed model, so called because it contains both fixed-effect and random-effect terms. A mixed model allows the presence of additional random-effect terms to be recognised in the full range of regression models, not just in balanced designs. Any statistical analysis that can be specified by a general linear model (the broadest form of linear regression model) or by anova can also be specified by a mixed model. However, the specification of a mixed model requires an additional step. The researcher must decide, for each term in the model, whether effects of that term (e.g. the deviations of group means from the grand mean) can be regarded as a random sample from some much larger population, or whether they are a fixed set. In some cases this decision is straightforward; in others, the distinction is subtle and the decision difficult. However, provided that an appropriate decision is made (see Chapter 6, Section 6.3), the mixed model specifies a statistical analysis which is of broader validity than regression analysis or anova, and which is nearly equivalent to those methods (though slightly less powerful) in the special cases where they are applicable.

It is fairly straightforward to specify the calculations required for regression analysis and anova, and this is done in many standard textbooks. For example, Draper and

Smith (1998) give a clear, thorough and extensive account of the methods of regression analysis, and Mead (1988) does the same for anova. To solve the equations that specify a mixed model is much less straightforward. The model is fitted – that is, the best estimates of its parameters are obtained – using the REML criterion, but the fitting process requires recursive numerical methods. It is largely because of this burden of calculation that mixed models are less familiar than regression analysis and anova: it is only in about the last two decades that the development of computer power and user-friendly statistical software has allowed them to be routinely used in research. This book aims to provide a guide to the use of mixed models that is accessible to the broad community of research scientists. It focuses not on the details of calculation, but on the specification of mixed models and the interpretation of the results.

The numerical examples in this book are presented and analysed using two statistical software systems, namely:

- GenStat, distributed by VSN International Ltd, Hemel Hempstead, via the web site http://www.vsni.co.uk/products/genstat/. Anyone who has bought this book can obtain free use of GenStat for a period of 12 months. Details, together with Genstat programs and data files for many of the examples in this book, can be found at www.wiley.com/go/mixed-modelling (as can the solutions to the end of chapter exercises).
- R, from The R Project for Statistical Computing. This software can be downloaded free of charge from the web site http://www.r-project.org/.

GenStat is a natural choice of software to illustrate the concepts and methods employed in mixed modelling because its facilities for this purpose are straightforward to use, extensive and well integrated with the rest of the system, and because their output is clearly laid out and easy to interpret. Above all, the recognition of random terms in statistical models lies at the heart of GenStat. GenStat's method of specifying anova models requires the distinction of random-effect (block) and fixed-effect (treatment) terms, which makes the interpretation of designed experiments uniquely reliable and straightforward. This approach extends naturally to REML models, and provides a firm framework within which the researcher can think and plan. Despite these merits, GenStat is not among the most widely used statistical software systems, and most of the numerical examples are therefore also analysed using the increasingly popular software R. Development of this software is taking place rapidly, and it is likely that its already-substantial power for mixed modelling will increase steadily in the future.

I am grateful to Mr Peter Lane, Dr Aruna Bansal and Dr Caroline Galwey for valuable comments and suggestions on parts of the manuscript of this book, and to the participants in the GenStat Discussion List for helpful responses to many enquiries. (Access to this lively forum can be obtained via the GenStat User Area at web site http://www.vsni.co.uk/products/genstat/user/) I am particularly grateful to Dr Roger Payne for an expert and sharp-eyed reading of the entire manuscript. Any errors or omissions of fact or interpretation that remain are the sole responsibility of the author. I would also like to express my gratitude to the many individuals and organisations who have given permission for the reproduction of data in the numerical examples presented. They are acknowledged individually in their respective places, but the high level of support that they have given me should be recognised here.

1

The need for more than one random-effect term when fitting a regression line

1.1 A data set with several observations of variable *Y* at each value of variable *X*

One of the commonest, and simplest, uses of statistical analysis is the fitting of a straight line, known for historical reasons as a *regression line*, to describe the relationship between an *explanatory variable*, X, and a *response variable*, Y. The departure of the values of Y from this line is the *residual variation*, which is regarded as random, and it is natural to ask whether the part of the variation in Y that is explained by the relationship with X is significantly greater than this residual variation. This is a simple *regression analysis*, and for many data sets it is all that is required. However, in some cases, several observations of Y are taken at each value of X. The data then form natural groups, and it may no longer be appropriate to analyse them as though every observation were independent: observations of Y at the same value of X may lie at a similar distance from the line. We may then be able to recognise two sources of random variation, namely:

- variation among groups
- variation among observations within each group.

This is one of the simplest situations in which it is necessary to consider the possibility that there may be more than a single *stratum* of random variation – or, in the language of mixed modelling, that a model with more than one *random-effect term* may be required. In this chapter we will examine a data set of this type, and explore how the usual regression analysis is modified by the fact that the data form natural groups.

Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance N. W. Galwey

We will explore this question in a data set that relates the prices of houses in England to their latitude. There is no doubt that houses cost more in the south of England than in the north: these data will not lead to any new conclusions, but they will illustrate this trend, and the methods used to explore it. The data are displayed in a spreadsheet in Table 1.1. The first cell in each column contains the name of the variable held in that column. The variables 'latitude' and 'price_pounds' are *variates* – lists of observations that can take any numerical value, the commonest kind of data for statistical analysis. However, the observations of the variable 'town' can only take certain permitted values – in this case, the names of the 11 towns under consideration. A variable of this type is called a *factor*, and the exclamation mark (!) after its name indicates that 'town' is of this type. The towns are the groups of observations: within each town, all the houses are at nearly the same latitude, and the latitude of the town is represented by a single nominal value in this data set. In contrast, the price of each house is potentially unique.

Before commencing a formal analysis of this data set, we should note its limitations. A thorough investigation of the relationship between latitude and house price would take into account many factors besides those recorded here – the number of bedrooms in each house, its state of repair and improvement, other observable indicators of the desirability of its location, and so on. To some extent such sources of variation have been eliminated from the present sample by the choice of houses that are broadly similar: they are all 'ordinary' houses (no flats, maisonettes, etc.), and all have 3, 4 or 5 bedrooms. The remaining sources of variation in price will contribute to the residual variation among houses in each town, and will be treated accordingly. We should also consider in what sense we can think of the latitudes of houses as 'explaining' the variation in their prices. The easily measurable variable 'latitude' is associated with many other variables, such as climate and distance from London, and it is probably some of these, rather than latitude per se, that have a causal connection with price. However, an explanatory variable does not have to be causally connected to the response in order to serve as a useful predictor. Finally, we should consider the value of studying the relationship between latitude and price in such a small sample. The data on this topic are extensive, and widely interpreted. However, this small data set, illustrating a simple, familiar example, is highly suitable for a study of the *methods* by which we judge the significance of a trend, estimate its magnitude, and place confidence limits on the estimate. Above all, this example will show that in order to do these things reliably, we must recognise that our observations – the houses – are not mutually independent, but form natural groups – the towns.

1.2 Simple regression analysis. Use of the software GenStat to perform the analysis

We will begin by performing a simple regression analysis on these data, before considering how this should be modified to take account of the grouping into towns. The standard linear *regression model* (Model 1.1) is

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij} \tag{1.1}$$

Table 1.1 Prices of houses in a sample of English towns, and their latitudes
Data obtained from an estate agents' web site in October 2004.

	A	B	C
1	town!	latitude	price_pounds
2	Bradford	53.7947	39950
3	Bradford	53.7947	59950
4	Bradford	53.7947	79950
5	Bradford	53.7947	79995
6	Bradford	53.7947	82500
7	Bradford	53.7947	105000
8	Bradford	53.7947	125000
9	Bradford	53.7947	139950
10	Bradford	53.7947	145000
11	Buxton	53.2591	120000
12	Buxton	53.2591	139950
13	Buxton	53.2591	149950
14	Buxton	53.2591	154950
15	Buxton	53.2591	159950
16	Buxton	53.2591	159950
17	Buxton	53.2591	175950
18	Buxton	53.2591	399950
19	Carlisle	54.8923	85000
20	Carlisle	54.8923	89950
21	Carlisle	54.8923	90000
22	Carlisle	54.8923	103000
23	Carlisle	54.8923	124950
24	Carlisle	54.8923	128500
25	Carlisle	54.8923	132500
26	Carlisle	54.8923	135000
27	Carlisle	54.8923	155000
28	Carlisle	54.8923	158000
29	Carlisle	54.8923	175000
30	Chichester	50.8377	199950
31	Chichester	50.8377	299250
32	Chichester	50.8377	350000
33	Crewe	53.0998	77500
34	Crewe	53.0998	84950
35	Crewe	53.0998	112500

(*Continued overleaf*)

Table 1.1 (*continued*)

	A	B	C
36	Crewe	53.0998	140000
37	Durham	54.7762	127950
38	Durham	54.7762	157000
39	Durham	54.7762	169950
40	Newbury	51.4037	172950
41	Newbury	51.4037	185000
42	Newbury	51.4037	189995
43	Newbury	51.4037	195000
44	Newbury	51.4037	295000
45	Newbury	51.4037	375000
46	Newbury	51.4037	400000
47	Newbury	51.4037	475000
48	Ripon	54.1356	140000
49	Ripon	54.1356	152000
50	Ripon	54.1356	187950
51	Ripon	54.1356	210000
52	Royal Leamington Spa	52.2876	147000
53	Royal Leamington Spa	52.2876	159950
54	Royal Leamington Spa	52.2876	182500
55	Royal Leamington Spa	52.2876	199950
56	Stoke-on-Trent	53.0041	69950
57	Stoke-on-Trent	53.0041	69950
58	Stoke-on-Trent	53.0041	75950
59	Stoke-on-Trent	53.0041	77500
60	Stoke-on-Trent	53.0041	87950
61	Stoke-on-Trent	53.0041	92000
62	Stoke-on-Trent	53.0041	94950
63	Witney	51.7871	179950
64	Witney	51.7871	189950
65	Witney	51.7871	220000

where

x_i = value of X (latitude) for the ith town,

y_{ij} = observed value of Y ($\log_{10}$(house price in pounds)) for the jth house in the ith town,

β_0, β_1 = constants to be estimated, defining the relationship between X and Y,

ε_{ij} = the residual effect, i.e. the deviation of y_{ij} from the value predicted on the basis of x_i, β_0 and β_1.

Note that in this model the house prices are transformed to logarithms, because preliminary exploration has shown that this gives a more linear relationship between latitude and price, and more uniform residual variation. The model is illustrated graphically in Figure 1.1.

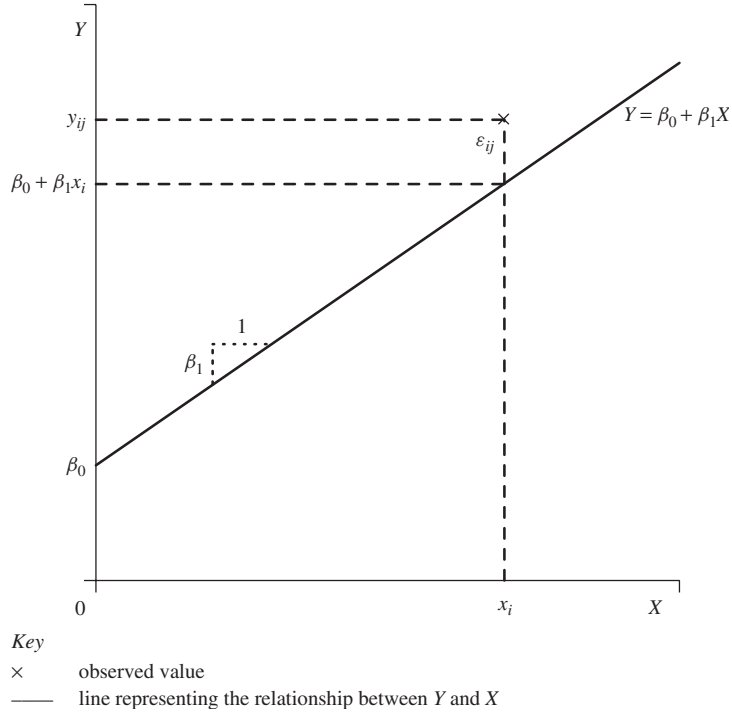

Figure 1.1 Linear relationship between an explanatory variable X and a response variable Y, with residual variation in the response variable.

The model specifies that a sloping straight line is to be used to describe the relationship between latitude and log(house price). The *parameters* β_0 and β_1 specify, respectively, the intercept and slope of this line. *Estimates* of these parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively, are to be obtained from the data, and these estimates will define the *line of best fit* through the data. An estimate of each of the ε_{ij}, designated $\hat{\varepsilon}_{ij}$, will be given by the deviation of the ijth data point, in a vertical direction, from the line of best fit. The parameter estimates chosen are those that minimise the sum of squares of the $\hat{\varepsilon}_{ij}$. It is assumed that the *true* values ε_{ij} are independent values of a variable E which is *normally distributed with mean zero and variance* σ^2. The meaning of this statement, which can be written in symbolic shorthand as

$$E \sim \mathrm{N}(0, \sigma^2),$$

is illustrated in Figure 1.2. The area under this curve between any two values of E gives the probability that a value of E will lie between these two values. For example,

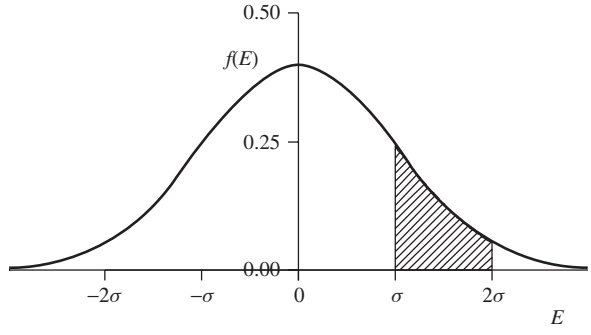

Figure 1.2 A normal distribution with mean zero and variance σ^2.
f(E) = probability density of E. Total area under curve = 1. Hatched area = probability that a value of E is greater than or equal to σ and less than 2σ.

the probability that a value of E will lie between σ and 2σ, represented by the hatched area in the figure, is 0.1359. Hence the total area under the curve is 1, as any value of E must be between minus infinity and plus infinity. The variable plotted on the vertical axis, f(E), is referred to as the *probability density*. It must be integrated over a range of values of E in order to give a value of probability, just as human population density must be integrated over an area of land in order to give a value of population. For the reader unfamiliar with such regression models, a fuller account is given by Draper and Smith (1998).

The calculations required in order to fit a regression model to data (i.e. to estimate the parameters of the model) can be performed by many pieces of computer software, and one of these, GenStat, will be referred to throughout this book. Information on obtaining access to GenStat is given in the preface of this book. The GenStat command language, used to specify the models to be fitted, provides a valuable tool for thinking clearly about these models, and the GenStat statements required will therefore be presented and discussed here. However, the details of a computer language should not be allowed to distract from the statistical concepts that are our central topic. We will therefore note only a few key points about these statements: a full introduction to the GenStat command language is given in Section 1.3 of GenStat's *Introduction* guide (Payne *et al*., 2003). This is available *via* GenStat's graphical user interface (GUI), which also gives access to the full specification of the language.

The following statements, in the GenStat command language, import the data into the GenStat environment and fit Model 1.1:

```
IMPORT \
   'Intro to Mixed Modelling\\Chapter 1\\house price, latitude.xls'; \
   SHEET = 'Sheet1'
CALCULATE logprice = log10(price_pounds)
MODEL logprice
FIT [FPROB = yes; TPROB = yes] latitude
```

The `IMPORT` statement specifies the file that contains the data, and makes the data available to GenStat. The `CALCULATE` statement performs the transformation to

logarithms, and stores the results in the variate 'logprice'. The `MODEL` statement specifies the response variable in the regression model (Y, logprice), and the `FIT` statement specifies the explanatory variable (X, latitude). The *option setting* '`FPROB = yes`' indicates that when an F statistic is presented, the associated probability is also to be given (see below). The option setting '`TPROB = yes`' indicates that this will also be done in the case of a t statistic. The same operations could be specified – perhaps more easily – using the menus and windows of GenStat's GUI: the use of these facilities is briefly illustrated in Section 1.12, and fully explained by Payne *et al*. (2003).

A researcher almost always receives the results of statistical analysis in the form of computer output, and the interpretation of this, the extraction of key pieces of information and their synthesis in a report are important statistical skills. The output produced by GenStat is therefore presented and interpreted here. That from the `FIT` statement is shown in the box below.

Regression analysis

Response variate: logprice
Fitted terms: Constant, latitude

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	0.710	0.70955	20.55	<.001
Residual	62	2.141	0.03453		
Total	63	2.850	0.04524		

Percentage variance accounted for 23.7
Standard error of observations is estimated to be 0.186.

Message: the following units have large standardized residuals.

Unit	Response	Residual
1	4.602	−2.72
17	5.602	2.46

Message: the residuals do not appear to be random; for example, fitted values in the range 5.009 to 5.074 are consistently smaller than observed values and fitted values in the range 5.162 to 5.231 are consistently larger than observed values.

Estimates of parameters

Parameter	estimate	s.e.	t(62)	t pr.
Constant	9.68	1.00	9.68	<.001
latitude	−0.0852	0.0188	−4.53	<.001

This output begins with a specification of the model fitted. Note that the fitted terms include not only the explanatory variable, latitude, but also a constant, although none was specified: any regression model includes a constant by default (β_0 in this case).

Next comes an analysis of variance (anova) table, which partitions the variation in log(house price) between the *terms* in Model 1.1, namely:

- the effect of latitude (represented in the row labelled 'Regression' in the anova table), and
- the residual effects (represented in the row labelled 'Residual').

After the names of the terms, the next two columns of the anova table hold the degrees of freedom (abbreviated to d.f. or DF) and the sum of squares (s.s. or SS) for each term. The methods for calculating these will not be given here (for an account, see Draper and Smith, 1998, Section 1.3, pp 28–34), but it should be noted that the degrees of freedom for each term represent the number of independent pieces of information to which that term is equivalent. Thus the effect of latitude is a single piece of information, and $\mathrm{DF}_{\mathrm{latitude}} = 1$. There are 64 houses in the sample, each of which gives a value of $\hat{\varepsilon}_{ij}$, so it might be thought that the 'Residual' term would comprise 64 pieces of information. However, two pieces of information have been 'used up' by the estimation of the intercept and the effect of latitude, so

$$\mathrm{DF}_{\mathrm{Residual}} = 64 - 2 = 62.$$

This reduction in the residual degrees of freedom is equivalent to that fact that a line of best fit based on only two observations passes exactly through the data points – such a data set provides no information on residual variation.

The mean square for each term (m.s. or MS) is given by SS/DF, and is an estimate of the part of the variance in log(house price) that is accounted for by the term. If there is no real effect of latitude on log(house price), the expected values of $\mathrm{MS}_{\mathrm{latitude}}$ and $\mathrm{MS}_{\mathrm{Residual}}$ are the same. Hence on this *null hypothesis* (H_0), the expected value of $\mathrm{MS}_{\mathrm{latitude}}/\mathrm{MS}_{\mathrm{Residual}}$ is 1, though the actual value will vary from one data set to another. This ratio, called the variance ratio (abbreviated to v.r.), thus provides a test of H_0. Provided that the residual variation is normally distributed (see Section 1.8), v.r. is also known as the F statistic. If H_0 is true, the distribution of F over an infinite population of samples (data sets) has a definite mathematical form. The precise shape of this distribution depends on the degrees of freedom in the numerator and denominator of the ratio: hence the variable F is referred to more precisely as $F_{\mathrm{DF}_{\mathrm{numerator}},\mathrm{DF}_{\mathrm{denominator}}}$. The distribution in the present case (i.e. the distribution of the variable $F_{1,62}$) is illustrated in Figure 1.3. This curve is interpreted in the same way as the normal distribution illustrated earlier. The area under the curve between any two values of F gives the probability that an observation of F will lie between these values. Hence again the total area under the curve is 1, as any observation of F must have some value between 0 and infinity. Again the variable plotted on the vertical axis is the probability density. This F distribution can be used to determine the probability P of obtaining by chance a value of F larger than that actually observed, as shown in the figure. For example, if $F_{1,62} > 4.00$, then $P < 0.05$, and it is said that the effect under consideration is significant at the 5 % level. Similarly, if $F_{1,62} > 7.06$, then $P < 0.01$, and it is said that the effect is significant at the 1 % level, i.e. highly significant. In the present case $F_{1,62} = 20.55$, and both the anova table and the figure show that P (called F pr. in the table) is less than 0.001: that is, the relationship between latitude and

log(house price) in this data set is very highly significant – provided that the model specified is correct. However, there is a diagnostic message which indicates that this may not be the case: GenStat has detected that the residuals do not appear to be random.

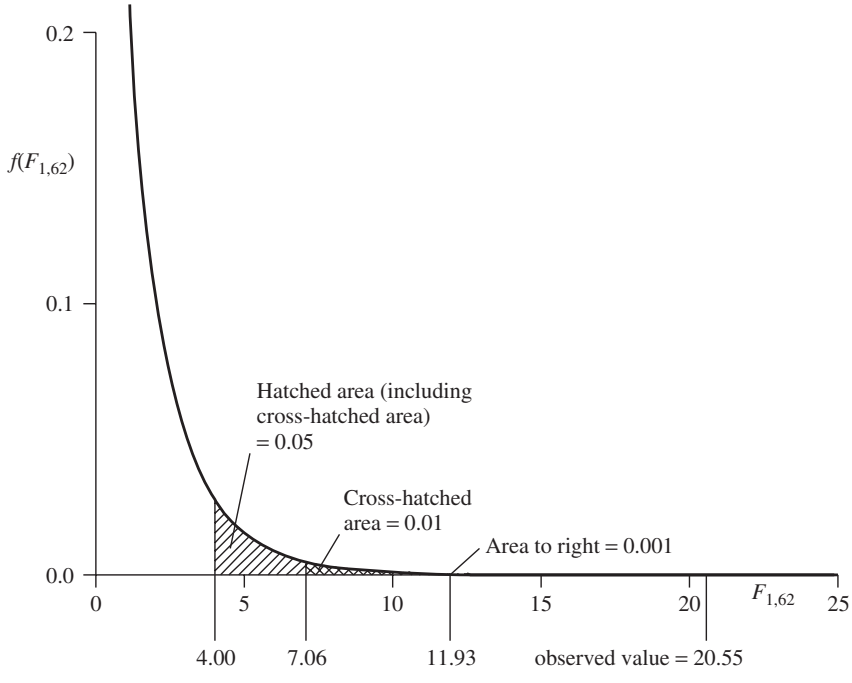

Figure 1.3 Distribution of the variable $F_{1,62}$, showing critical values for significance tests and the corresponding critical regions.

The next item in the output is the parameter estimates, which give the intercept ($\hat{\beta}_0$, the constant term) and slope ($\hat{\beta}_1$) of the line of best fit, with their standard errors (SEs). The negative slope indicates that house prices are higher in the south of England than in the north. The t statistic for the effect of latitude is given by estimate/$\text{SE}_{\text{estimate}} = -0.0852/0.0188 = -4.53$. Note that $t^2 = (-4.53)^2 = 20.55 = F$, and that for both these statistics the P value (when calculated to a greater degree of precision than is given by the GenStat output) is 0.0000271 – that is, the t test for the significance of the slope is equivalent to the F test in the analysis of variance.

The line of best fit is displayed, together with the data and the mean value for each town, in Figure 1.4. This figure shows that, overall, the regression line fits the data reasonably well. However, observations from the same town generally lie on the same side of the line – that is, the residual values within each town are not mutually independent. For example, as noted in GenStat's diagnostic message, the observations from Ripon, Durham and Carlisle generally lie above the line (GenStat identifies these by the fact that their fitted values of log(house price) lie in the range 5.009 to 5.074), whereas those from Stoke-on-Trent and Crewe all lie below the line (their fitted values lie in the range 5.162 to 5.231).

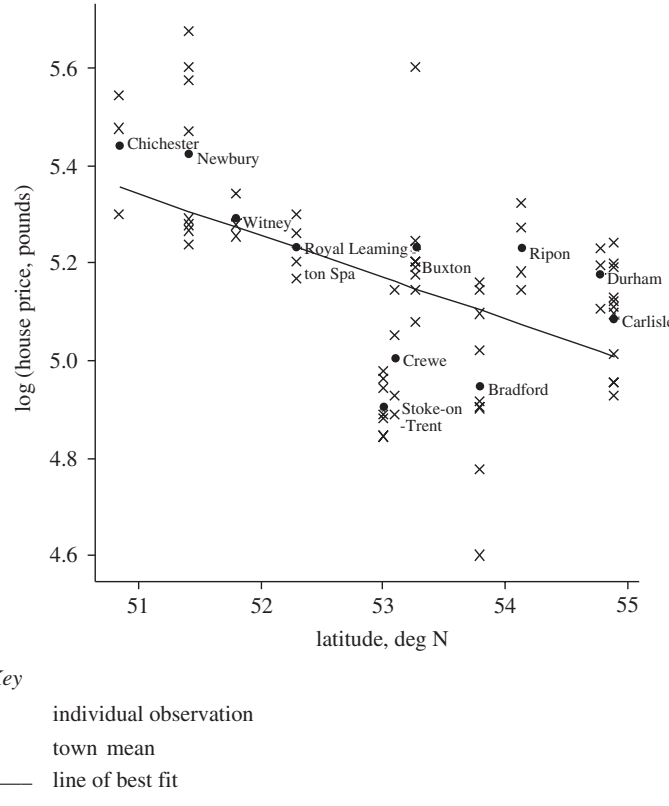

Figure 1.4 Relationship between latitude and house prices in a sample of English towns, showing individual observations, town means and line of best fit.

1.3 Regression analysis on the group means

Because the residual values are not mutually independent, the analysis presented above cannot be relied upon, even though the regression line appears reasonable. In particular, the residual degrees of freedom ($DF_{Residual} = 62$) are an overestimate: we deceive ourselves if we believe that the data comprise 64 independent observations. The simplest way to overcome this problem is to fit the regression model using the mean value of log(house price) for each town. These means are displayed in a spreadsheet in Table 1.2.

The following statements will perform a regression analysis on these mean values:

```
IMPORT \
   'Intro to Mixed Modelling\\Chapter 1\\price, lat, town means.xls'; \
   sheet = 'Sheet1'
MODEL meanlogprice
FIT [FPROB = yes; TPROB = yes] meanlatitude
```

Table 1.2 Mean values of $\log_{10}$(price) of houses in a sample of English towns, and their latitudes.

	A	B	C	D
1	town_unique	n_houses	meanlatitude	meanlogprice
2	Bradford	9	53.7947	4.94745
3	Buxton	8	53.2591	5.23083
4	Carlisle	11	54.8923	5.08552
5	Chichester	3	50.8377	5.44034
6	Crewe	4	53.0998	5.00394
7	Durham	3	54.7762	5.17775
8	Newbury	8	51.4037	5.42456
9	Ripon	4	54.1356	5.23106
10	Royal Leamington Spa	4	52.2876	5.23337
11	Stoke-on-Trent	7	53.0041	4.90642
12	Witney	3	51.7871	5.29207

The output of the `FIT` statement is as follows:

Regression analysis

Response variate: meanlogprice
Fitted terms: Constant, meanlatitude

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	0.1160	0.11601	5.19	0.049
Residual	9	0.2010	0.02234		
Total	10	0.3170	0.03170		

Percentage variance accounted for 29.5
Standard error of observations is estimated to be 0.149.

Estimates of parameters

Parameter	estimate	s.e.	t(9)	t pr.
Constant	9.44	1.87	5.04	<.001
meanlatitude	−0.0804	0.0353	−2.28	0.049

The residual degrees of freedom are much fewer ($DF_{Residual} = 9$), reflecting the more reasonable assumption that each town can be considered as an independent observation. Consequently, the relationship between latitude and house price is much less significant: $P = 0.049$, compared with $P < 0.001$ in the previous analysis. However, the estimates of the intercept and slope of the regression line are not much altered.

1.4 A regression model with a term for the groups

The simple method, presented in the previous section, for dealing with the problem of several observations in each town gives no account of the variation within the towns. Nor does it take account of the variation in the number of houses sampled, and hence in the precision of the mean, from one town to another. An alternative approach, which overcomes these deficiencies, is to add a term to the regression model to take account of the variation among towns that is not accounted for by latitude, namely:

$$y_{ij} = \beta_0 + \beta_1 x_i + \tau_i + \varepsilon_{ij} \tag{1.2}$$

where
τ_i = mean deviation from the regression line of observations from the ith town.

The following statements will fit this model (Model 1.2) using the ordinary methods of regression analysis:

```
IMPORT \
   'Intro to Mixed Modelling\\Chapter 1\\house price, latitude.xls'; \
   sheet = 'Sheet1'
CALCULATE logprice = log10(price_pounds)
MODEL logprice
FIT [FPROB = yes; TPROB = yes; \
   PRINT = model, estimates, accumulated] \
   latitude + town
```

The output of the `FIT` statement is as follows. It is quite voluminous, and each part will be discussed before presenting the next.

Message: term town cannot be fully included in the model because 1 parameter is aliased with terms already in the model.

(town Witney) = 26.80 − (latitude)*0.4981 − (town Buxton)*0.2668 + (town Carlisle)* 0.5467 − (town Chichester)*1.473 − (town Crewe)*0.3461 + (town Durham)*0.4889 −(town Newbury)*1.191 + (town Ripon)*0.1698 − (town Royal Leamington Spa)* 0.7507 − (town Stoke-on-Trent)*0.3938

First comes a message noting that because the variation among the town means is partly accounted for by latitude, the term 'town' cannot be fully included in the regression model. It is said to be *partially aliased* with latitude. (The technical consequence of this partial aliasing is that the effect of one of the towns – Witney, arbitrarily chosen because it comes last when the towns are arranged in alphabetical order – is a function of the effects of the other towns and of latitude, but the numerical details of this relationship, given in the message, need not concern us.)

Next come the statement of the regression model and the estimates of its parameters:

Regression analysis

Response variate: logprice
Fitted terms: Constant + latitude + town

Estimates of parameters

Parameter	estimate	s.e.	t(53)	t pr.
Constant	14.18	2.32	6.12	<.001
latitude	−0.1717	0.0435	−3.95	<.001
town Buxton	0.1914	0.0598	3.20	0.002
town Carlisle	0.3265	0.0885	3.69	<.001
town Chichester	−0.015	0.136	−0.11	0.914
town Crewe	−0.0628	0.0761	−0.83	0.413
town Durham	0.399	0.106	3.75	<.001
town Newbury	0.067	0.102	0.66	0.515
town Ripon	0.3421	0.0840	4.07	<.001
town Royal Leamington Spa	0.0272	0.0874	0.31	0.757
town Stoke-on-Trent	−0.1767	0.0636	−2.78	0.007
town Witney	0	*	*	*

Parameters for factors are differences compared with the reference level:
Factor Reference level
town Bradford

These parameter estimates correctly lead to the sample mean for each town when substituted into the formula

$$\text{mean (log(house price))} = 14.18 - 0.1717 \times \text{latitude} + \text{effect of town.}$$

For example, in Durham,

$$\text{mean (log(house price))} = 14.18 - 0.1717 \times 54.7762 + 0.399 = 5.1739.$$

However, the parameter estimates themselves are arbitrary and uninformative, as illustrated in Figure 1.5. The fitted line is arbitrarily specified to pass through the mean values for Bradford and Witney (the first and last towns in the alphabetic sequence), and the effects of the other towns – their vertical distances from the fitted line – are determined accordingly.

The option setting 'PRINT = accumulated' in the `FIT` statement specifies that the output should include an accumulated anova, which partitions the variation accounted for by the model between its two terms, namely:

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ latitude	1	0.70955	0.70955	41.35	<.001
+ town	9	1.23119	0.13680	7.97	<.001
Residual	53	0.90937	0.01716		
Total	63	2.85011	0.04524		

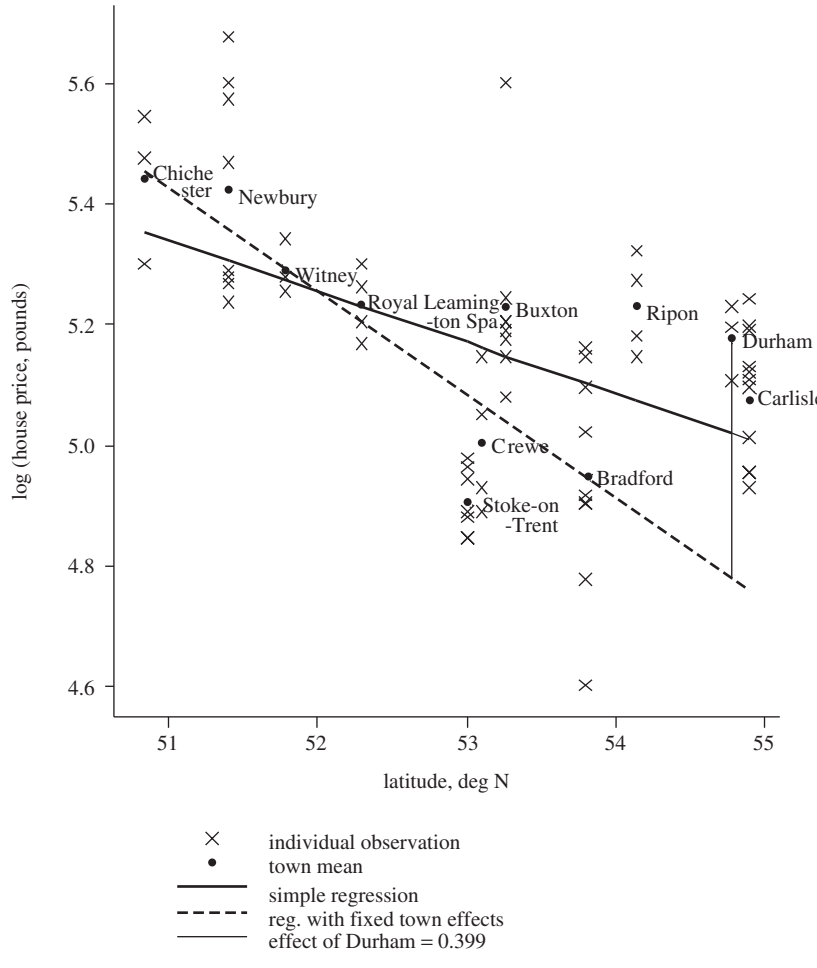

Figure 1.5 Relationship between latitude and house prices in a sample of English towns, comparing the lines of best fit from simple regression analysis and from analysis with town effects treated as fixed.

Despite the arbitrary nature of the parameter estimates, this anova is informative. The term 'latitude' represents the part of the variation in house price that is due to the effect of latitude, and the term 'town' represents the part due to the variation among towns after allowing for latitude; that is, the deviations of the town means from the original line of best fit (not the arbitrary fitted line presented above). Note that the mean square (MS) for latitude is the same as the corresponding value from Model 1.1:

$$\mathrm{MS}_{\text{latitude}} = \mathrm{MS}_{\text{Regression,Model 1.1}} = 0.70955.$$

That is, the amount of variation explained by latitude is consistent between Models 1.1 and 1.2. Note also that

$$\mathrm{SS}_{\text{Total,Model 1.1}} = \mathrm{SS}_{\text{Total,Model 1.2}} = 2.850, \text{ allowing for rounding.}$$

Hence Model 1.2 represents a partitioning of the residual term in Model 1.1. Thus

$$\mathrm{DF}_{\text{Residual,Model 1.1}} = \mathrm{DF}_{\text{town}} + \mathrm{DF}_{\text{Residual,Model 1.2}}$$
$$62 = 9 + 53$$

and

$$\mathrm{SS}_{\text{Residual,Model 1.1}} = \mathrm{SS}_{\text{town}} + \mathrm{SS}_{\text{Residual,Model 1.2}}$$
$$2.141 = 1.23119 + 0.90937, \text{ allowing for rounding.}$$

Part of the variation, formerly unexplained, is now attributed to the effects of towns.

1.5 Construction of the appropriate *F* test for the significance of the explanatory variable when groups are present

The accumulated anova can be adapted to provide a realistic assessment of the significance of the effect of latitude. In the GenStat output, both of the mean squares for model terms in this anova are tested against the residual mean square:

- $F_{\text{latitude}} = \mathrm{MS}_{\text{latitude}}/\mathrm{MS}_{\text{Residual}} = 0.70955/0.01716 = 41.35$
- $F_{\text{town}} = \mathrm{MS}_{\text{town}}/\mathrm{MS}_{\text{Residual}} = 0.13680/0.01716 = 7.97.$

F_{town} is highly significant ($P < 0.001$), confirming that there is real variation among the towns in addition to that accounted for by latitude. Consequently, F_{latitude} is misleading: it is much larger than the value of 5.19 obtained when the regression is fitted using the town means. In order to obtain the appropriate value from the present analysis, we need to calculate

$$F_{\text{latitude}} = \mathrm{MS}_{\text{latitude}}/\mathrm{MS}_{\text{town}} = 0.70955/0.13680 = 5.18677.$$

This is almost exactly equivalent to the F statistic for latitude in the analysis based on the town means, as shown in Table 1.3. (The GenStat output does not give these

Table 1.3 Comparison between the F statistics from regression analyses based on individual houses and on town means.

	Regression based on town means	Regression with term for towns (Model 1.2)
Formula for F_{latitude}	$\mathrm{MS}_{\text{Regression}}/\mathrm{MS}_{\text{Residual}}$	$\mathrm{MS}_{\text{latitude}}/\mathrm{MS}_{\text{town}}$
Numerical values	$0.11601/0.02234 = 5.19293$	$0.70955/0.13680 = 5.18677$
$\mathrm{DF}_{\text{Numerator}}$	1	1
$\mathrm{DF}_{\text{Denominator}}$	9	9
P	0.0487	0.0488

F and P values to sufficient precision to reveal the slight differences between them.) The reason why the two sets of values do not agree exactly will be explained in Section 1.9.

1.6 The decision to regard a model term as random: a mixed model

The F values presented in Table 1.3 are based on a comparison of the variation explained by the regression line with the variation of *the town means* about the line, whereas the much larger value in the accumulated anova ($F_{\text{latitude}} = 41.35$) is based on a comparison with the variation of *individual values of log(house price)* about their respective town means. When we use the variation of the town means as the basis of comparison, we are regarding these means as values of a *random variable* – that is, we are considering the towns in this study as a representative sample from a large population of towns. Formally speaking, we are assuming that the values τ_i are independent values of a variable T, such that

$$T \sim \mathrm{N}(0, \sigma_T^2)$$

– that is, they are very like the values ε_{ij}, except that their variance, σ_T^2, is different. This may seem a radical assumption, but it is necessary to assume that the τ_i are values of a random variable if we are to make any general statement about the relationship between latitude and house prices. (The assumption that this variable has a normal distribution is not a requirement, but other distributions require more advanced modelling methods – see Chapter 9.) If we were to drop one town from the study and replace it with newly chosen town, this would have an effect on the slope of the regression line – a larger effect than would be produced by simply taking a new sample of houses from within the same town. This is because there is more variation among town means than among individual houses from the same town, even after allowing for the effect of latitude: this is made clear by the highly significant value of F_{town}, 7.97. If we insist on regarding our choice of towns as fixed, any inference concerning the relationship between house price and latitude, its magnitude and our confidence that it is real, must be confined to these particular towns. This limitation is reflected by the arbitrary way in which GenStat defines the parameters of Model 1.2. In the language of mixed modelling, in order to obtain a realistic estimate of the effect of latitude on house prices, we must regard the effect of each town as a *random effect*, and town as a *random-effect term* in the regression model. The effect of latitude itself is a non-random or *fixed effect*. Since our model contains effects of both types, it is a *mixed model*.

When deciding whether to regard a factor as random, the essential question to ask is whether the levels studied can be regarded as a representative sample of some large population of levels. In the present case, can the towns sampled be considered representative of the population of English towns? This question of how to determine which model terms should be regarded as random will be discussed more fully, in the context of a wide range of models, in Chapter 6 (Section 6.3).

1.7 Comparison of the tests in a mixed model with a test of lack of fit

The partitioning of the variation around the regression line into two components, one due to the deviations of the town means from the line and the other to the deviations of individual houses from the town means, is similar to the test for lack of fit described by Draper and Smith (1998, Chapter 2, Section 2.1, pp 49–53). Indeed, the calculations performed to obtain the mean squares are identical in the two analyses. However, there is an important difference in the ideas that underlie them, and consequently in the F tests specified, as illustrated in Table 1.4. The angled lines in this table indicate the pairs of mean squares that are compared by the two F tests. The purpose of the test of lack of fit is to determine whether the variation among groups of observations at the same value of X (towns in the present case) is significant, or whether it can be absorbed into the 'Residual' term. In the mixed-model analysis, on the other hand, the reality of the variation among towns is not in doubt. The question is whether the effect of latitude is significant, or whether it can be absorbed into the 'town' term.

Even if the test of lack of fit leads to the conclusion that the variation among towns is significant, the two analyses are not equivalent. Because the test of lack of fit treats only the 'Residual' term as random, it leads to the use of this term as the denominator in all F tests, whereas the mixed model leads to the use of the random-effect term 'town' as the denominator against which to test the significance of the effect of latitude. The full set of tests specified by the two analyses is therefore as shown in Table 1.5. The dotted angled lines in this table indicate the pairs of mean

Table 1.4 Comparison between the F test for lack of fit and the mixed-model F test in the analysis of the effect of latitude on house prices in England.

Source of variation	DF	MS	Test of lack of fit		Mixed-model test	
			F	*P*	*F*	*P*
latitude	1	0.70955			5.19	0.0488
town	9	0.13680	7.97	<0.001		
Residual	53	0.01716				

Table 1.5 Comparison between the F tests conducted in ordinary multiple regression analysis and in mixed-model analysis of the effects of latitude and town on house prices in England.

Source of variation	DF	MS	Analysis with test of lack of fit		Mixed-model analysis	
			F	*P*	*F*	*P*
latitude	1	0.70955	41.35	<0.001	5.19	0.0488
town	9	0.13680	7.97	<0.001	7.97	<0.001
Residual	53	0.01716				

squares that are compared by the additional F tests. In the case considered by Draper and Smith the additional term required (equivalent to 'town' in the present example) is a quadratic one, to allow for curvature in the response to the explanatory variable, and in this situation the test of lack of fit is correct.

1.8 The use of residual maximum likelihood to fit the mixed model

So far we have taken an improvised approach to fitting the mixed model. We have:

- fitted two regression models, one without a term for the effects of towns and the other including such a term;
- taken the estimate of the slope from the first model and the mean squares from the second;
- obtained the F statistic to test the significance of the effect of latitude from the mean squares by hand.

However, the mixed-model analysis can be performed in a more unified manner using the criterion of *residual maximum likelihood* (REML, also known as *restricted maximum likelihood*). The formal meaning of this criterion will be explained in Chapter 10: here, we will simply apply it to the present data. The following GenStat statements specify the mixed model to be fitted:

```
VCOMPONENTS [FIXED = latitude; CADJUST = none] RANDOM = town
REML [PRINT = model, components, Wald, effects] logprice
```

The `VCOMPONENTS` statement specifies the terms in the model: it is equivalent to the `FIT` statement in an ordinary regression analysis. The option `FIXED` specifies the fixed-effect term or terms: in this case, 'latitude'. The constant (the intercept β_0) is also included in the model as a fixed-effect term by default: it does not have to be explicitly specified. The option setting '`CADJUST = none`' indicates that no adjustment is to be made to the covariate 'latitude' before analysis: by default, a covariate is *centred* by subtracting its mean value from each of its values (see Chapter 7, Section 7.2). The *parameter* `RANDOM` specifies the random-effect term(s): in this case, 'town'. The `REML` statement specifies the response variate whose variation is to be explained by the model: in this case, 'logprice'. It is equivalent to the `MODEL` statement in an ordinary regression analysis. Note that the `VCOMPONENTS` and `REML` statements are given in the opposite order to their equivalents in ordinary regression analysis. The `PRINT` option indicates what results from the model-fitting process are to be presented in the output.