5

Estimation of random effects in mixed models: Best Linear Unbiased Predictors (BLUPs)

5.1 The difference between the estimates of fixed and random effects

In ordinary regression analysis and analysis of variance, the random-effect terms are always *nuisance variables*: residual variation or block effects. The effects of individual levels of such terms are not of interest. However, we have seen that in a mixed model, effects that are of intrinsic interest may be specified as random – for example, the effects of the individual breeding lines in the barley field trial discussed in Chapter 3. The decision to specify a term as random causes a fundamental change in the way in which the effect of each level of that term is estimated. We will illustrate this change and its consequences in the context of the barley breeding lines. However, the concepts introduced and the arguments presented apply equally to any situation in which replicated, quantitative evaluations are available for the comparison of members of some population – for example, new chemical entities to be evaluated as potential medicines by a pharmaceutical company, or candidates for admission to a university on the basis of their examination scores.

It is easy to illustrate the relationship between the estimates of fixed and random effects in data that are grouped by a single factor – for example, a fully randomized design leading to a one-way anova. The data from the barley field trial are classified by two factors, line and block. Fortunately, however, the effects of blocks are negligible (Section 3.10). We will therefore treat this experiment as having a single-factor design, reducing Model 3.24 from

$$y_{ij} = \mu + \delta_i + \epsilon_{ij} + \phi_{k|ij}$$

to Model 5.1, namely

$$y_j = \mu + \phi_{k|j} + \epsilon_j \tag{5.1}$$

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

Companion website: http://www.wiley.com/go/beyond_regression

where the block effect δ_i is omitted and

y_j = the grain yield of the jth plot,
μ = the grand mean (overall mean) value of grain yield,
δ_i = the effect of the ith block,
$\phi_{k|j}$ = the effect of the kth breeding line, being the line sown in the jth plot,
ϵ_j = the residual effect of the jth plot.

We can then compare an analysis in which the variation among lines is specified as a fixed-effect term with one in which it is specified as a random-effect term. In the analysis of designed experiments, it is not normal practice to omit features of the design from the model fitted: it is justified in the present case solely in the interests of clarity.

The variation among lines in the original, 'unpadded' data set can be modelled as a fixed-effect term using GenStat's analysis of variance directives, as follows:

```
IMPORT 'IMM edn 2\\Ch 3\\barley progeny.xlsx'; SHEET = 'Sheet1'
BLOCKSTRUCTURE
TREATMENTSTRUCTURE line
ANOVA [FPROBABILITY = yes] yield_g_m2
```

The same model can be fitted using the mixed-modelling directives, with no random-effects model specified, as follows:

```
VCOMPONENTS [FIXED = line]
REML [PRINT = model, components, deviance, means; \
   PTERMS = 'constant' + line] yield_g_m2
```

In order to specify 'line' as a random-effects term, it is moved to the random-effects model in the VCOMPONENTS statement; thus:

```
VCOMPONENTS RANDOM = line
REML [PRINT = model, components, deviance, means, effects; \
   PTERMS = 'constant'+ line; METHOD = Fisher] yield_g_m2
```

In order to obtain standard errors (SEs) of the differences between means, the option setting 'METHOD = Fisher' must be used (see Sections 2.5 and 11.10).

The breeding-line means obtained from these two analyses are compared in Table 5.1. They are not the same: in the case of a high-yielding line such as Line 7, the random-effect mean is lower than the fixed-effect mean, whereas for a low-yielding line such as Line 16, the opposite is the case. The fixed-effect means are the simple means of the observations for the line in question. For example, the mean for Line 7 is

$$\frac{907.38 + 820.08}{2} = 863.73,$$

whereas that for Line 3, which occurs only in Block 1, is the single plot value 873.04. Each of these means is taken as an estimate of the true mean yield of the breeding line in question. But the plant breeder knows that if he/she selects the highest-yielding lines this year for further evaluation, he/she is selecting on both genetic and environmental variation. This year's environmental component will not contribute to the selected lines' yield next year, and consequently, their mean yield will generally be somewhat lower. We have already seen how this

Table 5.1 Comparison between the estimates of mean yields of breeding lines of barley obtained when breeding line is specified as either a fixed- or a random-effect model term.

Line	Fixed-effect mean	Random-effect mean	Line	Fixed-effect mean	Random-effect mean
1	654.5	639.9	43	757.6	724.8
2	483.3	510.2	44	695.4	673.6
3	873.0	782.5	45	547.8	552.2
4	719.1	674.9	46	675.6	657.3
5	799.0	730.7	47	192.4	306.9
6	802.4	761.6	48	715.6	690.2
7	863.7	812.1	49	826.9	781.8
8	468.2	486.7	50	803.4	762.4
9	681.0	661.7	51	719.3	693.3
10	760.8	727.4	52	603.0	593.8
11	580.8	579.4	53	559.1	563.2
12	436.6	477.5	54	834.4	787.9
13	508.1	519.5	55	555.3	560.5
14	922.7	860.6	56	479.1	495.7
15	600.1	591.7	57	182.0	299.6
16	225.4	286.9	58	478.5	495.2
17	755.5	723.1	59	594.0	590.2
18	315.8	361.3	60	428.8	454.3
19	523.2	532.0	61	542.0	547.4
20	632.5	614.4	62	730.5	682.9
21	514.6	532.1	63	659.4	644.0
22	340.7	410.5	64	664.8	648.4
23	711.6	686.9	65	766.1	731.7
24	567.2	568.2	66	767.0	732.5
25	428.6	472.0	67	554.9	558.0
26	609.8	603.2	68	355.7	394.2
27	747.1	716.2	69	423.1	449.6
28	679.9	647.5	70	950.5	883.5
29	769.4	734.4	71	582.7	580.9
30	353.3	419.3	72	212.2	276.1
31	196.4	263.1	73	355.3	420.7
32	753.1	721.0	74	520.4	529.6
33	808.5	766.7	75	617.3	603.8
34	656.3	641.4	76	344.5	413.2
35	464.5	483.6	77	357.9	422.6
36	927.5	864.5	78	260.1	315.5
37	724.0	697.1	79	378.4	412.8
38	591.0	587.8	80	483.0	498.9
39	179.3	249.0	81	309.7	356.3
40	489.6	504.3	82	251.3	348.1
41	429.9	455.2	83	280.4	368.4
42	673.0	655.2			

leads to an expected genetic advance under selection that is smaller than the difference between the mean of the selected lines and that of the full set of lines (Section 3.15). Mixed-model analysis provides a way of building the pessimism of the plant breeder more fully into the formal analysis of the data, giving a similar adjustment to the mean of each individual breeding line. This adjusted mean is the random-effect mean.

5.2 The method for estimation of random effects: The best linear unbiased predictor (BLUP) or 'shrunk estimate'

The adjustment to obtain the random-effect mean is made as follows. Following Model 5.1, the true mean of the kth breeding line is represented by

$$\mu_k = \mu + \phi_k. \tag{5.2}$$

In the fixed-effect means in Table 5.1, this value is estimated by

$$\widehat{\mu}_k = \frac{\sum_{j=1}^{r_k} y_{kj}}{r_k} \tag{5.3}$$

where

y_{kj} = the jth observation of the kth breeding line,
r_k = the number of observations of the kth breeding line.

The overall mean of the population of breeding lines, μ, is estimated by mixed modelling as about

$$\widehat{\mu} = 572.6$$

(Chapter 3 – GenStat: Section 3.10, Constant = 572.6; R: Section 3.16, Intercept = 572.47; SAS: Section 3.17, Intercept = 572.65). Note that this is not quite the same as the mean of all the observations (=581.6) or the mean of the line means (=569.1). Although we are treating this estimate as given, and using it to explain the estimation of the effects of individual lines, these two estimation steps are really interconnected in a single process. Something more will be said about the estimation of μ in a moment.

Rearranging Equation 5.2, we obtain

$$\phi_k = \mu_k - \mu. \tag{5.4}$$

Similarly, an estimate of ϕ_k is given by

$$\widehat{\phi}_k = \widehat{\mu}_k - \widehat{\mu}. \tag{5.5}$$

This ordinary estimate of the difference between a treatment mean and the overall mean is called the *Best Linear Unbiased Estimate* (*BLUE*). To allow for the expectation that high-yielding lines in the present trial will perform less well in a future trial – and that low-yielding lines will perform better – the BLUE can be replaced by a 'shrunk estimate' called the *Best Linear Unbiased Predictor* (*BLUP*). The formula for the required shrinkage is

$$\text{BLUP}_k = \text{BLUE}_k \cdot \text{shrinkage factor}_k = (\widehat{\mu}_k - \widehat{\mu}) \cdot \left(\frac{\widehat{\sigma}_G^2}{\widehat{\sigma}_G^2 + \frac{\widehat{\sigma}_E^2}{r_k}} \right). \tag{5.6}$$

The estimated variance components in this equation, $\widehat{\sigma}^2_G$ and $\widehat{\sigma}^2_E$, are obtained by mixed modelling as described in Section 3.10. Note that the shrinkage factor is the same as the estimated heritability defined in Equation 3.33, except that the average number of replications per line is replaced by the actual number for the line under consideration. The relationship in Equation 5.6, combined with the constraint

$$\sum_{k=1}^{p} \text{BLUP}_k = 0, \tag{5.7}$$

where
p = number of breeding lines,

determines the value of $\widehat{\mu}$ as well as those of the BLUPs. For Line 7, substituting the values given by Model 5.1, we obtain:

$$\text{BLUP}_7 = (863.7 - 572.5) \cdot \left(\frac{30667}{30667 + \frac{13226}{2}} \right) = 239.54$$

(Note that the variance component estimates are slightly different from those obtained in Section 3.10, because the block term has been dropped in Model 5.1.) A new estimate of the mean for the kth breeding line is then given by

$$\widehat{\mu}'_k = \widehat{\mu} + \text{BLUP}_k. \tag{5.8}$$

For Line 7,

$$\widehat{\mu}'_7 = 572.5 + 239.54 = 812.04.$$

As Line 7 is relatively high-yielding, its shrunk mean, 812.04, is lower than its unadjusted mean, 863.7.

The original estimates (BLUE_k and $\widehat{\mu}_k$) are compared with the shrunk estimates (BLUP_k and $\widehat{\mu}'_k$) for an arbitrary subset of the breeding lines (about a quarter of the total) in Figure 5.1. Note that BLUP_k is shrunk towards zero relative to BLUE_k, and $\widehat{\mu}'_k$ is correspondingly shrunk towards the estimate of the overall mean. In practice, it is usually the values at one extreme that are of interest – for example, the high-yielding breeding lines. Values far from the grand mean are shrunk more than those close to the mean. As should be expected, the amount of shrinkage specified by Equation 5.6 is large when

- the genetic variance, $\widehat{\sigma}^2_G$, is small
- the environmental variance, $\widehat{\sigma}^2_E$, is large
- the number of replications of the breeding line under consideration, r_k, is small.

Provided that the number of replications is constant over breeding lines, the shrinkage of BLUPs does not change the ranking of the means. However, if the number of replications is unequal, *crossovers* may occur, as in the present case, where Line 3 is estimated to be higher yielding than Line 7 on the basis of their BLUEs, but lower yielding on the basis of their BLUPs.

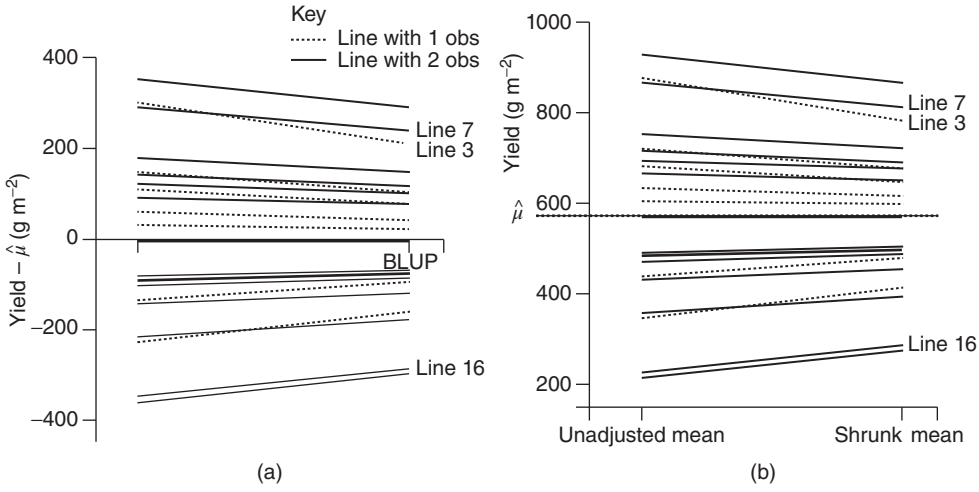

Figure 5.1 (a,b) Comparison between BLUEs and BLUPs, and between unadjusted and shrunk means, for yields of breeding lines of barley.

5.3 The relationship between the shrinkage of BLUPs and regression towards the mean

The relationship between the shrunk and unadjusted means can also be illustrated by a scatter diagram, as shown in Figure 5.2. Each point represents a breeding line. The shrinkage towards the overall mean is indicated by the fact that the points representing breeding lines that have an estimated yield above $\hat{\mu}$ lie below the line

$$\text{shrunk mean} = \text{unadjusted mean},$$

whereas those representing breeding lines with an estimated yield below $\hat{\mu}$ lie above this line. That is, the points lie approximately along a line that is flatter than this line. Moreover, points based on a single observation lie on a flatter line than those based on two observations. The crossover of Lines 3 and 7 is indicated by the fact that the point representing Line 3 lies below, but to the right of, that representing Line 7.

The flattening of the line of points on this scatter diagram is reminiscent of the commonly observed phenomenon of *regression towards the mean*, and this is no accident. An exploration of the connection between the two phenomena will help to clarify the distinction between the BLUE and the BLUP, and the sense in which each can be regarded as a 'best' statistic.

Suppose that the values of σ_G^2 and σ_E^2 for a population of breeding lines are known with considerable precision from experiments like that just described. Then suppose that the yield of a large number of new breeding lines, drawn at random from the same underlying population, is measured in an experiment with a single replication. Though much is known about the population, not much is known about the new lines individually – only the information from a single observation on each. If another experiment were performed on the same large sample of breeding lines, a second observation would be obtained on each. The relationship between the present and future observations (designated Y_{obs} and Y_{new}, respectively) would

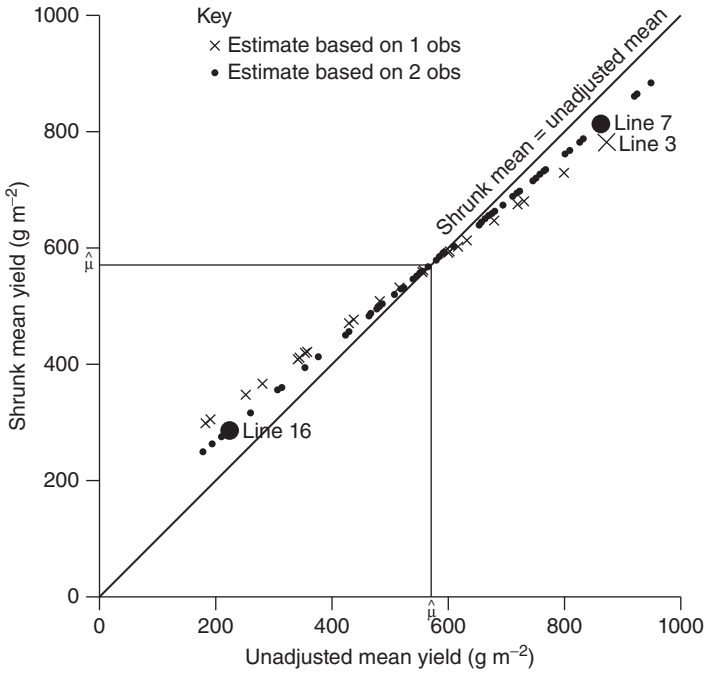

Figure 5.2 An alternative representation of the comparison between unadjusted and shrunk mean yields of breeding lines of barley.

then be as shown in Figure 5.3. Each point represents the value of the present observation on an individual breeding line, and a possible value for the future observation. The figure shows that the performance of each breeding line in the future experiment can be predicted from the available information, but not very accurately. High-yielding lines in the present experiment will generally give a high yield in the future experiment, and low-yielding lines a low future yield, but this relationship, which is due to the genetic value of each line, is blurred by the environmental component of each observation, present and future. The distribution of points in the figure – the probability distribution – is summarized by the ellipses. These are contour lines, each indicating a path along which the density of points (the probability density) is constant – a high density on the inner ellipse, a low density on the outer one. As all these density contours are the same shape, and concentric, a single contour is sufficient to indicate the general shape of the distribution, and this convention will be used in subsequent figures.

The criterion for the conversion of a BLUE to a BLUP, given in Equation 5.6, can also be represented graphically on this probability distribution, as follows. Consider Figure 5.4. The variance of the observed values in the present experiment is given by

$$\mathrm{var}(Y_{\mathrm{obs}}) = \sigma_G^2 + \sigma_E^2, \tag{5.9}$$

and in a future experiment $\mathrm{var}(Y_{\mathrm{new}})$ will be the same. The genetic component of each existing observation contributes to new observations on the same breeding line, but the environmental component does not: hence the covariance between Y_{obs} and Y_{new} is given by

$$\mathrm{cov}(Y_{\mathrm{obs}}, Y_{\mathrm{new}}) = \sigma_G^2. \tag{5.10}$$

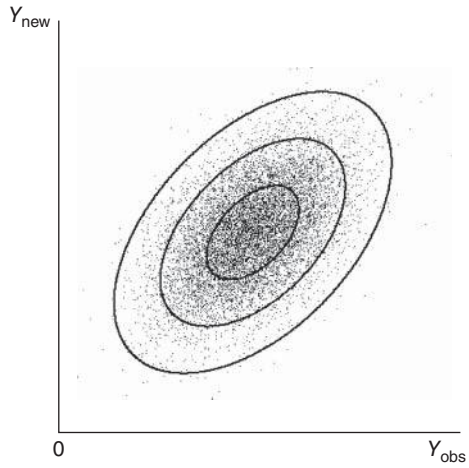

Figure 5.3 Joint distribution of present and future observations of yields of breeding lines of barley, when a single observation has been made on each line.
For explanation of conventions, see text.

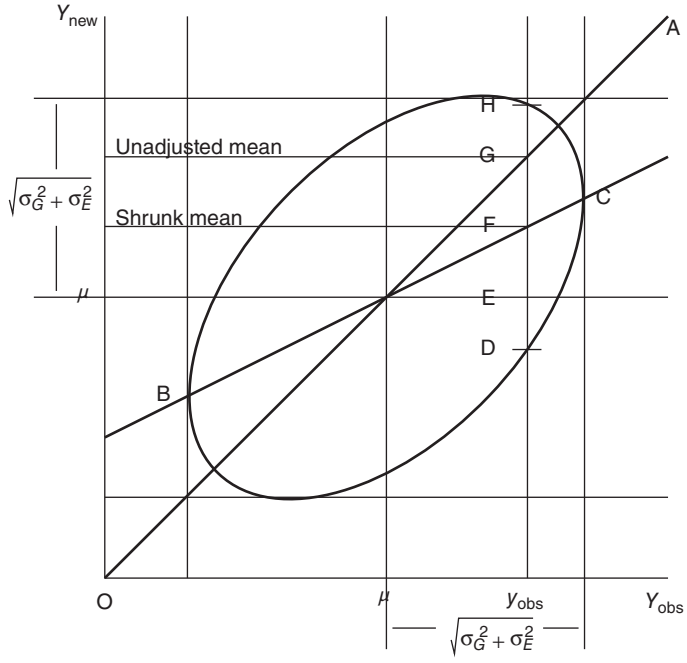

Figure 5.4 Summary of the relationship between present and future observations of yields of breeding lines of barley, when a single observation has been made on each line, showing the criterion for the conversion of a BLUE to a BLUP.
For an explanation of the annotations on this figure, see the text.

These variance and covariance values specify the bivariate-Normal distribution of Y_{obs} and Y_{new}, and the ellipse in the figure represents the 1-standard-deviation probability-density contour of this distribution. The simplest prediction of the future yield of any breeding line is its yield in the present experiment. This prediction is given by the equation

$$Y_{\text{new}} = Y_{\text{obs}},$$

which is represented by the line OA on the figure. For a breeding line that gives a yield y_{obs}, the point G is the corresponding position on OA, and the vertical coordinate of this point gives the prediction of the line's future yield. This is the unadjusted mean (based on a single observation in the present case), and the corresponding BLUE is given by its difference from the overall mean μ, that is, the distance EG. But the figure shows that this will not be the mean future yield of all the lines that give a yield of y_{obs} in the present experiment. The distances DF and FH are equal: hence the distribution of Y_{new}, *among those lines for which*

$$Y_{\text{obs}} = y_{\text{obs}},$$

is symmetrical about the point F, and the vertical coordinate of this point is the mean value of this distribution. This is the shrunk mean, and the corresponding BLUP is given by its difference from μ, that is, the distance EF. The point F lies on the line BC, which connects the left-most and right-most points on the ellipse, and which is the line of best fit obtained when Y_{obs} is treated as the explanatory variable, and Y_{new} as the response variable, in a regression analysis. For values of Y_{obs} above μ, BC lies below OA and the expected value of Y_{new} is less than Y_{obs}. Conversely, for values of Y_{obs} below μ, BC lies above OA and the expected value of Y_{new} is greater than Y_{obs}. This is the phenomenon of regression towards the mean noted by Francis Galton, which gave regression analysis its name and which was mistakenly interpreted as indicating that, over time, a biological population would converge to mediocrity unless steps were taken to prevent this. The connection between the shrinkage of BLUPs and regression towards the mean is also noted – together with connections to many other areas of statistics – in the classic paper by Robinson (1991, Section 5.2).

Now consider the situation when prediction is based not on a single observation of each breeding line, Y_{obs}, but on the mean of r observations, designated $\overline{Y}_{\text{obs}}$. This case is illustrated, together with the previous one, in Figure 5.5. We note that

$$\text{var}(\overline{Y}_{\text{obs}}) = \sigma_G^2 + \frac{\sigma_E^2}{r}, \tag{5.11}$$

less than $\text{var}(Y_{\text{obs}})$, due to the greater reliability of the mean of several observations, and the consequent smaller contribution of environmental variation. However, the variance of individual new observations, Y_{new}, is unchanged. Consequently, the ellipse is steeper, and so is the regression line B′C′: it is closer to the line OA. There is less regression towards the mean. Specifically, the regression line is given by

$$(Y_{\text{new}} - \mu) = (\overline{Y}_{\text{obs}} - \mu) \cdot \left(\frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_E^2}{r}} \right). \tag{5.12}$$

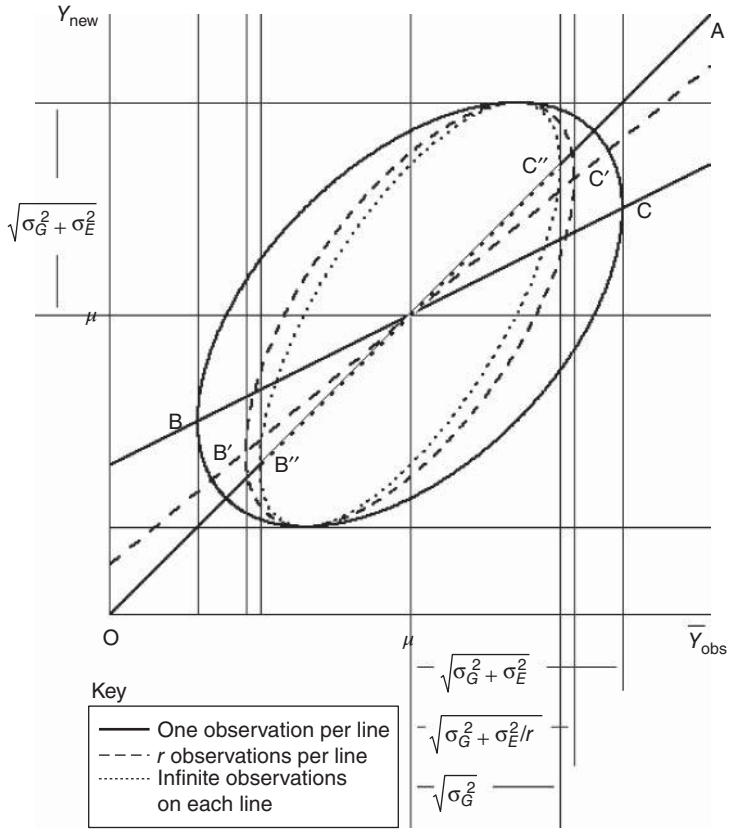

Figure 5.5 Summary of the relationship between present and future observations of yields of breeding lines of barley when r observations have been made on each line, showing the reduced discrepancy between BLUE and BLUP.

Comparison of this equation with Equation 5.6 shows that the amount of regression towards the mean is precisely equivalent to the shrinkage of the BLUP relative to the BLUE.

Finally, consider the situation when the prediction is based on a large – effectively infinite – number of observations, also illustrated in Figure 5.5. The effect of environmental variation on $\overline{Y}_{obs}$ is eliminated and var($\overline{Y}_{obs}$) is consequently reduced to σ_G^2, making the ellipse so steep that the regression line B″C″ is superimposed on the line OA. $\overline{Y}_{obs}$ now needs no adjustment to give the expected value of Y_{new}: there is no regression towards the mean, and no shrinkage of the BLUE is required to obtain the BLUP. If the plant breeder could attain this situation (which would require unlimited experimental resources), he/she would have no need of pessimism in his/her predictions: they would be based on full knowledge of the genetic potential of each line and would require no adjustment to indicate its future *mean* performance (though future individual observations would still vary as much as ever).

Except in this limiting case, the BLUE and the BLUP are different, so they cannot both be 'best' for the same purpose. For what purpose should each be preferred? The answer to this

question lies in a comparison of the following two equations:

$$E(Y_{\text{new}}|\mu_k) = \mu_k = E(\widehat{\mu} + \text{BLUE}_k) \tag{5.13}$$

$$E(Y_{\text{new}}|\overline{y}_k, \widehat{\mu}, \widehat{\sigma}^2_G, \widehat{\sigma}^2_E) = \mu + (\overline{y}_k - \mu) \cdot \left(\frac{\sigma^2_G}{\sigma^2_G + \frac{\sigma^2_E}{r_k}} \right) = E(\widehat{\mu} + \text{BLUP}_k). \tag{5.14}$$

Translated into words, these equations state the following:

- Equation 5.13. The expected value of a new observation on breeding line k, given that the true mean value for this breeding line is the unknown value μ_k, is also the expected value of the unadjusted mean of the sample of observations on this breeding line.
- Equation 5.14. The expected value of a new observation on breeding line k, given the unadjusted mean for this line, *together with the specified information about the population of lines to which it belongs* ($\widehat{\mu}, \widehat{\sigma}^2_G$ and $\widehat{\sigma}^2_E$), is also the expected value of the shrunk mean.

Thus if a factor is specified as fixed, each level sampled is considered to tell us only about itself, but if it is specified as random, each level is considered to provide some information concerning the other levels in the sample. Our best prediction about the future performance of Line k as a barley variety (or new chemical entity k as a medicine, or examination candidate k as a university student) therefore depends partly on our reason for taking an interest in it. If we have chosen Line k because of its relative performance among the set of factor levels under consideration – for example, because it is high-yielding relative to the other lines studied – then we should be guided by the BLUP. However, if we have chosen it for reasons that have nothing to do with its relative yield, for example, its brewing quality or its resistance to some disease – in short, if we have chosen it 'because it is Line k' – then we should ask ourselves whether the comparative performance of other lines is relevant, and if we decide that it is irrelevant, we should be guided by the BLUE. Unfortunately there is no simple, objective criterion for deciding whether the other levels of the factor under consideration are relevant. As we have seen (Section 1.6), it is not necessary that the levels comprise a representative sample from a population: it is sufficient that they form an exchangeable set, and more will be said about this situation below (Section 5.6). But how should we decide when this weaker criterion is met? 'There is nothing in (the mathematics) that requires the component problems (i.e. the factor levels) to have some sensible relation to each other' (Efron and Morris, 1977). This is *Stein's Paradox*, the BLUP being closely related to the *James–Stein estimator.* Efron and Morris give a fairly accessible account of this difficult issue, and Robinson (1991, Sections 5.3 and 6.1) robustly asserts that 'estimates of the characteristics of butterflies in Brazil, ball bearings in Birmingham and Brussels sprouts in Belgium ought not to be related to each other'.

However, sometimes common sense will suggest that factor levels should be regarded as exchangeable even though we have a special interest in one particular level. For example, suppose that we are interested in the performance of examination candidate k not because of his or her position in the distribution of students, but because we are his or her anxious parents. If Candidate k's performance is excellent on this occasion, we will of course hope for good results in the future. But we may prudently remember that there is an element of chance in these things, note his/her position in the overall distribution, resolve not to set our hopes too high, and specify 'candidate' as a random-effect term in our model. We can make

our assumption of exchangeability more robust by specifying that only levels of the same type as Level *k* should be considered – barley lines that share the brewing quality or disease resistance, chemical entities with a similar molecular structure or university students in the same socio-economic group. Or we may specify that all levels should be considered, but with an adjustment for such additional variables. This was our approach when we decided that the house prices in every town in the sample discussed in Chapter 1 were relevant to our prediction of house prices in Durham, but only after taking into account the latitude of each town, by including 'latitude' as a fixed-effect model term.

5.4 Use of R for the estimation of fixed and random effects

The following R commands import the original, 'unpadded' data on the barley breeding lines, fit the model omitting the block term and specifying 'line' as a fixed-effect term and print the results:

```
rm(list = ls())
barleyprogeny.unbalanced <- read.table(
   "IMM edn 2\\Ch 3\\barley progeny.txt",
   header=TRUE)
attach(barleyprogeny.unbalanced)
fline <- factor(line)
fblock <- factor(block)
barleyprogeny.model1lm <- lm(yield_g_m2 ~ fline)
summary(barleyprogeny.model1lm)
```

The output of the function `summary()` is as follows:

```
Call:
lm(formula = yield_g_m2 ~ fline)

Residuals:
    Min      1Q  Median      3Q     Max
-203.71  -43.41    0.00   43.41  203.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  654.485     81.442   8.036 4.75e-11 ***
fline2      -171.155    141.062  -1.213 0.229839
fline3       218.555    141.062   1.549 0.126645
fline4        64.655    141.062   0.458 0.648389
fline5       144.505    141.062   1.024 0.309825
fline6       147.870    115.177   1.284 0.204212
fline7       209.245    115.177   1.817 0.074339 .
fline8      -186.300    115.177  -1.618 0.111102
```

```
fline9         26.470    115.177   0.230 0.819026
fline10       106.270    115.177   0.923 0.359938
.
.
.
fline82      -403.185    141.062  -2.858 0.005878 **
fline83      -374.045    141.062  -2.652 0.010271 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 115.2 on 59 degrees of freedom
Multiple R-squared: 0.8733,     Adjusted R-squared: 0.6973
F-statistic: 4.961 on 82 and 59 DF,  p-value: 6.685e-10
```

The estimated mean for Line 1 is the intercept (654.485), and the effect of every other line is estimated relative to the value for Line 1: for example, for Line 7,

$$\text{fixed-effect mean} = 654.485 + 209.245 = 863.730.$$

The following commands fit the model omitting the block term and specifying 'line' as a random-effect term, and present the results:

```
library(nlme)
barleyprogeny.model1lme <- lme(yield_g_m2 ~ 1,
   data = barleyprogeny.unbalanced, random = ~ 1|fline)
summary(barleyprogeny.model1lme)
coef(barleyprogeny.model1lme)
```

The function `coef()` displays the estimates of the random effects in the model specified as its argument – in the present case, the shrunk means for the breeding lines. The coefficients for the fixed-effect terms – in this case, the intercept only – are included in the output of the function `summary()`.

The output of these commands is as follows:

```
Linear mixed-effects model fit by REML
 Data: barleyprogeny.unbalanced
       AIC     BIC   logLik
  1878.384 1887.23 -936.192

Random effects:
 Formula: ~1 | fline
        (Intercept) Residual
StdDev:    175.1198 115.0031
```

```
Fixed effects: yield_g_m2 ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 572.4734  21.67055 83 26.41711       0

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-1.62399201 -0.53822597 -0.02744873  0.46482593  2.05367544

Number of Observations: 142
Number of Groups: 83

   (Intercept)
1     639.9374
2     510.1906
3     782.4734
4     674.9465
5     730.7361
6     761.5776
7     812.0656
8     486.6841
9     661.7120
10    727.3568
.
.
.
82    348.0758
83    368.4353
```

The BLUP for each line can be obtained by rearranging Equation 5.8 to give

$$\text{BLUP}_k = \widehat{\mu}'_k - \widehat{\mu}. \tag{5.15}$$

For Line 7,

$$\text{BLUP}_7 = 812.0656 - 572.4734 = 239.5922.$$

This is almost, but not exactly, the same as the value given by GenStat (239.54 – Section 5.2).

5.5 Use of SAS for the estimation of random effects

The following SAS statements import the original, 'unpadded' data on the barley breeding lines, fit the model omitting the block term and specifying 'line' as a fixed-effect term and present the results:

```
PROC IMPORT OUT = barley DBMS = EXCELCS REPLACE
   DATAFILE = "&pathname.\IMM edn 2\Ch 3\barley progeny.xlsx";
   SHEET = "sheet1 SAS";
RUN;

ODS RTF;
```

```
PROC GLM;
   CLASS line;
   MODEL yield_g_m2 = line /SOLUTION;
RUN;
ODS RTF CLOSE;
```

Part of the output of PROC GLM is as follows:

Source	DF	Sum of squares	Mean square	*F* value	Pr > *F*
Model	82	5396990.540	65816.958	4.96	<0.0001
Error	59	782674.772	13265.674		
Corrected total	141	6179665.312			

R-square	Coeff Var	Root MSE	yield_g_m2 mean
0.873347	19.80432	115.1767	581.5735

Source	DF	Type I SS	Mean square	*F* value	Pr > *F*
line	82	5396990.540	65816.958	4.96	<0.0001

Source	DF	Type III SS	Mean square	*F* value	Pr > *F*
Line	82	5396990.540	65816.958	4.96	<0.0001

Parameter	Estimate		Standard error	*t* value	Pr > |*t*|
Intercept	280.4400000	B	115.1767082	2.43	0.0179
line 1	374.0450000	B	141.0620826	2.65	0.0103
line 2	202.8900000	B	162.8844627	1.25	0.2178
line 3	592.6000000	B	162.8844627	3.64	0.0006
line 4	438.7000000	B	162.8844627	2.69	0.0092
line 5	518.5500000	B	162.8844627	3.18	0.0023
line 6	521.9150000	B	141.0620826	3.70	0.0005
line 7	583.2900000	B	141.0620826	4.13	0.0001
line 8	187.7450000	B	141.0620826	1.33	0.1883
line 9	400.5150000	B	141.0620826	2.84	0.0062
line 10	480.3150000	B	141.0620826	3.40	0.0012
⋮					⋮
line 82	−29.1400000	B	162.8844627	−0.18	0.8586
line 83	0.0000000	B	.	.	.

Note: The X′X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The estimated mean for Line 83 is the intercept (280.44), and the effect of every other line is estimated relative to the value for Line 83: for example, for Line 7,

$$\text{fixed-effect mean} = 280.44 + 583.29 = 863.73.$$

The note reflects the fact that the choice of Line 83 as the factor level relative to which other effects are defined is arbitrary (cf. the arbitrary choice of two towns relative to which the effects of other towns are defined, described in Section 1.4). The following statements fit the model omitting the block term and specifying 'line' as a random-effect term, and present the results for Lines 1–4:

```
ODS RTF;
PROC MIXED ASYCOV NOBOUND;
   CLASS line;
   MODEL yield_g_m2 = /DDFM = KR HTYPE = 1 SOLUTION;
   RANDOM line /SOLUTION;
   ESTIMATE 'Shrunk mean, Line 1' INTERCEPT 1 | line 1;
   ESTIMATE 'Shrunk mean, Line 2' INTERCEPT 1 | line 0 1;
   ESTIMATE 'Shrunk mean, Line 3' INTERCEPT 1 | line 0 0 1;
   ESTIMATE 'Shrunk mean, Line 4' INTERCEPT 1 | line 0 0 0 1;
RUN;
ODS RTF CLOSE;
```

The `SOLUTION` option in the `RANDOM` statement indicates that the random-effect parameter estimate for each line – that is, its BLUP – is to be printed. The `ESTIMATE` statements obtain and print the shrunk means for each line, as a weighted sum of model parameters. Thus

$$\text{shrunk mean for Line 1} = 1 \times \text{intercept} + 1 \times \text{effect of Line 1},$$

$$\text{shrunk mean for Line 2} = 1 \times \text{intercept} + 0 \times \text{effect of Line 1} + 1 \times \text{effect of Line 2},$$

and so on. Similar, increasingly verbose statement could be used to obtain the shrunk means for the other lines. The shrunk means for all 83 lines can be obtained by replacing the `ESTIMATE` statements by the following:

```
%MACRO loop();
%DO i=1 %TO 83;
   ESTIMATE  "Shrunk mean, Line &i"
      INTERCEPT 1| line %DO j=1 %TO &i-1; 0 %END; 1;
%END;
%MEND loop;
OPTIONS MPRINT;
%loop;
```

However, a detailed explanation of the 'macro' syntax used in these statements is beyond the scope of the present account.

Part of the output of `PROC MIXED` is as follows:

Convergence criteria met.

Covariance parameter estimates	
Cov Parm	**Estimate**
Line	30655
Residual	13230

Asymptotic covariance matrix of estimates			
Row	**Cov Parm**	**CovP1**	**CovP2**
1	Line	38832845	−3532116
2	Residual	−3532116	5863037

Fit statistics	
−2 Res Log Likelihood	1872.4
AIC (smaller is better)	1876.4
AICC (smaller is better)	1876.5
BIC (smaller is better)	1881.2

AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion.

Null model likelihood ratio test		
DF	**Chi-square**	**Pr > ChiSq**
1	39.72	<0.0001

Solution for fixed effects					
Effect	**Estimate**	**Standard error**	**DF**	***t* value**	**Pr > |*t*|**
Intercept	572.48	21.6738	81.9	26.41	<0.0001

Solution for random effects						
Effect	**line**	**Estimate**	**Std Err Pred**	**DF**	***t* value**	**Pr > |*t*|**
line	1	67.4540	76.8380	106	0.88	0.3820
line	2	−62.2705	99.1142	127	−0.63	0.5310
Line	3	209.95	99.1142	127	2.12	0.0361
Line	4	102.45	99.1142	127	1.03	0.3033
Line	5	158.23	99.1142	127	1.60	0.1129
Line	6	189.08	76.8380	106	2.46	0.0155
Line	7	239.56	76.8380	106	3.12	0.0023
Line	8	−85.7798	76.8380	106	−1.12	0.2668

Solution for random effects						
Effect	**line**	**Estimate**	**Std Err Pred**	**DF**	***t* value**	**Pr > |*t*|**
Line	9	89.2259	76.8380	106	1.16	0.2482
Line	10	154.86	76.8380	106	2.02	0.0464
⋮						⋮
Line	82	−224.35	99.1142	127	−2.26	0.0253
Line	83	−204.00	99.1142	127	−2.06	0.0416

Estimates					
Label	**Estimate**	**Standard error**	**DF**	***t* value**	**Pr > |*t*|**
Shrunk mean, Line 1	639.93	74.8350	92.7	8.55	<0.0001
Shrunk mean, Line 2	510.20	98.1859	121	5.20	<0.0001
Shrunk mean, Line 3	782.43	98.1859	121	7.97	<0.0001
Shrunk mean, Line 4	674.92	98.1859	121	6.87	<0.0001
Shrunk mean, Line 5	730.70	98.1859	121	7.44	<0.0001
Shrunk mean, Line 6	761.55	74.8350	92.7	10.18	<0.0001
Shrunk mean, Line 7	812.04	74.8350	92.7	10.85	<0.0001
Shrunk mean, Line 8	486.70	74.8350	92.7	6.50	<0.0001
Shrunk mean, Line 9	661.70	74.8350	92.7	8.84	<0.0001
Shrunk mean, Line 10	727.34	74.8350	92.7	9.72	<0.0001
⋮					⋮
Shrunk mean, Line 82	348.12	98.1859	121	3.55	0.0006
Shrunk mean, Line 83	368.48	98.1859	121	3.75	0.0003

The column headed 'Estimate' in the table headed 'Solution for Random Effects' gives the BLUPs, and the corresponding column in the table headed 'Estimates' gives the shrunk means. These agree fairly closely with those produced by GenStat.

5.6 The Bayesian interpretation of BLUPs: Justification of a random-effect term without invoking an underlying infinite population

When the specification of a factor as a random-effect term is justified without invoking an underlying infinite population of levels, using the concept of exchangeability (Sections 1.6 and 2.6), an alternative justification can be given for the use of shrunk estimates of random effects. This is the *Bayesian* approach, in which the evidence about a parameter value (e.g. the mean of a factor level or the slope of a regression line) obtained from the data is combined with a prior belief concerning the value in order to obtain a posterior belief. We will examine this approach in general terms before considering its use to obtain BLUPs.

In Bayesian statistics, degrees of belief are represented by probability distributions, giving higher probability to values that are considered more credible. This interpretation of the concept of probability as a measure of belief is philosophically different from the *frequentist* interpretation, which considers the probability of an event to be the proportion of occasions on which that event occurs 'in the long run': a lucid account of the relationship between the two interpretations is given by Hacking (2001, Chapter 11, pp. 127–139, Chapter 16, pp. 189–200 and Chapter 21, pp. 256–260). The frequentist interpretation can be applied to estimates of a parameter, which will vary from one sample to the next, but cannot be applied to the true value, which is unknown but does not vary. However, the Bayesian interpretation can be applied to the true parameter value, because we have a belief about it – namely that it lies somewhere near our estimate. Fortunately the two concepts of probability obey the same laws, and can, with caution, be used together. In particular, the relationship between our belief concerning the true value of a parameter (in the present case, the mean yield of a particular breeding line) and an estimate of the same parameter (in the present case, the mean value from that breeding line in the field trial) is as follows.

We define Δ, the true value and D, the estimate of Δ from the data, both regarded as random variables. It can be shown that for any given value of Δ, say δ, and any given value of D, say d,

$$P(\Delta = \delta | D = d) = \frac{P(D = d | \Delta = \delta) \cdot P(\Delta = \delta)}{\sum_{\delta} P(D = d | \Delta = \delta) \cdot P(\Delta = \delta)} \tag{5.16}$$

where $P(\Delta = \delta | D = d)$ means 'The probability that Δ equals δ given that D equals d'. The symbol $\sum_{\delta}$ indicates summation over all possible values of δ: if δ is a continuous variable, this summation becomes an integration. This is Bayes' Theorem: its proof is straightforward, and is given, for example, by Hacking (2001, Chapter 7, pp. 69–71). It shows that in order to decide on our belief concerning the true value given the evidence (the *posterior probability distribution* of Δ, $P(\Delta = \delta | D = d)$), we must specify our belief before we see the evidence (the *prior distribution,* $P(\Delta = \delta)$).

The probability of our estimate given any particular true value, $P(D = d | \Delta = \delta)$, is given by the distribution

$$D | \delta, \sigma_D^2 \sim N(\delta, \sigma_D^2). \tag{5.17}$$

That is, if D were estimated repeatedly, the estimates would be clustered around the true value δ with a certain variance, σ_D^2. An estimate of this variance is provided by SE_d^2. It is convenient also to specify $P(\Delta = \delta)$ using a normally distributed variable,

$$\Delta | \mu_\Delta, \sigma_\Delta^2 \sim N(\mu_\Delta, \sigma_\Delta^2) \tag{5.18}$$

the values of μ_Δ and σ_Δ^2 being chosen to give a distribution that correctly describes our prior belief. We next define two *weights*, namely

$$w_{\text{prior}} = \frac{1}{\sigma_\Delta^2} \tag{5.19}$$

and

$$w_D = \frac{1}{\sigma_D^2}. \tag{5.20}$$

It can then be shown that our posterior belief concerning the true value, $P(\Delta = \delta | D = d)$, is expressed by the distribution

$$\Delta | \mu_\Delta, \sigma^2_\Delta, d, \sigma^2_D \sim N\left(\frac{w_{\text{prior}}\mu_\Delta + w_D d}{w_{\text{prior}} + w_D}, \frac{1}{w_{\text{prior}} + w_D} \right). \tag{5.21}$$

That is:

- the mean of the posterior distribution is a weighted mean of the prior mean and the estimate from the data and
- the variance of the posterior distribution is inversely related to the weights, and hence directly related to the variances of the component variables.

The mean of the posterior distribution is always intermediate between those of the component variables, and its variance is always smaller than that of either component variable: that is, precision is always gained by combining the two sources of information.

It can be shown that if all values of Δ from $-\infty$ to $+\infty$ were considered equally probable before inspecting the data, then on the basis of the data, Δ would have the distribution

$$\Delta | d, \sigma^2_D \sim N(d, \sigma^2_D) \tag{5.22}$$

Thus the posterior distribution (Distribution 5.21) is equivalent to the distribution of a new variable,

$$\Delta | \mu_\Delta, \sigma^2_\Delta, d, \sigma^2_D = \frac{w_{\text{prior}}(\Delta | \mu_\Delta, \sigma^2_\Delta) + w_D(\Delta | d, \sigma^2_D)}{w_{\text{prior}} + w_D}, \tag{5.23}$$

that is, a weighted mean of $\Delta | \mu_\Delta, \sigma^2_\Delta$ (which has Distribution 5.18) and $\Delta | d, \sigma^2_D$ (which has Distribution 5.22), the weights being inversely proportional to the variances of the component variables. Thus the Bayesian approach can be informally summarized as

$$\text{posterior belief} = \text{prior belief} + \text{belief on the basis of the data alone.}$$

In the Bayesian approach to statistical analysis, prior belief may come from expert opinion. For example, an expert in barley breeding may believe that the true mean yield of lines derived from the cross 'Chebec' × 'Harrington' is usually between 352.4 and 703.6 g/m^2, that is, 572.5 ± 175.1 g/m^2 (implausibly precise values: the reason for choosing them will become apparent shortly). This belief can be expressed by the statements

$$\mu_\Delta = 572,$$
$$\sigma^2_\Delta = 175.1^2,$$

and hence

$$\Delta | \mu_\Delta, \sigma^2_\Delta \sim N(572.5, 175.1^2)$$

(from which the units of measurement, g/m^2, have been dropped for convenience). The expert will therefore be sceptical about the high mean of Line 7, $d = 863.7$ g/m^2. The SE of this estimate is

$$\sqrt{\frac{\widehat{\sigma}^2_E}{r_k}} = \sqrt{\frac{13226}{2}} = 81.3 \text{ g/m}^2,$$

so approximately,

$$\sigma_D^2 = 81.3^2.$$

The expert's posterior belief concerning the true mean yield of this breeding line is obtained by combining his/her prior belief and the evidence, using the Equations and Distribution 5.19–5.21:

$$w_{\text{prior}} = \frac{1}{175.1^2} = 0.00003261$$

$$w_D = \frac{1}{81.3^2} = 0.00015122$$

$$\Delta|\mu_\Delta, \sigma_\Delta^2, d, \sigma_D^2 \sim$$

$$N\left(\frac{0.00003261 \times 572.5 + 0.00015122 \times 863.7}{0.00003261 + 0.00015122}, \frac{1}{0.00003261 + 0.00015122}\right)$$

$$\Delta|\mu_\Delta, \sigma_\Delta^2, d, \sigma_D^2 \sim N(812.04, 73.76^2).$$

But in the context of mixed modelling, we are interested in a prior distribution derived not from expert opinion but from a set of estimates of which the estimate under consideration is a member: in the present case, the estimated mean yields of the sample of breeding lines. The mean of all these estimates, and the variance of the estimates around this value, then provide the parameters of the prior distribution. This is sometimes known as the *empirical Bayesian approach* (see also Section 8.4). The first parameter of Distribution 5.21 can be rearranged as

$$\text{posterior mean}(\Delta|\mu_\Delta, \sigma_\Delta^2, d, \sigma_d^2) = \mu_\Delta + (d - \mu_\Delta) \cdot \frac{\sigma_\Delta^2}{\sigma_\Delta^2 + \sigma_d^2}, \tag{5.24}$$

and this expression bears a strong resemblance to the formula for the shrunk mean given in Equations 5.6 and 5.8: indeed, when the prior distribution is obtained in this way, we can re-write Equation 5.24 as

$$\text{posterior mean}(\Delta|\mu_\Delta, \sigma_\Delta^2, d, \sigma_d^2) = \mu_\Delta + \text{BLUP}(\Delta). \tag{5.25}$$

where

$$\text{BLUP}(\Delta) = (d - \mu_\Delta) \cdot \frac{\sigma_\Delta^2}{\sigma_\Delta^2 + \sigma_d^2} \tag{5.26}$$

In the present case, we substitute

$$\mu_\Delta = \hat{\mu} = 527.5$$

$$\sigma_\Delta^2 = \hat{\sigma}_G^2 = 30667 = 175.1^2$$

$$d = \hat{\mu}_k = 863.7$$

$$\sigma_D^2 = \frac{\hat{\sigma}_E^2}{r_k} = \frac{13226}{2}$$

and obtain

$$\text{posterior mean}(\Delta|\mu_\Delta, \sigma_\Delta^2, d, \sigma_d^2) = 527.5 + (863.7 - 527.5) \cdot \frac{30667}{30667 + \frac{13226}{2}} = 812.04,$$

which is the same as the value obtained from the prior distribution based on expert opinion, as the values specified for μ_Δ and σ^2_Δ are the same. The SE of this shrunk mean is given by

$$\sqrt{\text{posterior var}(\Delta|\mu_\Delta, \sigma^2_\Delta, d, \sigma^2_d)} = \sqrt{\frac{1}{w_{\text{prior}} + w_D}} = \sqrt{\frac{1}{0.00003261 + 0.00015122}} = 73.76.$$

In this Bayesian approach to BLUPs, the collective features of the exchangeable set of factor levels (in this case μ_Δ and σ^2_Δ) are regarded as prior knowledge, to be taken into account when interpreting the data on each individual level (in this case the values of D). If we decide that the factor levels do not form an exchangeable set, but are so disparate that each is irrelevant to the estimation of the effect of the others (see the discussion of Stein's paradox in Section 5.3), we can express this irrelevance by specifying $\sigma^2_\Delta = \infty$ instead of using the variance component for the factor under consideration. The prior distribution is then uniform from $-\infty$ to $+\infty$, and the posterior mean is the unshrunk mean for each factor level: that is, from a Bayesian point of view, a fixed-effect term is simply a random-effect term with a flat prior distribution.

It should be noted that the BLUP and its SE are only strictly valid when d and the estimate of σ^2_Δ are independent, whereas in practice they are almost always obtained from the same information. This does not matter much when the number of factor levels is large, as in the present case: the contribution of each breeding line to $\widehat{\sigma}^2_G$ is small, and hence the correlation (lack of independence) between $\widehat{\mu}_k$ and $\widehat{\sigma}^2_G$ is negligible. But when the number of factor levels is small, as in the sample of English towns considered in Chapter 1, this correlation starts to matter, and the BLUPs for individual towns, and their SEs, should not be taken too literally. In this context, it is helpful to distinguish two purposes for which a model term may be specified as random:

- to obtain BLUPs for the effects of the term itself or
- to ensure that it is taken into account when determining the precision of effect estimates in other terms.

For example, in the present case, our primary interest is in the yields of the individual breeding lines, and we will still be interested in shrinking our estimates of these as a defence against over-optimism, whether or not we feel justified in regarding them as a representative sample from a much larger population. For this purpose, exchangeability is enough, though we require a moderately large set of factor levels in order to obtain trustworthy BLUPs. In connection with the James–Stein estimator, Efron and Morris (1977) advise that as few as nine levels are tolerable and 15 or 20 are ample, and it seems reasonable to extend these judgements to BLUPs. However, in the study of the relationship between latitude and house prices, our main motive for specifying 'town' as a random-effect term was to ensure that this term contributed to the SE for the estimate of the effect of latitude, and for this purpose it is helpful to envisage a much larger population of towns for which this estimate will be valid, though a small sample of towns is enough to give us a usable SE. We took relatively little interest in the mean prices in individual towns, whether represented by BLUEs or by BLUPs. In the case of a randomized experimental design, the contribution of block and residual effects to the SE of a treatment effect is justified by the large number of possible permutations of these effects over the treatments, from which the actual permutation was randomly selected (see Sections 2.6 and 4.1). But it is probably not helpful to envisage the random permutation of English towns over their latitudes.

5.7 Summary

In ordinary regression analysis and analysis of variance, random-effect terms are always regarded as nuisance variables – residual variation or block effects – but this is not always appropriate. Effects that are of intrinsic interest may be specified as random.

The decision to specify a term as random causes a fundamental change in the way in which the effect of each of its levels is estimated.

This is illustrated in the field experiment to evaluate breeding lines of barley (Sections 3.8–3.17), but the same concepts and arguments apply equally to any situation in which replicated, quantitative evaluations are available for the comparison of members of some population, for example:

- new chemical entities to be evaluated as potential medicines
- candidates for admission to a university on the basis of their examination scores.

We saw earlier (Sections 3.15 and 3.20) that the predicted genetic advance due to selection among the barley breeding lines is less than the value given by a naïve prediction from the mean of the selected lines. Mixed-model analysis gives a similar adjustment to the mean of each individual breeding line.

When a term is specified as random, the ordinary difference between the mean for each level and the grand mean (the BLUE) is replaced by a 'shrunk estimate' (the BLUP).

The shrinkage of the BLUP, relative to the BLUE, is large when:

- the variance component for the term in question is small
- the residual variance is large
- the number of replications of the factor level under consideration is small.

Crossovers may occur as a result of the shrinkage, so that a level of a random-effect term that is ranked higher than another on the basis of the BLUEs is ranked lower on the basis of the BLUPs.

The shrinkage of the BLUP is equivalent to the phenomenon of regression towards the mean.

This is illustrated by considering the present observations of each level of a random-effect term as predictions of future observations. The line of best fit relating present observations to future observations (the regression line) is flatter than a line of unit slope passing through the origin.

The degree of flattening of the line of best fit (the amount of regression towards the mean) is equivalent to the shrinkage of the BLUP relative to the BLUE.

The discrepancy between the BLUE and the BLUP tells us that our best prediction about the future performance of any factor level of a random-effect term depends partly on our reason for taking an interest in it, as follows:

- if we have chosen the level in question because of its relative performance among the levels under consideration, we should be guided by the BLUP

- however, if we have chosen it for reasons unrelated to its relative performance, and consider the comparative performance of other levels to be irrelevant, we should be guided by the BLUE.

When the levels of a random-effect term are regarded as an exchangeable set rather than a sample from an infinite population, the BLUP can be justified by a *Bayesian* argument. This can be informally summarized as

$$\text{posterior belief} = \text{prior belief} + \text{belief on the basis of data alone.}$$

In the '*empirical Bayesian*' approach, the prior probability distribution of belief is specified by the grand mean and the variance component for the term in question. The estimated mean of the level under consideration (e.g. the mean for a particular barley breeding line) and its variance are combined with this prior distribution to obtain the posterior distribution. The mean of this posterior distribution is the shrunk mean for the level in question (grand mean + BLUP). The square root of the variance of the posterior distribution gives the SE of this shrunk mean.

It is helpful to distinguish two purposes for which a model term may be specified as random:

- to obtain BLUPs for the effects of the term itself. For this purpose, the criterion of exchangeability is sufficient.
- to ensure that the term is taken into account when determining the precision of effect estimates in other terms. For this purpose, it may be helpful to envisage a much larger population from which the levels of the term are sampled. In the case of a treatment effect in a randomized experimental design, the contribution of block and residual effects to the SE is justified by the large number of possible permutations of the treatments over these effects.

In order to achieve trustworthy BLUPs, a fairly large set of levels (at least 9 and ideally 15 or more) must be studied, so that the estimated variance component for the term in question and the estimated mean of each level are nearly independent.

5.8 Exercises

5.1 Return to the data set concerning the yield of F_3 wheat families in the presence of ryegrass, introduced in Exercise 3.3.

(a) Using mixed modelling, obtain an estimate of the mean yield of each of the F_3 families, specifying 'family' as a fixed-effect term.

(b) Using mixed modelling and specifying 'family' as a random-effect term, obtain the following:

(i) an estimate of the overall mean of the population of F_3 families

(ii) the BLUP for the effect of each family.

(c) From the output of the analysis performed in Section (b), obtain the following:

(i) an estimate of the variance component for 'family'

(ii) an estimate of the residual variance component.

Obtain also the number of observations of each family.

(d) From the information obtained in Parts (a)–(c), compare the relationship between the BLUPs and the estimates of family means obtained specifying 'family' as a fixed-effect term with that given in Equation 5.6.

(e) Obtain the shrunk mean for each family, and plot the shrunk means against the means obtained when 'family' is specified as a fixed-effect term (the unadjusted means).

The point representing one of the families deviates from the general relationship between these two types of mean.

(f) What is the distinguishing feature of this family?

5.2 In many types of plant, exposure to low temperature at an early stage of development causes flowering to occur more rapidly: this phenomenon is called *vernalization*. An inbred line of chickpea with a strong vernalization response and a line with little or no vernalization response were crossed, and the F_1 hybrid progeny were self-fertilized to produce the F_2 generation. Each F_2 plant was self-fertilized to produce an F_3 family. The seed of each F_3 family was divided into two batches. Germinating seeds of one were vernalized by exposure to low temperature (4 °C) for 4 weeks. The other batch provided a control. All F_3 seeds were then sown, and allowed to grow. The plants were arranged in groups of four: within each group the plants were of the same family and had received the same low-temperature treatment. Generally there were 12 plants (i.e. three groups of four) in each family and each exposed to low-temperature treatment, but in some families fewer or more plants were available. The number of days from sowing to flowering was recorded for each plant. The first and last few rows of the spreadsheet holding the data are presented in Table 5.2: the complete data set is held in the file 'chickpea vernalisation.xlsx', available from this book's website, at the web address given in the Preface. (Data reproduced by the kind permission of S. Abbo, Field Crops and Genetics, The Hebrew University of Jerusalem.)

(a) Divide the data between two spreadsheets, one holding only the results from the vernalized plants, the other, only those from the control plants.

(b) Analyse the results from the vernalized plants by mixed modelling, specifying 'family', 'group' and 'plant' as random-effect terms.

(c) Obtain an estimate of the component of variance for each of the following terms:

(i) family

(ii) group within family

(iii) plant within group.

Which term in your mixed model represents residual variation?

(d) Estimate the heritability of time to flowering in vernalized plants from this population of families. (N.B. The estimate obtained using the methods described in Chapter 3 is slightly biased downwards, as some of the residual variance is due to genetic differences among plants of the same family.)

(e) Obtain the unadjusted mean and the shrunk mean, and the BLUP, for the number of days from sowing to flowering in each family.

(f) Extend Equation 5.6 to the present situation, in which two components of variance contribute to the shrinkage of the BLUPs. Use the values obtained above to check your equation.

(g) Repeat the steps indicated in Sections (b), (c) and (d) of this exercise for the control plants. Comment on the difference between the estimates of variance components and heritability obtained from the vernalized plants and the control plants.

Table 5.2 Time from sowing to flowering of F_3 chickpea plants with and without exposure to a vernalizing stimulus.

	A	B	C	D	E
1	plant_group	plant	family	low_T	days_to_flower
2	60	1	88	vernalized	63
3	60	2	88	vernalized	64
4	60	3	88	vernalized	61
5	60	4	88	vernalized	70
6	1	1	14	vernalized	75
7	1	2	14	vernalized	
8	1	3	14	vernalized	
9	1	4	14	vernalized	
10	103	1	29	control	78
11	103	2	29	control	82
12	103	3	29	control	82
13	103	4	29	control	88
⋮					⋮
1092	41	3	44	vernalized	70
1093	41	4	44	vernalized	64

Source: Data reproduced by kind permission of S. Abbo, Field Crops and Genetics, The Hebrew University of Jerusalem.

(h) Plot the shrunk mean for each family obtained from the control plants against the corresponding value obtained from the vernalized plants. Comment on the relationship between the two sets of means.

5.3 Return to the house price data analysed in Chapter 1.

(a) Fit the mixed model introduced in Chapter 1 to these data. Obtain the intercept and slope of the line of best fit relating log(house price) to latitude. Obtain the BLUP for the effect of each town. Do you consider this data set to be adequate to permit interpretation of the BLUPs?

(b) Obtain the fitted value of log(house price) for each town, that is, the value on the line of best fit at the latitude of each town.

(c) Hence obtain the shrunk mean value of log(house price) for each town.

(d) Obtain the BLUE for each town.

(e) Produce a figure similar to Figure 1.4, but with the line of best fit from the mixed model instead of the simple regression line. Add the shrunk means to this plot. Comment on their distribution relative to the simple means.

(f) Plot the shrunk means against the simple means. Identify any crossovers among the towns for these variables.

(g) Plot the shrunk means against the simple means. Again, identify any crossovers.

References

Efron, B. and Morris, C. (1977) Stein's paradox in statistics. *Scientific American*, **236**, 119–127.

Hacking, I. (2001) *An Introduction to Probability and Inductive Logic*, Cambridge University Press, Cambridge, 302 pp.

Robinson, G.K. (1991) That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–51.