Dr Jing Xu

Statistical Modelling

4 Randomized Blocks Designs and Two-Way ANOVA

4.1 Randomized blocks designs and principles of experimental design

Randomized blocks designs originated in agricultural experimentation, but they may be used in many other applications. In an agricultural experiment where several treatments on a certain crop are to be compared, the plots of land to which the treatments are to be applied may be laid out in blocks, where, for example, different blocks might have different soil types or different levels of exposure to sunlight or wind, or different levels of drainage. Different blocks will in general have different levels of fertility, but the aim is to make plots within blocks as uniform as possible. In the simplest examples of randomized blocks designs, the number of plots in each block is made equal to the number of treatments and the treatments can then be applied in a random arrangement to the plots in each block.

Randomized Blocks Design

Blocks

I	II	III
1	4	3
3	3	1
4	2	2
2	1	4

(Treatments 1, 2, 3, 4)

Some principles of experimental design:

Replication – the taking of repeated observations (here growing the crop on several plots) for each treatment. This enables (i) an assessment of the size of the experimental errors to be made and, in particular, an estimate of their variance to be obtained and (ii) more accurate estimates of the treatment effects to be found.

Randomization – the random allocation of the experimental units (here the plots) to the treatments in order to avoid biasing the results through any extraneous factors which might be present.

Blocking – a partition of the experimental units into groups, known as blocks, such that units in the same block are as similar as possible. Consequently, variation between units in the same block should be small in comparison with variation between units in different blocks.

The concept of blocking is very important in experimental design. Blocking is a restriction on randomization designed to increase the precision of an experiment. It is a controlled allocation of the treatments to the experimental units in a balanced way with the intention of removing the possibility that any treatment should have a monopoly of the extreme units. Blocking should have the effect of reducing the error variance. In the analysis of the data, the variation due to differences among blocks will be eliminated to reduce the residual sum of squares.

Another example of blocking, which arises in a natural way, is where the effects of different diets or different drugs are being compared using experimental animals, for example, rats. The experimental units are the animals which are being fed the diets or drugs, and blocks correspond to sets of siblings, that is, animals from the same litter. We would expect animals from the same litter to be more similar in their response to any given treatment than animals chosen at random.

In the basic *randomized blocks design* for comparing a treatments, experimental material is divided into b blocks each containing a experimental units, and the a treatments applied in a random arrangement to the a units in each block. Thus the total number of experimental units is $n \equiv ab$.

4.2 The statistical model for a randomized blocks design

Let y_{ij} represent the observation for Treatment i in Block j $(1 \leq i \leq a, 1 \leq j \leq b)$. Each observation has two factors associated with it, treatment and block. We set up the linear statistical model

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \qquad (1 \leq i \leq a, 1 \leq j \leq b). \tag{1}$$

The parameter μ is an overall mean, the parameters τ_i $(1 \leq i \leq a)$ are the *treatment effects*, the parameters β_j $(1 \leq j \leq b)$ are the *block effects* and the ϵ_{ij} are random errors, the parameters being unknown. It will be assumed that the ϵ_{ij} are independently and identically distributed $\mathrm{N}(0, \sigma^2)$, where the error variance σ^2 is unknown.

The model (1) is the statistical model for a randomized blocks design, the data from which will be analysed using a *two-way analysis of variance (ANOVA)*.

- In setting up the model of equation (1) we assume that the treatment and block effects are *additive*, i.e., that there is no *interaction* between treatment and block effects.
- More generally, model (1) is the model for an experiment with two factors which are additive in their effects and with just one observation for each combination of factor levels.

The model (1) may be written in the form of the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} = (\mu, \tau_1, \tau_2, \ldots, \tau_a, \beta_1, \beta_2, \ldots, \beta_b)'$ is an $(a + b + 1) \times 1$ vector of parameters and $\mathbf{X}$ is an $n \times (a + b + 1)$ design matrix.

For illustration we take the special case shown in Section 4.1, where $a = 4$ and $b = 3$, so that $n = 12$. We take

$$\mathbf{y} = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}, y_{31}, y_{32}, y_{33}, y_{41}, y_{42}, y_{43})'$$

The columns of the 12×8 design matrix $\mathbf{X}$ shown below are the coefficients for the parameters μ, $\tau_1, \tau_2, \tau_3, \tau_4$, $\beta_1, \beta_2, \beta_3$, respectively:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

A MLE, equivalent to a least squares estimate, $\hat{\boldsymbol{\beta}}$ of the vector of parameters is given by a solution of the normal equations

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}, \tag{2}$$

where

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 3 & 3 & 3 & 3 & 4 & 4 & 4 \\ 3 & 3 & 0 & 0 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 0 & 0 & 1 & 1 & 1 \\ 3 & 0 & 0 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 0 & 0 & 3 & 1 & 1 & 1 \\ 4 & 1 & 1 & 1 & 1 & 4 & 0 & 0 \\ 4 & 1 & 1 & 1 & 1 & 0 & 4 & 0 \\ 4 & 1 & 1 & 1 & 1 & 0 & 0 & 4 \end{pmatrix}$$

The 8×8 matrix $\mathbf{X}'\mathbf{X}$ is not invertible, since it is not of full rank 8. In fact it is of rank 6: for both $\mathbf{X}$ and $\mathbf{X}'\mathbf{X}$, adding columns 2, 3, 4 and 5 gives column 1, as does adding columns 6, 7 and 8. So the normal equations (2) do not have a unique solution.

This conclusion holds more generally for the model (1), in which there are two redundant parameters. To obtain a unique MLE we may impose two constraints, one on the treatment effects τ_i and one on the block effects β_j. The traditional constraints that give the neatest results algebraically are

$$\sum_{i=1}^{a} \tau_i = 0 \tag{3}$$

and

$$\sum_{j=1}^{b} \beta_j = 0. \tag{4}$$

Algebraic solution of the normal equations shows that, using the standard dot notation, the least squares estimates of μ and the τ_i are then given by

$$\hat{\mu} = \bar{y}_{..}$$

and

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..} \qquad (1 \leq i \leq a).$$

The block effects β_j are often regarded as "nuisance parameters", which we are not particularly interested in estimating. However, should we wish to estimate them,

$$\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..} \qquad (1 \leq j \leq b).$$

The basic null hypothesis, H_0, to be tested is that the treatment effects τ_i are all equal, i.e., with the constraint (3),

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0.$$

The alternative hypothesis is

$$H_1 : \tau_i \neq 0 \text{ for at least one } i.$$

4.3 The partition of the total corrected sum of squares

The total (corrected) sum of squares SS_T for the experiment is defined by

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2.$$

The total sum of squares may be partitioned as

$$SS_T = SS_{Treatments} + SS_{Blocks} + SS_{Residual}, \tag{5}$$

where

$$SS_{Treatments} = b \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 = b \sum_{i=1}^{a} \hat{\tau}_i^2,$$

$$SS_{Blocks} = a \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 = a \sum_{j=1}^{b} \hat{\beta}_j^2$$

and

$$SS_{Residual} = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2.$$

The sums of squares on the right hand side of equation (5) are independently distributed. Irrespective of the values of the treatment and block effects,

$$\frac{SS_{Residual}}{\sigma^2} \sim \chi^2_{(a-1)(b-1)} \ .$$

Under H_0,

$$\frac{SS_{Treatments}}{\sigma^2} \sim \chi^2_{a-1} \ ,$$

and if all the block effects are equal then

$$\frac{SS_{Blocks}}{\sigma^2} \sim \chi^2_{b-1} \ .$$

The partition of the total sum of squares and the associated degrees of freedom together with the subsequent analysis may be laid out in an ANOVA table.

ANOVA TABLE

Source	DF	SS	MS
Treatments	$a-1$	$b\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2$	$\frac{SS_{Treatments}}{a-1}$
Blocks	$b-1$	$a\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2$	$\frac{SS_{Blocks}}{b-1}$
Residual	$(a-1)(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2$	$s^2 \equiv \frac{SS_{Residual}}{(a-1)(b-1)}$
Total	$n-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2$	

Whether or not H_0 is true, an unbiased estimator of σ^2 is given by $s^2 \equiv MS_{Residual}$. If H_0 is true then the F-statistic,

$$F = \frac{MS_{Treatments}}{MS_{Residual}},$$

has the $F_{a-1,(a-1)(b-1)}$ distribution. This statistic is used for a one-tail test of H_0.

The sum of squares for blocks is that part of the total sum of squares accounted for by differences among blocks. If the blocking is effective, the error mean square is substantially reduced from what it would have been if blocking had not been used in the experiment or differences among blocks had not been taken into account in the analysis. Thus the blocking increases the sensitivity of the F-test for differences among treatment effects. The importance of the block effects may be assessed using the ratio

$$F = \frac{MS_{Blocks}}{MS_{Residual}}$$

and the $F_{b-1,(a-1)(b-1)}$ distribution.

4.4 An example from manufacturing

The table below gives the number of units produced in a day by 4 different machines, A, B, C and D, on each of 5 different days. The days may be regarded as a nuisance factor. We wish to compare the production levels of the machines and consider the days as blocks, although, since no randomization within blocks has taken place, we do not strictly speaking have a randomized blocks design.

Units produced

Day	Machine			
	A	B	C	D
1	293	308	323	333
2	298	353	343	363
3	280	323	350	368
4	288	358	365	345
5	260	343	340	330

The data are analyzed in the following R output.

```
> manu <- read.table("manu.txt")
> names(manu) <- list("Machine","Day","Output")
> manu
   Machine Day Output
1        A   1    293
2        A   2    298
3        A   3    280
4        A   4    288
5        A   5    260
6        B   1    308
7        B   2    353
8        B   3    323
9        B   4    358
10       B   5    343
11       C   1    323
12       C   2    343
13       C   3    350
14       C   4    365
15       C   5    340
16       D   1    333
17       D   2    363
18       D   3    368
19       D   4    345
20       D   5    330
> attach(manu)
> Day <- factor(Day)
> manu.aov <- aov(Output ~ Machine + Day)
> summary(manu.aov)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Machine      3 13444.8  4481.6 20.4780 5.178e-05 ***
Day          4  2146.2   536.6  2.4517    0.1027
Residuals   12  2626.2   218.8
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> tapply(Output,Machine,mean)
    A     B     C     D
283.8 337.0 344.2 347.8
> tapply(Output,Day,mean)
     1      2      3      4      5
314.25 339.25 330.25 339.00 318.25
> coef(manu.aov)
(Intercept)    MachineB    MachineC    MachineD        Day2        Day3
     269.85       53.20       60.40       64.00       25.00       16.00
       Day4        Day5
      24.75        4.00
```

Note first the layout of the data in the dataframe `manu`. The first variable, `Machine`, for each of the 20 observations specifies the machine label, the second variable, `Day`, specifies the day, and the third variable, `Output`, gives the output for the given machine on the given day.

The `aov` function carries out the analysis of variance, whose results in the present case are stored in the object `manu.aov`. The model formula

```
Output ~ Machine + Day
```

specifies that the `Output` variable is explained by the two factors `Machine` and `Day`, which are additive in their effects. In other words, this R formula specifies a linear model of the form set out in detail in equation (1).

The `summary` function, applied to the object `manu.aov`, produces the ANOVA table, which shows that there are highly significant differences among the estimated machine effects, that is, among the observed average outputs from each machine ($p < 0.0001$). The differences among the machine outputs on the different days are not very significant ($p > 0.1$). Nevertheless, the inclusion of `Day` as a factor has reduced the residual mean square and improved the sensitivity of the tests for differences among machines.

Although at this point we do not carry out a formal test of individual differences among the machines and among the days, it is important to note the nature of the differences.

- The `tapply` function is used here to apply the function `mean` to the vector `Output` separately for each level of a factor, in the first case the factor `Machine` and in the second case the factor `Day`, so that the mean output is calculated for each machine and for each day.

We may note that Machine A has a much lower average output than the others. As we shall see, this difference does turn out to be highly significant. Even though overall the differences among the days are not significant, we may note that the mean outputs are somewhat smaller on Day 1 and on Day 5 than on the other days. If the days are the working days of a given week in their natural order then the output from the machines is somewhat lower on Monday and Friday than on Tuesday, Wednesday and Thursday.

The function `coef` when applied to the object `manu.aov` lists the estimated parameter values.

- In R, instead of the constraints of equations (3) and (4), the constraints $\tau_1 = 0$ and $\beta_1 = 0$ are used, so that the first level of each factor is used as the basis for comparisons. This alters the values of the parameter estimates, so that

$$\begin{aligned} \hat{\tau}_i &= \bar{y}_{i\cdot} - \bar{y}_{1\cdot} \qquad (1 \le i \le a), \\ \hat{\beta}_j &= \bar{y}_{\cdot j} - \bar{y}_{\cdot 1} \qquad (1 \le j \le b), \\ \hat{\mu} &= \bar{y}_{1\cdot} + \bar{y}_{\cdot 1} - \bar{y}_{\cdot\cdot} \end{aligned}$$

  However, the ANOVA table and other aspects of the analysis are unchanged.

In the present case, the machine effect is zero for Machine A, the day effect is zero for Day 1. The other machine effects are measured relative to Machine A and the day effects relative to Day 1. The `Intercept` is the value of $\hat{\mu}$.

The R function `lm` may be used more generally for fitting the various forms of the general linear model, such as multiple linear regression, but `aov` is designed to provide a more useful summary output for the case of what are known as *balanced* designs, where there is an equal number of observations for each combination of factor levels. (In the present case there is exactly one observation for each combination of machine and day.) For comparison, we show below some output generated by using the function `lm`. The functions `lm` and `aov` use exactly the same method for fitting the linear model, but the results are presented in different ways. The `anova` function, when applied to the object

`manu.lm`, gives exactly the same ANOVA table as we saw in the summary output for the `aov` function.

In the `lm` summary output, instead of the ANOVA table, we have a list of the estimated parameter values together with their standard errors, t-values and p-values. As for the `aov` function, the constraints adopted are that the parameter values for Machine A and Day 1 are zero. Hence the estimated parameter values for Machines and Days are the estimated effects relative to Machine A and Day 1, respectively. In particular, we may observe that the highly significant and positive estimated effects of Machines B, C and D imply that each of them has a significantly higher output than Machine A.

```
> manu.lm <- lm(Output ~ Machine + Day)
> summary(manu.lm)

Call:
lm(formula = Output ~ Machine + Day)

Residuals:
    Min      1Q  Median      3Q     Max
-16.050  -8.950   1.150   6.812  23.150

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  269.850      9.356  28.842 1.88e-12 ***
MachineB      53.200      9.356   5.686 0.000101 ***
MachineC      60.400      9.356   6.456 3.13e-05 ***
MachineD      64.000      9.356   6.840 1.80e-05 ***
Day2          25.000     10.461   2.390 0.034141 *
Day3          16.000     10.461   1.530 0.152054
Day4          24.750     10.461   2.366 0.035659 *
Day5           4.000     10.461   0.382 0.708862
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.79 on 12 degrees of freedom
Multiple R-squared: 0.8558, Adjusted R-squared: 0.7717
F-statistic: 10.18 on 7 and 12 DF,  p-value: 0.000322

> anova(manu.lm)
Analysis of Variance Table

Response: Output
          Df  Sum Sq Mean Sq F value    Pr(>F)
Machine    3 13444.8  4481.6 20.4780 5.178e-05 ***
Day        4  2146.2   536.6  2.4517    0.1027
Residuals 12  2626.2   218.8
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Looking ahead, in the main body of the course the `glm` function will be used for fitting *generalized linear models*.