STATISTICAL MODELLING

III. Completely Randomized Design

(ref. Myers and Milton, sec. 5.2 and 6.2; Mead, sec. 6.2)

III.A Design of a CRD III-1
III.B Models and estimation for a CRD III-2
a) Maximal model III-2
b) Alternative indicator-variable, expectation models III-6
III.C Hypothesis testing using the ANOVA method III-8
a) Analysis of the rat example III-8
b) Sums of squares for the analysis of variance III-8
c) Expected mean squares III-12
d) Summary of the hypothesis test III-15
e) Comparison with traditional one-way ANOVA III-16
f) Computation of the ANOVA in R III-16
III.D Diagnostic checking III-18
III.E Treatment differences III-25
a) Multiple comparisons procedures for comparing all treatments III-25
b) Fitting submodels III-27
c) Comparison of treatment parametrizations III-34
III.F Summary III-35
III.G Exercises III-36

III.A Design of a CRD

Definition III.1: An experiment is set up using a Completely Randomized Design (CRD) when each treatment is applied a specified, possibly unequal, number of times, the particular units to receive a treatment being selected completely at random. ■

Generally we will use R to obtain randomized layouts and how to do this is described in Appendix B, *Randomized layouts and sample size computations in R*, for all the designs that will be covered in this course, and more besides.

Example III.1 Rat experiment

For example, consider an experiment in which the effects of three diets on rats is to be investigated. Suppose I have 6 rats to be fed one of three diets, 3 rats to be fed diet A, 2 diet B and 1 diet C.

The following output shows the use of the R function `fac.layout` from the `dae` package to produce the randomized layout for this example. The `unrandomized` argument gives the single unrandomized factor indexing the units in the experiment. The `randomized` argument specifies the factor, Diets, that is to be randomized. The `seed` argument is used so that the same randomized layout for a particular

experiment can be generated at a later date. In general a different random small integer value, say between 0 and 1023, should be supplied as for this argument each time a new experimental layout is being obtained.

```
> #
> # Obtaining randomized layout for a CRD
> #
> n <- 6
> CRDRat.unit <- list(Rat = n)
> Diet <- factor(rep(c("A","B","C"), times = c(3,2,1)))
> CRDRat.lay <- fac.layout(unrandomized=CRDRat.unit, randomized=Diet, seed=695)
> CRDRat.lay
  Units Permutation Rat Diet
1     1           4   1    A
2     2           1   2    C
3     3           5   3    B
4     4           3   4    A
5     5           6   5    A
6     6           2   6    B
> #remove Diet object in workspace to avoid using it by mistake
> remove(Diet)
```

■

III.B Models and estimation for a CRD

The analysis of CRD experiments uses least-squares or maximum likelihood estimation of the parameters of a linear model and hypothesis testing based on the ANOVA method or maximum likelihood ratio testing. We will investigate linear models and the estimation of its parameters using our rat experiment.

a) Maximal model

Definition III.2: The **maximal expectation model** is the most complicated model for the expectation that is to be considered in analysing an experiment. ■

We first consider the maximal model for the CRD.

Example III.1 Rat experiment (continued)

Suppose, the experimenter measured the liver weight as a percentage of total body weight at the end of the experiment. The results of the experiment, in standard order, are as follows:

Rat	1	4	5	3	6	2
Diet	A	A	A	B	B	C
Liver wt.	3.3	3.1	2.9	3.2	3.4	2.7

The analysis of this experiment will be based on a linear model, that is

$$E\left[\mathbf{Y}\right] = \mathbf{X\theta} \text{ and } \text{var}\left[\mathbf{Y}\right] = \mathbf{V_Y} = \sigma^2\mathbf{I}_n .$$

Our model also involves assuming $\mathbf{Y} \sim N\left(\mathbf{X\theta}, \mathbf{V}_Y\right)$.

Now, the trick here is what are **X** and **θ** going to be? I suppose an intuitively obvious thing to do might be to allocate codes for diet (A = 1; B = 2; C = 3) and use these as values to form the explanatory variable whose values make up the columns of **X**. Thus, regression equations for the example, including a term for the intercept, are as follows:

$$\begin{bmatrix} E[Y_1] \\ E[Y_2] \\ E[Y_3] \\ E[Y_4] \\ E[Y_5] \\ E[Y_6] \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

Note that the numbering of the Y's does not correspond to that of the Rats. It does not affect the model and to order the observations like this is neater.

The above model can then be fitted using simple linear regression techniques. The fitted equation is:

$$\widehat{E[Y]} = 3.30 - 0.120x$$

That is, $\hat{\mu} = 3.30$, $\hat{\gamma} = -0.12$.

Now the model states that $\widehat{E[Y_i]} = \hat{\mu} + \hat{\gamma} \times x_i$

so that $\widehat{E[Y_i]} = 3.30 + (-0.12) \times x_i$

Hence $\widehat{E[Y_1]} = 3.30 + (-0.12) \times 1 = 3.18$

and $\widehat{E[Y_5]} = 3.30 + (-0.12) \times 2 = 3.06$

That is, for each unit increase in diet, liver weight decreases by 0.120. However, does this make sense? The fitted model says that, for each unit increase in diet, % liver weight decreases by 0.120. It is sensible **only if** the diets differences are based on equally spaced levels of some component; for example, if the diets represent 2, 4 and 6 mg of copper added to each 100g of food, then the unit increase in diet is 2 mg of copper. But, there is no unit increase if the diets are unequally spaced, such as 2,4, and 10 mg copper added, or the diets differ qualitatively, such as if the diets represented the addition of three different amino acids.

To overcome this problem the regression is performed on explanatory variables indicator variables. In this method the explanatory variables are also called *factors* and the possible values they take *levels* (remember they are limited). Thus, in our example, we have a factor Diet and it has three levels: A, B and C.

Definition III.3: **Indicator variables** are formed from a factor by creating a variable for each level of the factor. The values of a variable are either 1 or 0, 1 when the unit has the particular level for that variable and 0 otherwise. ■

Thus, our model becomes:

$$\begin{bmatrix} E[Y_1] \\ E[Y_2] \\ E[Y_3] \\ E[Y_4] \\ E[Y_5] \\ E[Y_6] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

Hence,

$$\mathrm{E}[Y_i] = \alpha_k \text{ and } var[Y_i] = \sigma^2,\ cov[Y_i, Y_j] = 0,\ (i \neq j).$$

These can be written as $\boldsymbol{\psi}_\mathrm{D} = E[\mathbf{Y}] = \mathbf{X}_\mathrm{D}\boldsymbol{\alpha}$ and $\mathbf{V} = \sigma^2\mathbf{I}_n$

where $\mathbf{X}_\mathrm{D}$ is a 6x3 matrix containing the indicator variables for the Diets and $\boldsymbol{\psi}_\mathrm{D}$ is a 6-vector containing the parameters (αs) under the model involving $\mathbf{X}_\mathrm{D}$.

This model suggests that there are 3 different expected (or mean) values for the diets. This contrasts with the previous model that says that liver weight increase linearly as the Diet increases. ■

Indeed a little reflection will allow one to see that $\mathbf{X}_\mathrm{T}$ for a set of t Treatments can be written generally in the form of the following partitioned matrix, provided **Y** is ordered so that all the observations for the first treatment occur in the first r_1 rows, those for the second treatment occur in the next r_2 rows and so on with the last treatment occurring in the last r_t rows.

$$\mathbf{X}_\mathrm{T} = \begin{bmatrix} \mathbf{1}_{r_1} & \mathbf{0}_{r_1\times 1} & \cdots & \mathbf{0}_{r_1\times 1} \\ \mathbf{0}_{r_2\times 1} & \mathbf{1}_{r_2} & \cdots & \mathbf{0}_{r_2\times 1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{r_t\times 1} & \mathbf{0}_{r_t\times 1} & \cdots & \mathbf{1}_{r_t} \end{bmatrix}$$

where $\mathbf{1}_{r_i}$ is the $r_i \times 1$ column vector of ones and $\mathbf{0}_{r_i\times 1}$ is the $r_i \times 1$ column vector of zeroes (the latter requires the number of rows and columns because **0** occurs both as vectors and matrices, whereas **1** occurs only as a vector with **J** being used for a matrix of ones). This order of treatments is that for the systematic layout that is obtained prior to randomization and is one of the possible randomized layouts. Hence it is as good as any other layout for providing the $\mathbf{X}_\mathrm{T}$ matrix

So, in general, the model for the expected values from a CRD is still of the general form $E[\mathbf{Y}] = \mathrm{X}\boldsymbol{\theta}$ and on assuming Y is distributed N($\boldsymbol{\psi}_\mathrm{D}$, $\sigma^2\mathbf{I}_n$), we can use standard least squares or maximum likelihood estimation. Note that N($\boldsymbol{\psi}_\mathrm{D}$, $\sigma^2\mathbf{I}_n$) stands for (multivariate) normal with expected value $\boldsymbol{\psi}_\mathrm{D}$ and variance-covariance $\sigma^2\mathbf{I}_n$.

It can be shown, by examining the OLS equation $\hat{\boldsymbol{\alpha}} = \left(\mathbf{X}_\mathrm{T}'\mathbf{X}_\mathrm{T}\right)^{-1}\mathbf{X}_\mathrm{T}'\mathbf{Y}$, that the estimates of the elements of α are the means of the treatments, as are those of the elements of ψ.

Example III.1 Rat experiment (continued)

The estimates of α are

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = \begin{bmatrix} 3.10 \\ 3.30 \\ 2.70 \end{bmatrix}$$

so that the estimates of the expected values, the fitted values, are given by:

$$\hat{\boldsymbol{\psi}}_\mathrm{D} = \mathbf{X}_\mathrm{D}\hat{\boldsymbol{\alpha}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3.10 \\ 3.30 \\ 2.70 \end{bmatrix} = \begin{bmatrix} 3.10 \\ 3.10 \\ 3.10 \\ 3.30 \\ 3.30 \\ 2.70 \end{bmatrix}$$

■

In general, $\hat{\boldsymbol{\psi}}_\mathrm{T} = \mathbf{X}_\mathrm{T}\hat{\boldsymbol{\alpha}} = \bar{\mathbf{T}}$ where $\bar{\mathbf{T}}$ is the *n*-vector consisting of the treatment means for each unit. As is generally the case with least squares, the estimator of the expectation parameters can be written as a linear combination of **Y**. That is, $\bar{\mathbf{T}}$ can be obtained as the product of an $n \times n$ matrix and the *n*-vector **Y**. Let us write $\bar{\mathbf{T}} = \mathbf{M}_\mathrm{T}\mathbf{Y}$. Now $\mathbf{M}_\mathrm{T}$ is the matrix that replaces each value of **Y** with the mean of the corresponding treatment. It can be shown that the general form of $\mathbf{M}_\mathrm{T}$ is

$$\mathbf{M}_\mathrm{T} = \begin{bmatrix} r_1^{-1}\mathbf{J}_{r_1} & \mathbf{0}_{r_1 \times r_2} & \cdots & \mathbf{0}_{r_1 \times r_t} \\ \mathbf{0}_{r_2 \times r_1} & r_2^{-1}\mathbf{J}_{r_2} & \cdots & \mathbf{0}_{r_2 \times r_t} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{r_t \times r_1} & \mathbf{0}_{r_t \times r_2} & \cdots & r_t^{-1}\mathbf{J}_{r_t} \end{bmatrix}$$

so that $\bar{\mathbf{T}} = \mathbf{M}_\mathrm{T}\mathbf{Y}$ and it is clear from the above expression that the first r_1 elements of $\bar{\mathbf{T}}$ are the mean of the Y_i s for the first treatment, the next r_2 elements are the mean of those for the second treatment and so on. That is,

$$\bar{\mathbf{T}}' = \begin{bmatrix} \bar{T}_1 & \cdots & \bar{T}_1 & \bar{T}_2 & \cdots & \bar{T}_2 & \cdots & \bar{T}_t & \cdots & \bar{T}_t \end{bmatrix}$$

So $\mathbf{M}_\mathrm{T}$ could be called the treatment mean operator as it computes the treatment means from the vector to which it is applied and replaces each element of this vector with its treatment mean.

As before the residuals are the estimates of the random errors in the observed values of **Y**. In this case, the estimator for the random errors is given by $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\boldsymbol{\psi}}_\mathrm{T} = \mathbf{Y} - \bar{\mathbf{T}}$.

Example III.1 Rat experiment (continued)

We know that the estimator of the expected values is $\hat{\boldsymbol{\psi}}_D = \bar{\mathbf{T}} = \mathbf{M}_D\mathbf{Y}$, where $\bar{\mathbf{T}}' = \begin{bmatrix} \bar{T}_1 & \bar{T}_1 & \bar{T}_1 & \bar{T}_2 & \bar{T}_2 & \bar{T}_3 \end{bmatrix}$, so that alternative expression for the fitted values is:

$$\begin{aligned}
\hat{\boldsymbol{\psi}}_D = \bar{\mathbf{t}} &= \mathbf{M}_D\mathbf{y} \\
&= \begin{bmatrix} \frac{1}{3}\mathbf{J}_3 & \mathbf{0}_{3\times 2} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{2\times 3} & \frac{1}{2}\mathbf{J}_2 & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 2} & \frac{1}{1}\mathbf{J}_1 \end{bmatrix}\mathbf{y} \\
&= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 3.3 \\ 3.1 \\ 2.9 \\ 3.2 \\ 3.4 \\ 2.7 \end{bmatrix} = \begin{bmatrix} 3.10 \\ 3.10 \\ 3.10 \\ 3.30 \\ 3.30 \\ 2.70 \end{bmatrix}
\end{aligned}$$

Note the use of $\bar{\mathbf{t}}$ instead of $\bar{\mathbf{T}}$ to indicate that these are actual estimates rather than estimators.

Also, the residuals are

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{e} = \mathbf{y} - \bar{\mathbf{t}} = \begin{bmatrix} 3.3 \\ 3.1 \\ 2.9 \\ 3.2 \\ 3.4 \\ 2.7 \end{bmatrix} - \begin{bmatrix} 3.10 \\ 3.10 \\ 3.10 \\ 3.30 \\ 3.30 \\ 2.70 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.0 \\ -0.2 \\ -0.1 \\ 0.1 \\ 0.0 \end{bmatrix}$$

■

It turns out that the fitted values for orthogonal experiments, a large number of the experiments discussed in this course, are functions of means. However, nonorthogonal experiments are a practically important class of experiments and the fitted values are not so easily obtained.

b) Alternative indicator-variable, expectation models

For the completely randomized design, two indicator-variable expectation models are considered:

1. $E[Y_i] = \mu$ *or* $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$ where $\mathbf{X}_G = \mathbf{1}_n$;
2. $E[Y_i] = \alpha_k$ *or* $\boldsymbol{\psi}_T = \mathbf{X}_T\boldsymbol{\alpha}$.

The first model, referred to as the minimal expectation model, says that the population mean response is the same for all observations, irrespective of diet. Of course, the second model is the maximal expectation model.

Definition III.4: The **minimal expectation model** is the simplest model for the expectation that is to be considered in analysing an experiment. ■

The minimal expectation model is the same as the intercept-only model given for the single sample in chapter I, *Statistical inference*, and this will be the case for all the analyses that we consider. Now as we saw, the estimator of the expected values for the intercept-only model is $\hat{\boldsymbol{\psi}}_G = \bar{\mathbf{G}}$ where $\bar{\mathbf{G}}$ is the *n*-vector each of whose elements is the grand mean. For the Rat experiment, $\hat{\boldsymbol{\psi}}_G = 3.1\mathbf{1}_6$.

Note that, as for simple regression, there is no interest in the model $\boldsymbol{\psi} = \mathbf{0}$ and only a minimal and a maximal model are under consideration for the CRD experiment. Now the minimal model is said to be marginal to the maximal model.

A marginal model can be obtained from the model to which it is marginal by the imposition of linear constraints. In the context of linear regression we obtained such a model by imposing the constraint that certain parameters are zero. Here, rather than setting certain parameters to zero, equality constraints are imposed on the parameters. In particular, the intercept-only model is simply obtained from the maximal model by setting $\alpha_k = \mu$. That is, this model is a special case in which all the α_ks are equal. Ho does this affect the matrix expression? Here, the intercept-only model is obtained by replacing each of the elements of α with μ i.e replacing $\boldsymbol{\alpha}$ with $\mathbf{1}_t\mu$ so that $\boldsymbol{\psi}_T = \mathbf{X}_T\boldsymbol{\alpha} = \mathbf{X}_T\mathbf{1}_t\mu = \mathbf{X}_G\mu$.

A consequence of imposing such constraints is that the columns of the **X** matrix for the marginal model can be written as a linear combination of those of the **X** matrix for the original model. This leads to the following general definition for marginality.

Definition III.5: Let $\mathcal{C}(\mathbf{X})$ denote the column space of **X**. For two models, $\boldsymbol{\psi}_1 = \mathbf{X}_1\boldsymbol{\theta}_1$ and $\boldsymbol{\psi}_2 = \mathbf{X}_2\boldsymbol{\theta}_2$, the first model is **marginal** to the second if $\mathcal{C}(\mathbf{X}_1) \subseteq \mathcal{C}(\mathbf{X}_2)$ irrespective of the replication of the levels in the columns of the **X**s. That is if the columns of $\mathbf{X}_1$ can always be written as linear combinations of the columns of $\mathbf{X}_2$. We write $\boldsymbol{\psi}_1 \leq \boldsymbol{\psi}_2$. ■

Note that the marginality relationship is not symmetric — it is directional, like the less-than relation that we use to symbolize it. So while $\boldsymbol{\psi}_1 \leq \boldsymbol{\psi}_2$, $\boldsymbol{\psi}_2$ is **not** marginal to $\boldsymbol{\psi}_1$ unless $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_2$.

For the two models being considered for the CRD, $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$ is marginal to the model $\boldsymbol{\psi}_T = \mathbf{X}_T\boldsymbol{\alpha}$ or $\boldsymbol{\psi}_G < \boldsymbol{\psi}_T$ because $\mathcal{C}(\mathbf{X}_G) \subset \mathcal{C}(\mathbf{X}_T)$ in that an element from a row of $\mathbf{X}_G$ is the sum of the elements in the corresponding row of $\mathbf{X}_T$ and this will occur irrespective of the replication of the levels in the columns of $\mathbf{X}_T$ and $\mathbf{X}_G$. So while $\boldsymbol{\psi}_G < \boldsymbol{\psi}_T$, $\boldsymbol{\psi}_T$ is **not** marginal to $\boldsymbol{\psi}_G$ as $\mathcal{C}(\mathbf{X}_T) \not\subset \mathcal{C}(\mathbf{X}_G)$ so that $\boldsymbol{\psi}_T \not< \boldsymbol{\psi}_G$. In geometrical terms, $\mathcal{C}(\mathbf{X}_T)$ is a three-dimensional space and $\mathcal{C}(\mathbf{X}_G)$ is a line, the equiangular line, that is a subspace of $\mathcal{C}(\mathbf{X}_T)$.

III.C Hypothesis testing using the ANOVA method

Often an experimenter will be interested in determining if there are significant differences between the treatment means; in the example, whether or not there are significant differences between the diet means. This is equivalent to deciding which of our two expectation models best describes the data. We now perform the hypothesis test to do this for the example.

a) Analysis of the rat example

Example III.1 Rat experiment (continued)

Step 1: Set up hypotheses

H_0: $\alpha_A = \alpha_B = \alpha_C = \mu$ $\quad (\boldsymbol{\psi}_G = \mathbf{X}_G\mu)$

H_1: not all population Diet means are equal $\quad (\boldsymbol{\psi}_D = \mathbf{X}_D\boldsymbol{\alpha})$

Set $\alpha = 0.05$.

Step 2: Calculate test statistic

Source	df	SSq	MSq	F	Prob
Rats	5	0.34			
Diets	2	0.24	0.1200	3.60	0.1595
Residual	3	0.10	0.0332		

From this table we can see that we have taken the (corrected) total variation amongst the six Rats and partitioned it into two parts: variance of difference between diet means and the left-over (residual) rat variation.

Step 3: Decide between hypotheses

As the probability of exceeding an F of 3.60 with $\nu_1 = 2$ and $\nu_2 = 3$ is 0.1595 and this is greater than $\alpha = 0.05$, there is not much evidence of a diet difference and the expectation model that appears to provide an adequate description of the data is $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$.

b) Sums of squares for the analysis of variance

As we are using the ANOVA method, an analysis of variance table is formulated based on partitioning the Total corrected sum of squares into two sums of squares, one reflecting Treatment differences and the other Residual variation. We then compute an F statistic to be used in deciding which model best describes the data.

Now we saw in chapter I, *Statistical inference*, that an SSq is the sum of squares of the elements of a vector and that we can write the sum of squares as the product of the transpose of a column vector with the original column vector. The estimators of the sum of squares for the CRD ANOVA are the sums of squares of the following vectors:

Total or Units SSq: $\mathbf{D}_G = \mathbf{Y} - \bar{\mathbf{G}}$

Treatments SSq: $\mathbf{T}_e = \bar{\mathbf{T}} - \bar{\mathbf{G}}$

Residual SSq: $\mathbf{D}_T = \mathbf{Y} - \bar{\mathbf{T}}$

where the **D**s are *n*-vectors of **d**eviations from **Y** and $\mathbf{T}_e$ is the *n*-vector of **T**reatment ***e***ffects.

Definition III.6: An effect is a linear combination of means with a set of effects summing to zero. ■

Now, as in chapter I, *Statistical Inference*, $\mathbf{D}_G$ is just the observations minus the grand mean estimator that yields the Corrected Total sum of squares, $\mathbf{T}_e$ is difference between the expected values for the alternative and null models that yields the Model SSq, and $\mathbf{D}_T$ is the deviations of the observations from the expected values under the alternative model.

We want to show that estimators of all the sums of squares can be written in the following alternative form $\mathbf{Y}'\mathbf{Q}\mathbf{Y}$. Note that, as this is the product of $1\times n$, $n\times n$ and $n\times 1$ vectors and matrix, it is 1×1 or a scalar.

Definition III.7: A **quadratic form** in a vector **Y** is a scalar function of **Y** of the form $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ where **A** is called the matrix of the quadratic form. ■

Firstly write $\mathbf{Y} = \mathbf{I}_n\mathbf{Y} = \mathbf{M}_U\mathbf{Y}$, $\bar{\mathbf{G}} = \frac{1}{n}\mathbf{J}_n\mathbf{Y} = \mathbf{M}_G\mathbf{Y}$ and $\bar{\mathbf{T}} = \mathbf{M}_T\mathbf{Y}$. That is, each of the individual vectors on which the sums of squares are based can be written as an **M** matrix times **Y**. These **M** matrices are mean operators that are symmetric and idempotent in that $\mathbf{M}' = \mathbf{M}$ and $\mathbf{M}^2 = \mathbf{M}$ in all cases.

Then

$$\mathbf{D}_G = \mathbf{Y} - \bar{\mathbf{G}} = \mathbf{M}_U\mathbf{Y} - \mathbf{M}_G\mathbf{Y} = (\mathbf{M}_U - \mathbf{M}_G)\mathbf{Y} = \mathbf{Q}_U\mathbf{Y} \quad \text{with } \mathbf{Q}_U = \mathbf{M}_U - \mathbf{M}_G$$

$$\mathbf{T}_e = \bar{\mathbf{T}} - \bar{\mathbf{G}} = \mathbf{M}_T\mathbf{Y} - \mathbf{M}_G\mathbf{Y} = (\mathbf{M}_T - \mathbf{M}_G)\mathbf{Y} = \mathbf{Q}_T\mathbf{Y} \quad \text{with } \mathbf{Q}_T = \mathbf{M}_T - \mathbf{M}_G$$

$$\mathbf{D}_T = \mathbf{Y} - \bar{\mathbf{T}} = \mathbf{M}_U\mathbf{Y} - \mathbf{M}_T\mathbf{Y} = (\mathbf{M}_U - \mathbf{M}_T)\mathbf{Y} = \mathbf{Q}_{U_{Res}}\mathbf{Y} \quad \text{with } \mathbf{Q}_{U_{Res}} = \mathbf{M}_U - \mathbf{M}_T$$

It is relatively straightforward to show that the three **Q** matrices are symmetric and idempotent. It can also be shown that $\mathbf{Q}_T\mathbf{Q}_{U_{Res}} = \mathbf{Q}_{U_{Res}}\mathbf{Q}_T = \mathbf{0}$.

Consequently we obtain the following expressions for the sums of squares:

$$\begin{aligned}
\mathbf{D}_G'\mathbf{D}_G &= \left(\mathbf{Y} - \bar{\mathbf{G}}\right)'\left(\mathbf{Y} - \bar{\mathbf{G}}\right) \\
&= \left(\mathbf{Q}_U\mathbf{Y}\right)'\left(\mathbf{Q}_U\mathbf{Y}\right) \\
&= \mathbf{Y}'\mathbf{Q}_U'\mathbf{Q}_U\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Q}_U\mathbf{Q}_U\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Q}_U\mathbf{Y}
\end{aligned}$$

$$\begin{aligned}\mathbf{T}_e'\mathbf{T}_e &= \left(\overline{\mathbf{T}} - \overline{\mathbf{G}}\right)'\left(\overline{\mathbf{T}} - \overline{\mathbf{G}}\right)\\ &= \left(\mathbf{Q}_\mathrm{T}\mathbf{Y}\right)'\left(\mathbf{Q}_\mathrm{T}\mathbf{Y}\right)\\ &= \mathbf{Y}'\mathbf{Q}_\mathrm{T}'\mathbf{Q}_\mathrm{T}\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{Q}_\mathrm{T}\mathbf{Q}_\mathrm{T}\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{Q}_\mathrm{T}\mathbf{Y}\end{aligned}$$

$$\begin{aligned}\mathbf{D}_\mathrm{T}'\mathbf{D}_\mathrm{T} &= \left(\mathbf{Y} - \overline{\mathbf{T}}\right)'\left(\mathbf{Y} - \overline{\mathbf{T}}\right)\\ &= \left(\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}\right)'\left(\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}\right)\\ &= \mathbf{Y}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}\end{aligned}$$

Theorem III.1: For a completely randomized design, the sums of squares in the analysis of variance for Units, Treatments and Residual are given by the quadratic forms

$$\mathbf{Y}'\mathbf{Q}_\mathrm{U}\mathbf{Y},\ \mathbf{Y}'\mathbf{Q}_\mathrm{T}\mathbf{Y} \text{ and } \mathbf{Y}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}, \text{ respectively,}$$

where $\mathbf{Q}_\mathrm{U} = \mathbf{M}_\mathrm{U} - \mathbf{M}_\mathrm{G}$, $\mathbf{Q}_\mathrm{T} = \mathbf{M}_\mathrm{T} - \mathbf{M}_\mathrm{G}$ and $\mathbf{Q}_{\mathrm{U}_\mathrm{Res}} = \mathbf{M}_\mathrm{U} - \mathbf{M}_\mathrm{T}$ and

$$\mathbf{M}_\mathrm{U} = \mathbf{I}_n,\ \mathbf{M}_\mathrm{G} = \tfrac{1}{n}\mathbf{J}_n \text{ and } \mathbf{M}_\mathrm{T} = \begin{bmatrix} r_1^{-1}\mathbf{J}_{r_1} & \mathbf{0}_{r_1\times r_2} & \cdots & \mathbf{0}_{r_1\times r_t} \\ \mathbf{0}_{r_2\times r_1} & r_2^{-1}\mathbf{J}_{r_2} & \cdots & \mathbf{0}_{r_2\times r_t} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{r_t\times r_1} & \mathbf{0}_{r_t\times r_2} & \cdots & r_t^{-1}\mathbf{J}_{r_t} \end{bmatrix}$$

Proof: follows the argument given above. ■

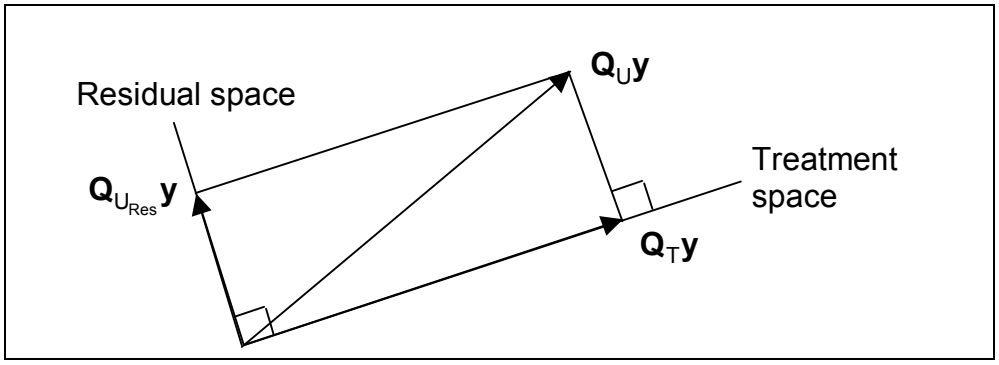
Note that $\mathbf{Q}_\mathrm{U} - \mathbf{Q}_\mathrm{T} = \left(\mathbf{M}_\mathrm{U} - \mathbf{M}_\mathrm{G}\right) - \left(\mathbf{M}_\mathrm{T} - \mathbf{M}_\mathrm{G}\right) = \mathbf{M}_\mathrm{U} - \mathbf{M}_\mathrm{T} = \mathbf{Q}_{\mathrm{U}_\mathrm{Res}}$ so that $\mathbf{y}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{y} = \mathbf{y}'\left(\mathbf{Q}_\mathrm{U} - \mathbf{Q}_\mathrm{T}\right)\mathbf{y} = \mathbf{y}'\mathbf{Q}_\mathrm{U}\mathbf{y} - \mathbf{y}'\mathbf{Q}_\mathrm{T}\mathbf{y}$. That is, the Residual SSq is the Units SSq minus the Treatments SSq.

So the analysis of variance table is constructed as follows:

Source	df	SSq	MSq (s^2)	F	p
Units	$n-1$	$\mathbf{Y}'\mathbf{Q}_\mathrm{U}\mathbf{Y}$			
Treatments	$t-1$	$\mathbf{Y}'\mathbf{Q}_\mathrm{T}\mathbf{Y}$	$\dfrac{\mathbf{Y}'\mathbf{Q}_\mathrm{T}\mathbf{Y}}{t-1} = s_\mathrm{T}^2$	$s_\mathrm{T}^2 / s_{\mathrm{U}_\mathrm{Res}}^2$	p_T
Residual	$n-t$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}$	$\dfrac{\mathbf{Y}'\mathbf{Q}_{\mathrm{U}_\mathrm{Res}}\mathbf{Y}}{n-t} = s_{\mathrm{U}_\mathrm{Res}}^2$		

Now the situation here is exactly the same as outlined for regression in chapter I, *Statistical Inference*. The **Q** matrices, being symmetric and idempotent, are

orthogonal projection matrices. The matrix $\mathbf{Q}_U$ orthogonally projects the data vector into the $n-1$ dimensional part of the n-dimensional data space that is orthogonal to equiangular line. The matrix $\mathbf{Q}_T$ orthogonally projects the data vector into the $t-1$ dimensional part of the t-dimensional Treatment space that is orthogonal to equiangular line. Here the Treatment space is the column space of the matrix $\mathbf{X}_T$. Finally, the matrix $\mathbf{Q}_{U_{Res}}$ orthogonally projects the data vector into the $n-t$ dimensional Residual subspace. That is, the Units space is divided into the two orthogonal subspaces, the Treatments and Residual subspaces. This is represented diagrammatically in the following diagram.



Of course, the sums of squares are just the squared lengths of these vectors and, according to Pythagoras' theorem, the Treatments and Residual sums of squares must sum to the Units sum of squares.

Example III.1 Rat experiment (continued)

The vectors for computing the sums of squares are given in the following table.

Diet	Liver wt. $\mathbf{y}$	Grand mean $\bar{\mathbf{g}} = \mathbf{M}_G\mathbf{y}$	Total Rat deviations $\mathbf{d}_G = \mathbf{Q}_R\mathbf{y} = \mathbf{y} - \bar{\mathbf{g}}$	Diet means $\bar{\mathbf{t}} = \mathbf{M}_D\mathbf{y}$	Diet effects $\mathbf{t}_e = \mathbf{Q}_D\mathbf{y} = \bar{\mathbf{t}} - \bar{\mathbf{g}}$	Residual Rat deviations $\mathbf{d}_D = \mathbf{Q}_{R_{Res}}\mathbf{y} = \mathbf{y} - \bar{\mathbf{t}}$
A	3.3	3.1	0.2	3.1	0.0	0.2
A	3.1	3.1	0.0	3.1	0.0	0.0
A	2.9	3.1	−0.2	3.1	0.0	−0.2
B	3.2	3.1	0.1	3.3	0.2	−0.1
B	3.4	3.1	0.3	3.3	0.2	0.1
C	2.7	3.1	−0.4	2.7	−0.4	0.0
SSq			0.34		0.24	0.10

In particular, we are interested in the Total Rat deviations, the Diet Effects and the Residual Rats deviations, these being the projections into the Rats, Diets and Residual subspaces that are of dimension 5, 2 and 3 respectively. The squared lengths of the projections into these subspaces are obtained as the sums of squares of the elements of the vectors. The Rats SSq is $\mathbf{Y}'\mathbf{Q}_R\mathbf{Y} = 0.34$, the Diets SSq is $\mathbf{Y}'\mathbf{Q}_D\mathbf{Y} = 0.24$ and the Residual SSq is $\mathbf{Y}'\mathbf{Q}_{R_{Res}}\mathbf{Y} = 0.10$. ■

c) Expected mean squares

So we have an analysis of variance in which we use an F value that is the ratio of two mean squares. But why is this appropriate? To answer this question we look at what the two mean squares measure and this is done using the expected values of the mean squares in the analysis of variance table. We will need to determine the expected mean squares (E[MSq]s) under the maximal model:

$$\boldsymbol{\psi}_T = E\left[\mathbf{Y}\right] = \mathbf{X}_T\boldsymbol{\alpha} \text{ and } \mathbf{V}_Y = \sigma^2\mathbf{I}_n$$

and the minimal model:

$$\boldsymbol{\psi}_G = E\left[\mathbf{Y}\right] = \mathbf{X}_G\mu \text{ and } \mathbf{V}_Y = \sigma^2\mathbf{I}_n .$$

The expected mean squares are the mean values of the mean squares in populations that are described by the model for which they are derived i.e. an expected mean square is its true mean value and it turns out that it depends on the parameters of the model. This is analogous to saying that the expected value of an observation in the population is its mean value. That is, if we ask "what is E[Y_i]?", the answer is E[Y_i] = α_k. So, analogously, what is E[MSq] for the Treatment mean square?

The expected mean squares under the maximal model are given in the following table:

Source	df	MSq (s^2)	E[MSq]	F
Units	$n-1$			
Treatments	$t-1$	$\frac{\mathbf{Y}'\mathbf{Q}_T\mathbf{Y}}{t-1} = s_T^2$	$\sigma^2 + q_T(\boldsymbol{\psi})$	$s_T^2 / s_{U_{Res}}^2$
Residual	$n-t$	$\frac{\mathbf{Y}'\mathbf{Q}_{U_{Res}}\mathbf{Y}}{n-t} = s_{U_{Res}}^2$	σ^2	

So if we had the complete populations for all Treatments and computed the mean squares from this complete data, then the values of the mean squares would be as given in the table. Notice that the average value, in the populations, of both mean squares involves σ^2, the uncontrolled variation amongst units from the same treatment. However, the average value for the Treatments mean square also depends on the quantity $q_T(\boldsymbol{\psi})$. So what is $q_T(\boldsymbol{\psi})$ about?

Firstly, the subscript T in $q_T(\boldsymbol{\psi})$ indicates that it involves $\mathbf{Q}_T$. Indeed $q_T(\boldsymbol{\psi}) = \boldsymbol{\psi}'\mathbf{Q}_T\boldsymbol{\psi}/(t-1)$ whose numerator is the same as the sum of squares except that it is a quadratic form in $\boldsymbol{\psi}$ instead of $\mathbf{Y}$. Now, so far we have not included a subscript on the $\boldsymbol{\psi}$ in $q_T(\boldsymbol{\psi})$, because we will determine expressions for it under both the maximal model $(\boldsymbol{\psi}_T)$ and the alternative model $(\boldsymbol{\psi}_G)$. That is, $\boldsymbol{\psi}$ in $q_T(\boldsymbol{\psi})$ will vary.

To further understand what $q_T(\boldsymbol{\psi}) = \boldsymbol{\psi}'\mathbf{Q}_T\boldsymbol{\psi}/(t-1)$ means we derive expressions for it in terms of the individual parameters of expectation model. We will show that under

the maximal model, $(\boldsymbol{\psi}_T)$, $q_T(\boldsymbol{\psi}_T) = \boldsymbol{\psi}_T'\mathbf{Q}_T\boldsymbol{\psi}_T/(t-1) = \sum_{k=1}^{t} r_k(\alpha_k - \bar{\alpha}_.)^2 \Big/ (t-1)$, where $\bar{\alpha}_. = \sum_{k=1}^{t} r_k\alpha_k/n = (3\alpha_1 + 2\alpha_2 + \alpha_3)/6$, that under the minimal model $(\boldsymbol{\psi}_G = \mu\mathbf{1}_n)$ $q_T(\boldsymbol{\psi}_G) = 0$. Thus, if we knew the values for the α_ks and for σ^2, we could compute the population mean of the mean squares.

To derive the required expressions, observe that as $\mathbf{Q}_T$ is symmetric and idempotent, $\boldsymbol{\psi}'\mathbf{Q}_T\boldsymbol{\psi} = (\mathbf{Q}_T\boldsymbol{\psi})'\mathbf{Q}_T\boldsymbol{\psi}$, and $q_T(\boldsymbol{\psi})$ is the sum of squares of $\mathbf{Q}_T\boldsymbol{\psi} = (\mathbf{M}_T - \mathbf{M}_G)\boldsymbol{\psi} = \mathbf{M}_T\boldsymbol{\psi} - \mathbf{M}_G\boldsymbol{\psi}$, divided by $t-1$. Now $\mathbf{M}_G\boldsymbol{\psi}$ replaces each element of $\boldsymbol{\psi}$ with the grand mean of the elements of $\boldsymbol{\psi}$ and $\mathbf{M}_T\boldsymbol{\psi}$ replaces each element of $\boldsymbol{\psi}$ with the mean of the elements of $\boldsymbol{\psi}$ that received the same treatment as the element being replaced.

Under the maximal model $(\boldsymbol{\psi}_T)$, $\mathbf{M}_G\boldsymbol{\psi}_T = \bar{\alpha}_.\mathbf{1}_n$ and $\mathbf{M}_T\boldsymbol{\psi}_T = \boldsymbol{\psi}_T$ so that $\boldsymbol{\psi}_T'\mathbf{Q}_T\boldsymbol{\psi}_T$ is the sum of squares of the elements of $\boldsymbol{\psi}_T - \bar{\alpha}_.\mathbf{1}_n$, that is, the sum of squares of $(\alpha_k - \bar{\alpha}_.)$. From this, it is a straightforward matter to show that $q_T(\boldsymbol{\psi}_T) = \boldsymbol{\psi}_T'\mathbf{Q}_T\boldsymbol{\psi}_T/(t-1) = \sum_{k=1}^{t} r_k(\alpha_k - \bar{\alpha}_.)^2 \Big/ (t-1)$. On the other hand, under the minimal model $(\boldsymbol{\psi}_G = \mu\mathbf{1}_n)$, we have that $\mathbf{M}_G\boldsymbol{\psi}_G = \mathbf{M}_T\boldsymbol{\psi}_G = \mu\mathbf{1}_n$ so that $\boldsymbol{\psi}_G'\mathbf{Q}_T\boldsymbol{\psi}_G = 0$ and $q_T(\boldsymbol{\psi}_G) = 0$. Put another way, we have that $\alpha_k = \mu$ so that $\bar{\alpha}_. = \sum_{k=1}^{t} r_k\alpha_k/n = \mu\sum_{k=1}^{t} r_k/n = \mu$ and $\alpha_k - \bar{\alpha}_. = \mu - \mu = 0$. Hence, $q_T(\boldsymbol{\psi}_G) = 0$.

Example III.1 Rat experiment (continued)

In the liver weight example,

$$\mathbf{Q}_D\boldsymbol{\psi}_D = \mathbf{M}_D\begin{bmatrix}\alpha_1\\ \alpha_1\\ \alpha_1\\ \alpha_2\\ \alpha_2\\ \alpha_3\end{bmatrix} - \mathbf{M}_G\begin{bmatrix}\alpha_1\\ \alpha_1\\ \alpha_1\\ \alpha_2\\ \alpha_2\\ \alpha_3\end{bmatrix}$$

$$= \begin{bmatrix}\alpha_1\\ \alpha_1\\ \alpha_1\\ \alpha_2\\ \alpha_2\\ \alpha_3\end{bmatrix} - \begin{bmatrix}\bar{\alpha}_.\\ \bar{\alpha}_.\\ \bar{\alpha}_.\\ \bar{\alpha}_.\\ \bar{\alpha}_.\\ \bar{\alpha}_.\end{bmatrix}$$

$$= \begin{bmatrix}\alpha_1 - \bar{\alpha}_.\\ \alpha_1 - \bar{\alpha}_.\\ \alpha_1 - \bar{\alpha}_.\\ \alpha_2 - \bar{\alpha}_.\\ \alpha_2 - \bar{\alpha}_.\\ \alpha_3 - \bar{\alpha}_.\end{bmatrix}$$

so that $q_{\mathrm{D}}(\boldsymbol{\psi}_{\mathrm{D}})$ is the sum of squares of the latter vector. Hence,

$$q_{\mathrm{D}}(\boldsymbol{\psi}_{\mathrm{D}}) = \boldsymbol{\psi}_{\mathrm{D}}'\mathbf{Q}_{\mathrm{D}}\boldsymbol{\psi}_{\mathrm{D}}/(t-1)$$
$$= \left\{3(\alpha_1 - \bar{\alpha}_\cdot)^2 + 2(\alpha_2 - \bar{\alpha}_\cdot)^2 + (\alpha_3 - \bar{\alpha}_\cdot)^2\right\}\Big/(3-1) = \sum_{k=1}^{t} r_k(\alpha_k - \bar{\alpha}_\cdot)^2 \Big/ (t-1).$$

Also,

$$\mathbf{Q}_{\mathrm{D}}\boldsymbol{\psi}_{\mathrm{G}} = \mathbf{M}_{\mathrm{D}}\begin{bmatrix}\mu\\ \mu\\ \mu\\ \mu\\ \mu\\ \mu\end{bmatrix} - \mathbf{M}_{\mathrm{G}}\begin{bmatrix}\mu\\ \mu\\ \mu\\ \mu\\ \mu\\ \mu\end{bmatrix}$$
$$= \begin{bmatrix}\mu\\ \mu\\ \mu\\ \mu\\ \mu\\ \mu\end{bmatrix} - \begin{bmatrix}\mu\\ \mu\\ \mu\\ \mu\\ \mu\\ \mu\end{bmatrix} = \begin{bmatrix}0\\ 0\\ 0\\ 0\\ 0\\ 0\end{bmatrix}$$

so that

$$q_{\mathrm{D}}(\boldsymbol{\psi}_{\mathrm{G}}) = \boldsymbol{\psi}_{\mathrm{G}}'\mathbf{Q}_{\mathrm{D}}\boldsymbol{\psi}_{\mathrm{G}}/(t-1) = 0$$

■

Under the maximal model, the expected value of the Treatment mean square differs from that for the Residual mean square by $q_{\mathrm{T}}(\boldsymbol{\psi})$. This is a quadratic form and is basically a sum of squares so that it must be nonnegative. Indeed the magnitude of $q_{\mathrm{T}}(\boldsymbol{\psi})$ depends on the size of the differences between the population treatment means, the α_ks — if all the α_ks are similar they will be close to their mean, $\bar{\alpha}$, whereas if they are widely scattered several will be some distance from their mean. Consequently, the Treatment mean square will on average be greater than the Residual mean square as it is influenced by both uncontrolled variation and the magnitude of treatment differences.

The quadratic form $q_{\mathrm{T}}(\boldsymbol{\psi})$ will only be zero when all the αs are equal, that is when the null hypothesis is true. Then the expected mean squares under the minimal model are equal so that the F value will be approximately one.

The above interpretation is not surprising if we consider the differences that are likely to contribute to differences between treatment means.

Example III.1 Rat experiment (continued)

In the liver weight example, the data and treatment means are as follows:

	Diet		
	A	B	C
	3.3	3.2	2.7
	3.1	3.4	
	2.9		
Mean	3.1	3.3	2.7

So what can potentially contribute to the observed difference 3.1 and 2.7? **Answer**: Obviously, the different diets. But what is not so obvious is that differences arising from uncontrolled variation also contribute as two different groups of rats are involved. This is then reflected in the expected means squares in that it involves σ^2 and the "variance" of the 3 effects. ■

Thus, the F test involves asking the question "Is the variance in the sample treatment means greater than can be expected from uncontrolled variation alone?".

If the variance is no greater, it is concluded that $q_T(\boldsymbol{\psi})$ is zero and the minimal model is the correct model since the expected Treatment mean square under this model is just σ^2. Otherwise, if the variance is greater, $q_T(\boldsymbol{\psi})$ is nonzero and the maximal model is required to describe the data. This is essentially the same argument as was used in deciding whether Concentration affected Strength in Example II.2, *Paper bag experiment*.

d) Summary of the hypothesis test

We summarize the ANOVA-based hypothesis test for a CRD involving t treatments and a total of n observed units.

Step 1: Set up hypotheses

H_0: $\alpha_1 = \alpha_2 = ... = \alpha_t = \mu$ (or $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$)

H_1: not all population Treatment means are equal (or $\boldsymbol{\psi}_T = \mathbf{X}_T\boldsymbol{\alpha}$)

Set α.

Step 2: Calculate test statistic

The analysis of variance table for a CRD is:

Source	df	MSq (s^2)	E[MSq]	F	p
Units	$n-1$				
Treatments	$t-1$	$\frac{\mathbf{Y}'\mathbf{Q}_{\mathrm{T}}\mathbf{Y}}{t-1} = s^2_{\mathrm{T}}$	$\sigma^2 + q_{\mathrm{T}}(\boldsymbol{\psi})$	$s^2_{\mathrm{T}} / s^2_{\mathrm{U_{Res}}}$	p_{T}
Residual	$n-t$	$\frac{\mathbf{Y}'\mathbf{Q}_{\mathrm{U_{Res}}}\mathbf{Y}}{n-t} = s^2_{\mathrm{U_{Res}}}$	σ^2		

Note that the expected mean squares under the maximal model are included in this table, it being recognized that when the null hypothesis is true $q_{\mathrm{T}}(\boldsymbol{\psi}) = 0$.

Step 3: Decide between hypotheses

Determine probability of observed F value that has ν_1 = numerator d.f. = t-1 and ν_2 = denominator d.f. = $n-t$.

If $\Pr\{F_{t-1,n-t} \geq F_O\} = p_{\mathrm{T}} \leq \alpha$ then the evidence suggests that the null hypothesis should be rejected and the alternative model be used to describe the data.

e) Comparison with traditional one-way ANOVA

The above analysis of variance table is essentially the same as the traditional one-way ANOVA table — the values of the F statistic from each table are exactly the same. As illustrated in the table below, the labelling differs and the Total is normally be placed at the bottom of the table, not at the top. The difference between the two tables is symbolic. The Units term explicitly represents a source of uncontrolled variation: differences between Units. Further, our table exhibits the confounding in the experiment. The indenting of Treatments under Units signifies that treatment differences are confounded or "mixed-up" with unit differences. Also, the Residual reflects differences between the units once the treatment differences have been removed or subtracted off. This is clear from the expected mean squares for this analysis.

Source	df	Source in traditional one-way ANOVA
Units	$n-1$	Total
Treatments	$t-1$	Between Treatments
Residual	$n-t$	Within Treatments

f) Computation of the ANOVA in R

Generally you would begin with entering the data and an initial graphical exploration of it. Data entry is covered in the *Appendix A, Introduction to R* and, in more detail, in

Appendix C, Analysis of designed experiments in R. The initial graphical exploration of the data would involve say boxplots. We will defer their use to a second example as there is insufficient data for *Example III.1, Rat experiment*, to employ boxplots.

Here we will concentrate on obtaining the analysis of variance. While the function `lm` could be used, the function `aov` is preferred for analysing data from a designed experiment. Both of these are based on specifying a *model formula* of the form:

Response variable ~ *explanatory variables* (and *operators*)

So far the expressions on the right have been fairly simple — one or two explanatory variables separated by a "+". However, there is a subtlety that arises in connection with the analysis of designed experiments. If the explanatory variable is a numeric, such as a numeric vector, then R fits just one coefficient for it. So for a single explanatory variable, a straight-line relationship between the response and explanatory variables is fitted. On the other hand, if the explanatory variable is categorical, such as a factor, a coefficient is fit for each level of the variable.

So in analyzing a completely randomized design it is important that the treatment factor is stored in a factor object and not a vector object. Being stored in a factor object signals to R that it should use indicator variables, or their equivalent, instead of values for the variable itself in estimating the parameters.

There are two ways in which this analysis can be obtained using the `aov` function: without and with an `Error` function in the model formula. The `Error` function is used in the model formula to specify a model for the Error in the experiment, that is a model for the uncontrolled variation. As usual the `summary` function is used and this produces the analysis of variance table. Also, the `model.tables` function can be used to obtain tables of means.

Example III.1 Rat experiment (continued)

The following commands are used to perform the two analyses of the data.

```
# AOV without Error
#
Rat.NoError.aov <- aov(LiverWt ~ Diet, CRDRat.dat)
summary(Rat.NoError.aov)
#
# AOV with Error
#
Rat.aov <- aov(LiverWt ~ Diet + Error(Rat), CRDRat.dat)
summary(Rat.aov)
model.tables(Rat.aov, type = "means")
```

The output of these commands is as follows:

```
> # AOV without Error
> #
> Rat.NoError.aov <- aov(LiverWt ~ Diet, CRDRat.dat)
> summary(Rat.NoError.aov)
            Df   Sum Sq  Mean Sq F value Pr(>F)
Diet         2 0.240000 0.120000     3.6 0.1595
Residuals    3 0.100000 0.033333
```

```
> #
> # AOV with Error
> #
> Rat.aov <- aov(LiverWt ~ Diet + Error(Rat), CRDRat.dat)
> summary(Rat.aov)

Error: Rat
          Df   Sum Sq  Mean Sq F value Pr(>F)
Diet       2 0.240000 0.120000     3.6 0.1595
Residuals  3 0.100000 0.033333
> model.tables(Rat.aov, type="means")
Tables of means
Grand mean

3.1

 Diet
      A   B   C
    3.1 3.3 2.7
rep 3.0 2.0 1.0
```

The analysis without the Error function parallels the traditional analysis and the analysis with Error is similar to our table. In this course we will use the Error function to do the analysis in R.

III.D Diagnostic checking

(Box, Hunter and Hunter sec.6.5)

In performing the analysis we assumed the data followed a certain model, namely, that $\mathbf{Y}$ is distributed N($\boldsymbol{\psi}$, $\sigma^2\mathbf{I}_n$) where $\boldsymbol{\psi} = E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\theta}$ and σ^2 is the variance of an observation. The maximal expectation model is used in diagnostic checking: $\boldsymbol{\psi}_T = E[\mathbf{Y}] = \mathbf{X}_T\boldsymbol{\alpha}$. For the complete model to be appropriate requires that:

a) the response is operating additively, that is, that the individual deviations in the response to a treatment are similar for all treatments;
b) that the sets of units assigned the treatments are comparable in that the amount of uncontrolled variation exhibited by them is the same for each treatment;
c) each observation is independent of other observations; and
d) that the response of the units is normally distributed.

Now, it may be that the data do not conform to the assumptions inherent in the model and that the conclusions we have drawn based on the model are incorrect. Thus, it would seem highly desirable to check the adequacy of our model in describing the data.

Example III.2 Caffeine effects on students

To illustrate the procedures, I am going to use an experiment in which the effect of orally ingested caffeine on a physical task was investigated (Draper and Smith, 1981, sec.9.1). Thirty healthy male college students were selected and trained in finger tapping. Ten men were randomly assigned to receive one of three doses of caffeine (0, 100 or 200 mg). The number of finger taps after ingesting the caffeine was recorded for each student and the data were as follows:

Caffeine Dose (mg)		
0	100	200
242	248	246
245	246	248
244	245	250
248	247	252
247	248	248
248	250	250
242	247	246
244	246	248
246	243	245
242	244	250

The first task is to set up a data frame containing the data. Since we have the data already arranged in standard order, we will enter it in this order. So, the data frame needs to contain a factor Students with values 1–30, a factor Dose with levels 0, 100 and 200 and values depending on whether the data is entered by rows or columns, and a numeric vector Taps with the 30 observed values of the response variable.

The following expressions could be used to set up the factors in a `data.frame` that I have named *CRDCaff.dat*.

```
> #set up data.frame with factors Students and Dose and response variable Taps
> CRDCaff.dat <- data.frame(Students = factor(1:30),
+                           Dose = factor(rep(c(0,100,200), times=10)))
> CRDCaff.dat$Taps <-
+        c(242,248,246,245,246,248,244,245,250,248,247,252,247,248,248,
+          248,250,250,242,247,246,244,246,248,246,243,245,242,244,250)
```

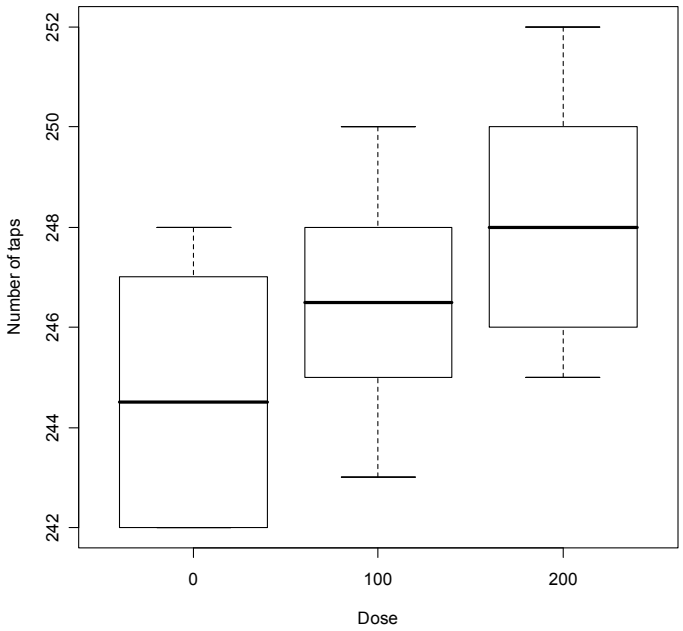
The data in the resulting data frame would be as follows:

```
> CRDCaff.dat
   Students Dose Taps
1         1    0  242
2         2  100  248
3         3  200  246
4         4    0  245
5         5  100  246
6         6  200  248
7         7    0  244
8         8  100  245
9         9  200  250
10       10    0  248
11       11  100  247
12       12  200  252
13       13    0  247
14       14  100  248
15       15  200  248
16       16    0  248
17       17  100  250
18       18  200  250
19       19    0  242
20       20  100  247
21       21  200  246
22       22    0  244
23       23  100  246
24       24  200  248
25       25    0  246
```

```
26       26  100  243
27       27  200  245
28       28    0  242
29       29  100  244
30       30  200  250
```

Next we produce boxplots for each level of Dose. This can be done by using the function `boxplot(split(Taps, Dose), xlab="Dose", ylab="Number of taps")`. In the `split` function we are specifying that that the response (*Dependent*) variable is Taps and the explanatory (*Independent*) variable is Dose.

The boxplot produced is shown in the following diagram.



The average number of taps would appear to be increasing as the dose increases. There is some evidence of dose 0 having more variability than dose 100.

Here is the analysis of variance for this data:

```
> Caffeine.aov <- aov(Taps ~ Dose + Error(Students), CRDCaff.dat)
> summary(Caffeine.aov)

Error: Students
          Df  Sum Sq Mean Sq F value   Pr(>F)
Dose       2  61.400  30.700  6.1812 0.006163
Residuals 27 134.100   4.967
> model.tables(Caffeine.aov, type="means")
Tables of means
Grand mean

246.5

 Dose
Dose
    0   100   200
244.8 246.4 248.3
```

The analysis of this data is as follows:

Step 1: Set up hypotheses

H_0: $\alpha_1 = \alpha_2 = ... = \alpha_t = \mu$ (or $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$)

H_1: not all population Dose means are equal (or $\boldsymbol{\psi}_D = \mathbf{X}_D\boldsymbol{\alpha}$)

Set $\alpha = 0.05$.

Step 2: Calculate test statistic

The analysis of variance table for the example is:

Source	df	SSq	MSq	F	Prob
Students	29	195.50			
Doses	2	61.40	30.70	6.18	0.0062
Residual	27	134.10	4.97		

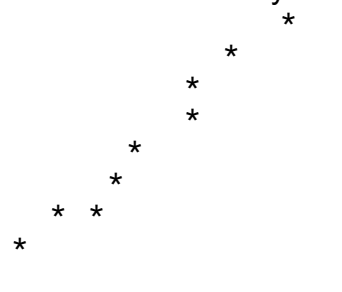
Step 3: Decide between hypotheses

$P(F_{2,27} \geq 6.18) = p = 0.006 < \alpha = 0.05$. The evidence suggests that there is a dose difference and that the expectation model that best describes the data is $\boldsymbol{\psi}_D = \mathbf{X}_D\boldsymbol{\alpha}$.

To look at how well the data conform to the assumed model one examines the residuals, that is the vector $\mathbf{e}_T$. We shall use the *Residuals-versus-fitted-values plot* and the *Normal Probability plot*.

In using these plots one should take into account that the analysis of variance is robust to variance heterogeneity, provided the treatments are equally replicated, and to moderate departures from normality. So we would only take action in respect of apparent variance heterogeneity when the treatments are not equally replicated or, perhaps when there is an extremely large difference in variances. Similarly, there would have to be evidence of marked nonnormality before remedial action would be required. The most common form of discrepancy revealed by these plots is that there is an unusually large or small residual. The cause of such extreme values requires investigation. It may be a mistake in recording or entering the data or a catastrophe affecting that unit can be identified. In the absence of an explanation, serious consideration of the possibility that the result is valid and that it is the result of some unanticipated, but important effect.

The *Normal Probability plot* should show a broadly straight-line trend.



The *Residuals-versus-fitted-values plot* involves plotting $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\theta}$ against $\mathbf{X}\hat{\theta}$. Generally, the points on the scatter diagram should be spread across plot evenly, that is, displaying no particular pattern.

no particular pattern

Patterns such as shown below indicate problems:

systematic trend in residuals

variance increases as level increases

variance peaks in middle

Actually, for the CRD, the *Residuals-versus-fitted-values plot* will have a vertical scatter of points for each treatment located along the X-axis according to the treatment mean. Each cloud will be centred on zero and should be of the same width. Unacceptable patterns in this plot indicates possible violation of the first three assumptions. However, for violations to be detected they must affect the treatments differently — if they are not treatment related they may not show up in the plot.

The R functions used in producing these plots are:

`resid.errors`: extract the residuals from an `aov` object when Error function used
`fitted.errors`: extract the fitted values from an `aov` object when Error function used
`plot`: to plot the fitted values against the residuals
`qqnorm`: to plot the residuals against the normal quantiles
`qqline`: to add a line to the plot produced by `qqnorm`.

The first two functions are nonstandard functions from the DAE package.

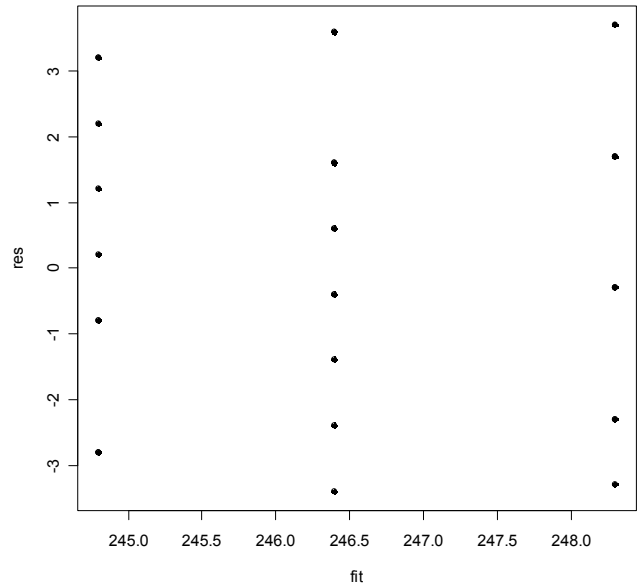
Example III.2 Caffeine effects on students (continued)

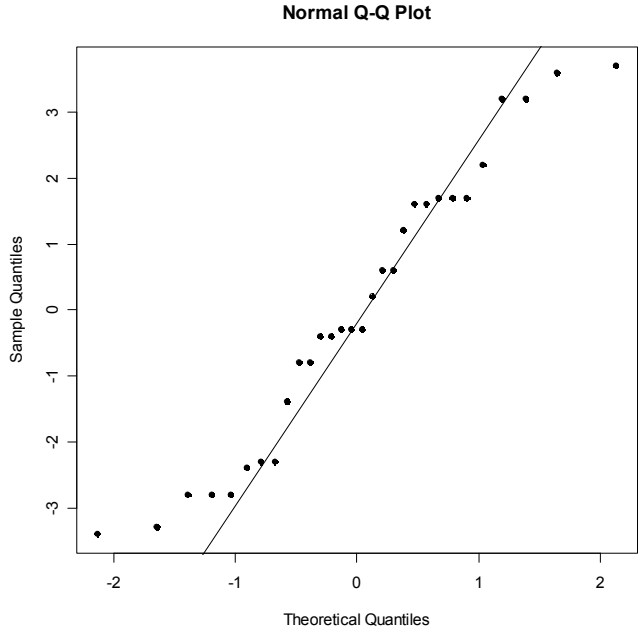
For this example, a violation of the assumption would occur if all the students were in the same room and the presence of other students caused anxiety to just the students that had no caffeine. That is, the response of the students is not independent. It may be that the inhibition of this group resulted in less variation in their response which would be manifest in the plot. Another situation that would lead to an unacceptable pattern in the plot is if the effect becomes more variable as the level of the response variable increases. For example, caffeine increases the tapping but at higher levels the variability of increase from student to student is greater. That is there is a lack of unit-treatment additivity.

For this example, the commands to obtain the two plots and the output are as follows. Note the use of `data.frame` to produce a printed list of the original data with the residuals and fitted values. This function forms a data frame from the listed factors and vectors which then printed by default.

```
> res <- resid.errors(Caffeine.aov)
> fit <- fitted.errors(Caffeine.aov)
> data.frame(Students,Dose,Taps,res,fit)
   Students Dose Taps  res   fit
1         1    0  242 -2.8 244.8
2         2  100  248  1.6 246.4
3         3  200  246 -2.3 248.3
4         4    0  245  0.2 244.8
5         5  100  246 -0.4 246.4
6         6  200  248 -0.3 248.3
7         7    0  244 -0.8 244.8
8         8  100  245 -1.4 246.4
9         9  200  250  1.7 248.3
10       10    0  248  3.2 244.8
11       11  100  247  0.6 246.4
12       12  200  252  3.7 248.3
13       13    0  247  2.2 244.8
14       14  100  248  1.6 246.4
15       15  200  248 -0.3 248.3
16       16    0  248  3.2 244.8
17       17  100  250  3.6 246.4
18       18  200  250  1.7 248.3
19       19    0  242 -2.8 244.8
20       20  100  247  0.6 246.4
21       21  200  246 -2.3 248.3
22       22    0  244 -0.8 244.8
23       23  100  246 -0.4 246.4
24       24  200  248 -0.3 248.3
25       25    0  246  1.2 244.8
26       26  100  243 -3.4 246.4
27       27  200  245 -3.3 248.3
28       28    0  242 -2.8 244.8
29       29  100  244 -2.4 246.4
30       30  200  250  1.7 248.3
> plot(fit, res, pch = 16)
> qqnorm(res, pch = 16)
> qqline(res)
```

Here are the plots for the example:





The *Residuals-versus-fitted-values plot* appears to be fine. The *Normal Probability plot* is displaying some evidence of curvature at both ends. This indicates that the data is perhaps heavier in the tails and flatter in the peak than one would expect for a normal distribution. Given that normality is not a critical assumption and that it involves only a few observations, we will use the analysis that we have performed.

III.E Treatment differences

So far all that our analysis has accomplished is that we have decided whether or not there appears to be a difference between the population treatment means. Of greater interest to the researcher is how the treatment means differ. We look at addressing that question now.

There are two alternatives available: **Multiple comparisons procedures** and **Fitting submodels**. The first technique is used when the treatment factors are all qualitative so that it is appropriate to test for differences between treatments. When one (or more) of the treatment factors is quantitative the fitting of smooth curves to the trend in the means is likely to lead to a more appropriate and concise description of the effects of the factors. Often, for reasons explained in chapter II, a low order polynomial will provide an adequate description of the trend.

It is important to note that multiple comparison procedures should *not* be used when the test for treatment differences is not significant. On the other hand, submodels should be fitted *irrespective* of whether the overall test for treatment differences is significant. The difference in usage has to do with one being concerned with mean differences and the other with deciding between models.

a) Multiple comparisons procedures for comparing all treatments

Multiple **c**omparisons for **a**ll treatments are known as MCA procedures and there are many available. We will use Tukey's HSD procedure.

In general MCA procedures divide into those that are based on family-wise error rates and those that are based on comparison-wise error rates. For those based on family-wise error rates, the Type I error rate is specified and controlled for over all comparisons, often at 0.05. Those that control for the comparison-wise error rate do so for each comparison. The problem with the latter procedures is that the probability of an incorrect conclusion gets very high as the number of comparisons increases. This is, of course, related to the number of means. It can be shown that, if a procedure with a comparison-wise error rate of 0.05 is used, the family-wise error rate is as specified in the following table:

No. means	No. Compared	Family-wise error rate
5	10	0.40
10	45	0.90
15	105	0.995

Clearly, the probability of drawing an erroneous conclusion with even 5 means is high and with 10 means is getting very high. It is for this reason that we recommend the use of an MCA procedure based on family-wise error rates.

Tukey's HSD procedure

Tukey's Honestly Significant Difference procedure determines for every pair of means whether they are significantly different and is based on a family-wise error rate. It is basically a technique for equal numbers of observations for each mean. However, a modification that provides an approximation will be provided for when the numbers of observations in the different means are not equal.

Each application of the procedure is based on the hypotheses:

H_o: $\alpha_A = \alpha_B$
H_1: $\alpha_A \neq \alpha_B$

One calculates the statistic

$$w(\alpha\%) = \frac{q_{t,\upsilon,\alpha}}{\sqrt{2}} s_{\bar{x}_d}$$

where $q_{t,\nu,\alpha}$ is the studentized range with $t =$ number of means, $\nu =$ Residual degrees of freedom and $\alpha =$ significance level,

$s_{\bar{x}_d}$ is the standard error of the difference

$$= \sqrt{\text{Residual MSq} \left(1/r_A + 1/r_B \right)}$$

r_A, r_B are the number of replicates for each of a pair of means being compared.

Note: Note that strictly speaking r_A and r_B should be equal. Using the values of r_A and r_B when they are not is sometimes called the Tukey-Kramer procedure. When they are not equal, $w(\alpha\%)$ will depend on which means are being compared.

If the treatments are all equally replicated, the formula for $s_{\bar{x}_d}$ reduces to

$$s_{\bar{x}_d} = \sqrt{\text{Residual MSq} \left(2/r \right)}$$

where r is the number of replicates in a treatment mean.

The R function `aov` provides the Residual MSq and df and `model.tables` is used to produce the tables of means. To obtain $q_{t,\nu,\alpha}$, use the R function `qtukey` as follows:

```
q <- qtukey(1 - α, t, ν)
```

Example III.2 Caffeine effects on students (continued)

In this experiment we have already concluded that the evidence suggests that there is a dose difference. But which doses are different? While we will use Tukey's HSD procedure to investigate this, this is done for illustration only as it is more appropriate to fit submodels given that Dose is a quantitative factor.

The output from the functions to get the means, standard errors and studentized range is as follows:

```
> model.tables(Caffeine.aov, type="means")
Tables of means
Grand mean

246.5

 Dose
Dose
    0   100   200
244.8 246.4 248.3
> q <- qtukey(0.95, 3, 27)
> q
[1] 3.506426
```

Note that to get the studentized range we had $\alpha = 0.05$, $t = 3$ and $\nu = 27$ and from the output $q_{3,27,0.05} = 3.506$. Hence,

$$w(5\%) = \frac{3.506}{\sqrt{2}} \times \sqrt{\frac{4.967 \times 2}{10}} = \frac{3.506}{\sqrt{2}} \times 0.997 = 2.47$$

Next we examine all pairs of treatment mean differences. This most easily accomplished in a table with both rows and columns corresponding to the Treatments — the rows and columns are ordered in ascending order of the values for the means. The body of this table should contain the differences between row means and column means — in particular, the row mean minus the column mean for entries below the diagonal. We can then determine which pairs of means are different by comparing the differences with the $w(\alpha\%)$

Differences between all pairs of Dose means

Dose	Mean	0 244.8	100 246.4	200 248.3
0	244.8			
100	246.4	1.6		
200	248.3	3.5	1.9	
	w(5%)		2.47	

Any two means different by more than 2.47 are significantly different. Our conclusion is that the mean for 0 and 200 are different but that for 100 is somewhat intermediate.

b) Fitting submodels

So far we have investigated differences between means or linear combinations of means. However, when the levels of a factor are quantitative, it is often better to examine the relationship between the response and the levels of the factor. This is commonly done using polynomials. Now, of course a polynomial of degree $t-1$ will fit

exactly t points. So that, in our example, a quadratic will fit exactly the three means. Thus, we are unable to consider polynomials of higher order than 2. In practice, one would usually does not want to consider polynomials of order greater than 2; however, more than 3 points may be desirable so that deviations from the fitted curve or lack of fit can be tested.

Polynomial models

Thus, to investigate polynomial models up to order 2, the following models for the expectation are investigated:

$$\begin{array}{ll} E[Y_i] = \mu & E[\mathbf{Y}] = \mathbf{X}_G\mu \\ E[Y_i] = \mu + \gamma_1 x_k & E[\mathbf{Y}] = \mathbf{X}_1\boldsymbol{\gamma}_1 \text{ where } \boldsymbol{\gamma}_1' = [\mu \quad \gamma_1] \\ E[Y_i] = \mu + \gamma_1 x_k + \gamma_2 x_k^2 & E[\mathbf{Y}] = \mathbf{X}_2\boldsymbol{\gamma}_2 \text{ where } \boldsymbol{\gamma}_2' = [\mu \quad \gamma_1 \quad \gamma_2] \\ E[Y_i] = \alpha_k & E[\mathbf{Y}] = \mathbf{X}_T\boldsymbol{\alpha} \text{ where } \boldsymbol{\alpha}' = [\alpha_0 \quad \alpha_{100} \quad \alpha_{200}] \end{array} \quad \text{or}$$

where x_k is the value of the kth level of the treatment factor,
μ is the intercept of the fitted equation and
γ_1 is the slope of the fitted equation and
γ_2 is the quadratic coefficient of the fitted equation.

To fit such models involves doing least squares fitting where, for the polynomial models, the $\mathbf{X}_1$ and $\mathbf{X}_2$ matrices are made up of columns that consist of the values of the levels of the factor and powers of those levels — not the indicator variables of before.

Note that each model in the sequence of models given above is marginal to all models before it as $\mathcal{C}(\mathbf{X}_G) \subseteq \mathcal{C}(\mathbf{X}_1) \subseteq \mathcal{C}(\mathbf{X}_2) \subseteq \mathcal{C}(\mathbf{X}_T)$. That is, the columns of each $\mathbf{X}$ matrix in the above list are a linear combination of those of any of the $\mathbf{X}$ matrices to its right in the list. Marginality is not a symmetric relationship in that, if a model is marginal to second model, the second model is not necessarily marginal to the first; for example, $E[\mathbf{Y}] = \mathbf{X}_1\boldsymbol{\gamma}_1$ is marginal to $E[\mathbf{Y}] = \mathbf{X}_2\boldsymbol{\gamma}_2$ but not vice-a-versa, except when $t = 2$.

Example III.2 Caffeine effects on students (continued)

The $\mathbf{X}$ matrices for the example are:

$$\mathbf{X}_G = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{X}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 100 \\ 1 & 100 \\ \vdots & \vdots \\ 1 & 100 \\ 1 & 200 \\ 1 & 200 \\ \vdots & \vdots \\ 1 & 200 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 100 & 10000 \\ 1 & 100 & 10000 \\ \vdots & \vdots & \vdots \\ 1 & 100 & 10000 \\ 1 & 200 & 40000 \\ 1 & 200 & 40000 \\ \vdots & \vdots & \vdots \\ 1 & 200 & 40000 \end{bmatrix}, \quad \mathbf{X}_D = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix}$$

Note that for the example $\mathcal{C}(\mathbf{X}_G) \subset \mathcal{C}(\mathbf{X}_1) \subset \mathcal{C}(\mathbf{X}_2) = \mathcal{C}(\mathbf{X}_D)$. That is, the columns of each **X** matrix in the above list are a linear combination of those of any of the **X** matrices to its right in the list. In this case $\mathcal{C}(\mathbf{X}_2) = \mathcal{C}(\mathbf{X}_D)$ as the three columns of one matrix can be written as 3 linearly independent combinations of the columns of the other matrix; both matrices span the same space. Hence the models $E[\mathbf{Y}] = \mathbf{X}_2\boldsymbol{\gamma}_2$ and $E[\mathbf{Y}] = \mathbf{X}_D\boldsymbol{\alpha}$ are marginal to each other and are equivalent. However, while the fitted values are the same, the estimates and interpretation of the parameters are different. The parameters corresponding to $\mathbf{X}_2$ are interpreted as the intercept, slope and curvature coefficient; those corresponding to $\mathbf{X}_D$ are interpreted as the expected (mean) value for that dose treatment.

Also, in spite of being marginal, the estimators of the same parameter differ depending on the model that has been fitted. For example, $\hat{\mu} = \bar{Y}$ for the model $E[\mathbf{Y}] = \mathbf{X}_G\mu$, but $\hat{\mu} = \bar{Y} - \hat{\gamma}_1\bar{x}$ for $E[\mathbf{Y}] = \mathbf{X}_1\boldsymbol{\gamma}_1$. Similar comments apply to $\hat{\gamma}_1$. The models might be marginal, but they are not orthogonal.

Hypothesis test incorporating submodels

Once again the test statistics on which the hypothesis tests are based can then be conveniently computed in an analysis of variance table. The hypothesis tests are conducted as follows:

Step 1: Set up hypotheses

a) H_0: $\alpha_k - \mu - \gamma_1 x_k - \gamma_2 x_k^2 = 0$ for all *k* (Differences between fitted models or deviations from quadratic are zero)
H_1: $\alpha_k - \mu - \gamma_1 x_k - \gamma_2 x_k^2 \neq 0$ for all *k*

b) H_0: $\gamma_2 = 0$
H_1: $\gamma_2 \neq 0$

c) H_0: $\gamma_1 = 0$
H_1: $\gamma_1 \neq 0$

Set α.

Step 2: Calculate test statistics

The analysis of variance table for a CRD is:

Source	df	SSq	MSq		F	p
Units	$n-1$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{U}}\mathbf{Y}$				
Treatments	$t-1$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T}}\mathbf{Y}$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T}}\mathbf{Y}/(t-1)$			
Linear	1	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T_L}}\mathbf{Y}$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T_L}}\mathbf{Y}$	$\left(=s^2_{\mathrm{T_L}}\right)$	$s^2_{\mathrm{T_L}}/s^2_{\mathrm{U_{Res}}}$	$p_{\mathrm{T_L}}$
Quadratic	1	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T_Q}}\mathbf{Y}$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T_Q}}\mathbf{Y}$	$\left(=s^2_{\mathrm{T_Q}}\right)$	$s^2_{\mathrm{T_Q}}/s^2_{\mathrm{U_{Res}}}$	$p_{\mathrm{T_Q}}$
Deviations	$t-3$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T_{Dev}}}\mathbf{Y}$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T_{Dev}}}\mathbf{Y}/(t-3)$	$\left(=s^2_{\mathrm{T_{Dev}}}\right)$	$s^2_{\mathrm{T_{Dev}}}/s^2_{\mathrm{U_{Res}}}$	$p_{\mathrm{T_{Dev}}}$
Residual	$n-t$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{U_{Res}}}\mathbf{Y}$	$\mathbf{Y}'\mathbf{Q}_{\mathrm{U_{Res}}}\mathbf{Y}/(n-t)$	$\left(=s^2_{\mathrm{U_{Res}}}\right)$		

Note that $\mathbf{Q}_{\mathrm{T_L}}$, $\mathbf{Q}_{\mathrm{T_Q}}$ and $\mathbf{Q}_{\mathrm{T_{Dev}}}$ are not simple linear functions of **M** (mean operator) matrices, whereas the other **Q** matrices are. The test statistics correspond to the hypothesis pairs given in Step 1.

Step 3: Decide between hypotheses

We begin with the first hypothesis pair and determine its significance and continue down the sequence until a significant result is obtained.

A significant Deviations F $\left(p_{\mathrm{T_{Dev}}} \le \alpha\right)$ indicates that the linear and quadratic terms provide an inadequate description of the trend in the treatment means and so a model based on them would be unsatisfactory. There is no point in continuing to test if this is the case.

If the Deviations F is not significant $\left(p_{\mathrm{T_{Dev}}} > \alpha\right)$, then a significant Quadratic F $\left(p_{\mathrm{T_Q}} \le \alpha\right)$ indicates that a second-degree polynomial is required to adequately describe the trend. As a linear coefficient is necessarily incorporated in a second-degree polynomial there is no point to further testing in this case.

If both the Deviations and Quadratic Fs are not significant $\left(p_{\mathrm{T_{Dev}}} > \alpha \;\&\; p_{\mathrm{T_Q}} > \alpha\right)$, then a significant Linear F $\left(p_{\mathrm{T_L}} \le \alpha\right)$ indicates a linear relationship describes the trend in the treatment means.

If the Deviations are significant, then two options are: a) increase the degree of the polynomial (often not desirable) or b) use multiple comparisons to investigate the differences between the means.

Fitting polynomials in R involves realizing that the default contrasts for `ordered` (factor) objects are polynomial contrasts, assuming the levels are equally spaced: linear transformations of all columns of **X**, except the first, such that each column a) is orthogonal to all other columns in **X** and b) has sums of squares equal to one. They facilitate the computations. So, if a factor is quantitative, set it up as an ordered object from the start. However, if you have not made it an ordered object, then redefine the `factor` to be an `ordered`.

Example III.2 Caffeine effects on students (continued)

The R output for fitting the polynomial submodels is as follows:

```
> # fit polynomials
> #
> t <- 3
> Dose.lev <- c(0,100,200)
> CRDCaff.dat$Dose <- ordered(CRDCaff.dat$Dose, levels=Dose.lev)
> contrasts(CRDCaff.dat$Dose) <- contr.poly(t,
scores=Dose.lev)
> Caffeine.aov <- aov(Taps ~ Dose + Error(Students), CRDCaff.dat)
> contrasts(CRDCaff.dat$Dose)
                .L          .Q
0   -7.071068e-01  0.4082483
100 -1.190326e-16 -0.8164966
200  7.071068e-01  0.4082483
> summary(Caffeine.aov, split = list(Dose = list(L = 1, Q = 2)))

Error: Students
          Df  Sum Sq Mean Sq F value   Pr(>F)
Dose       2  61.400  30.700  6.1812 0.006163
  Dose: L  1  61.250  61.250 12.3322 0.001585
  Dose: Q  1   0.150   0.150  0.0302 0.863331
Residuals 27 134.100   4.967
```

Notes:
1. Dose levels have been stored in a numeric vector to save repeatedly entering them;
2. The expression `contrasts(CRDCaff.dat$Dose) <- contr.poly t, scores=Dose.lev)` is redundant for `ordered`s with equally-spaced levels, but is required when the levels are unequally spaced.
3. The computed contrasts are coded values of dose and dose squared, respectively, with the properties that the coefficients of each contrast sum to 0, their sum of squares is 1, and the sum of the cross-products is 0. Other information associated with the contrasts is printed out as well — it can be ignored.
4. Coefficients can be obtained using the `coef` function, but these are not suitable for obtaining the fitted values.

The hypothesis test for the example is given below. Note that in this case there are only three treatments so that a quadratic will follow the trend in the treatment means exactly. This is reflected in the fact that $\mathcal{C}(\mathbf{X}_2) = \mathcal{C}(\mathbf{X}_D)$. Thus, the Deviations line is redundant in this example. When required, the Deviations line is computed by assigning extra powers to a single line named, say, Dev. For example, if there were 7 treatments then the following function will fit a cubic and compute the Deviations line:

```
> summary(Experiment.aov, split = list(Treatment =
                    list(L = 1, Q = 2, C = 3, Dev = 4:6)))
```

Step 1: Set up hypotheses

a) H_0: $\gamma_2 = 0$
   H_1: $\gamma_2 \neq 0$

b) H_0: $\gamma_1 = 0$
   H_1: $\gamma_1 \neq 0$

Set $\alpha = 0.05$.

Step 2: Calculate test statistics

Source	df	SSq	MSq	F	Prob
Students	29	195.50			
Doses	2	61.40	30.70	6.18	0.006
Linear	1	61.25	61.25	12.33	0.002
Quadratic	1	0.15	0.15	0.03	0.863
Residual	27	134.10	4.97		

Step 3: Decide between hypotheses

The Quadratic source has a probability of 0.863 > 0.05 and so the null hypothesis is not rejected in this case. The linear source has a probability of 0.002 < 0.05 and so the null hypothesis is rejected in this case. It is clear that the quadratic term is not significant but that the linear term is highly significant so that the appropriate model for the expectation is the linear model $\boldsymbol{\psi} = \mathbf{X}_1\gamma_1$ where $\gamma_1' = [\mu \quad \gamma_1]$.

Fitted equation

The fitted equation is obtained by putting the values of the levels into a numeric vector and using the `lm` function to fit a polynomial of the order indicated by the hypothesis test as necessary for describing the pattern in the treatment means.

For the example, linear equation was adequate and so the analysis is redone with 1 for the order of the polynomial.

```
> #
> #get fitted equation
> #
> D <- as.vector(Dose)
> D <- as.numeric(D)
> Caffeine.lm <- lm(Taps ~ D)
> coef(Caffeine.lm)
(Intercept)           D
   244.7500      0.0175
```

The fitted equation is $Y = 244.75 + 0.0175X$ where X is the number of taps.

The slope of this equation is 0.0175. That is, taps increase 0.0175 x 100 = 1.75 with each 100 mg of caffeine. This conclusion seems a more satisfactory summary of the results than that the response at 200 is significantly greater than at 0 with 100 being intermediate. It is the reason that we prefer fitting submodels over multiple comparison procedures when the treatment factor is quantitative.

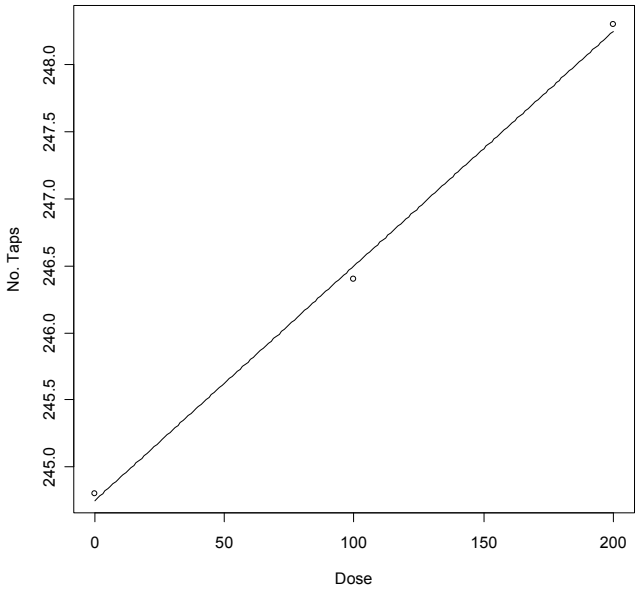
The command to fit a quadratic would be:

```
D2 <- D*D
Caffeine.lm <- lm(Taps ~ D + D2)
```

Next we obtain a plot of the mean number of taps versus the Dose and the fitted straight line. First a data.frame containing the means and the levels values of the factor is set up. Then the `plot` function is used to plot the means and the `abline` function to add the fitted line.

```
> #
> # plot means and fitted line
> #
> Caffeine.tab <- model.tables(Caffeine.aov, type="means")
> Dose.Mean <- Caffeine.tab$tables$Dose
> plot(x=Dose.lev, y=Dose.Mean, xlab="Dose", ylab="No. Taps")
> Caffeine.coef <- coef(Caffeine.lm)
> dosex <- seq(0, 200, 1)
> Dose.Fit <- Caffeine.coef[[1]] + Caffeine.coef[[2]]*dosex
> lines(x=dosex, y=Dose.Fit, type="l")
```

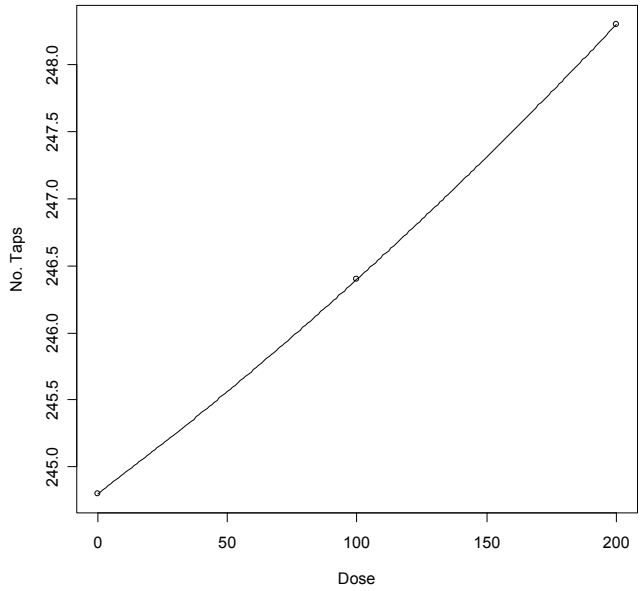
The plot produced is as follows:



If a quadratic had been appropriate (it is not), it could be fitted and plotted by modifying the expressions for the linear case to adding a quadratic term to the `lm` expression and to the calculation of fits. The modified expressions for the example are as follows:

```
> #doing quadratic regression
> D2 <- D*D
> Caffeine.lm <- lm(Taps ~ D + D2)
> #plot means and fitted quadratic
> Caffeine.tab <- model.tables(Caffeine.aov, type="means")
> Dose.Mean <- Caffeine.tab$tables$Dose
> plot(x=Dose.lev, y=Dose.Mean, xlab="Dose", ylab="No. Taps")
> Caffeine.coef <- coef(Caffeine.lm)
> dosex <- seq(0, 200, 1)
> Dose.Fit <- Caffeine.coef[[1]] + Caffeine.coef[[2]]*dosex +
+                                  Caffeine.coef[[3]]*dosex*dosex
> lines(x=dosex, y=Dose.Fit, type="l")
```

The plot produced is as follows:



While the curve does follow the points more closely, our analysis tells us that this slight curvature is just random variation (this is indicated by the nonsignificant quadratic term) and so not likely to recur in a repeat of the experiment. The linear equation and plot should be used in reporting the results of the experiment.

c) Comparison of treatment parametrizations

One point that arises from the analyses that have been performed so far is that the reduction in sums of squares for treatments can be obtained in several ways. We have investigated two for the analysis of a CRD: fitting indicator variables and polynomial contrasts. These amount to different bases for the treatment space with the unique elements of the **X** matrices whose columns form the basis being:

Indicator variables	Polynomials
$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{bmatrix}$

They lead to different parameter estimates with different interpretations and different partitions of the treatment sums of squares, but the total treatment sums of squares and fitted values for treatments remain the same while the contrasts span the treatment space. That is, in the example,

$$SSq_{\mathrm{T}} = SSq_{\mathrm{L}} + SSq_{\mathrm{Q}}$$

III.F Summary

In this chapter we have:

- described how to design an experiment using a completely randomized design;
- formulated a linear model using indicator variables to describe the results from a completely randomized design; given the estimators of the parameters in the linear model, the expected values and the errors as functions of **M** or mean operator matrices;
- outlined an hypothesis test for choosing between two expectation models, termed the minimal and maximal models, using the ANOVA hypothesis test procedure outlined in chapter I;
  - the concept of a marginal model was introduced to describe a model that is a submodel of another model;
  - the partition of the total sums of squares into treatment and residual sums of squares was given; the sums of squares were expressed as the sums of squares of the elements of vectors and as quadratic forms where the matrices of the quadratic forms, **Q** matrices, are symmetric idempotents;
  - the expected mean squares under the minimal and maximal expectation models are used to justify the choice of F test statistic;
- shown how to obtain a layout and the analysis of variance in R;
- discussed procedures for checking the adequacy of the proposed models;
- outlined two methods for investigating in more detail the differences between the treatments: multiple comparisons procedures for investigating all possible pairs of differences when the factor is qualitative; polynomial models for describing the trend in the means when the factor is quantitative. The latter involves extending the analysis of variance to partition the treatment sums of squares into single-degree-of-freedom sources, one for each coefficient in the polynomial model. In addition a source for deviations (from the fitted polynomial) may also be included.

III.G Exercises

III.1 Show that $\mathbf{Q}_{\mathrm{U}}$, $\mathbf{Q}_{\mathrm{T}}$ and $\mathbf{Q}_{\mathrm{U}_{\mathrm{Res}}}$ are symmetric and idempotent where $\mathbf{Q}_{\mathrm{U}} = \mathbf{M}_{\mathrm{U}} - \mathbf{M}_{\mathrm{G}}$, $\mathbf{Q}_{\mathrm{T}} = \mathbf{M}_{\mathrm{T}} - \mathbf{M}_{\mathrm{G}}$ and $\mathbf{Q}_{\mathrm{U}_{\mathrm{Res}}} = \mathbf{M}_{\mathrm{U}} - \mathbf{M}_{\mathrm{T}}$. You are given that $\mathbf{M}_{\mathrm{U}} = \mathbf{I}_n$, $\mathbf{M}_{\mathrm{T}}$ and $\mathbf{M}_{\mathrm{G}}$ are symmetric and idempotent and that $\mathbf{M}_{\mathrm{T}}\mathbf{M}_{\mathrm{G}} = \mathbf{M}_{\mathrm{G}}\mathbf{M}_{\mathrm{T}} = \mathbf{M}_{\mathrm{G}}$.

III.2 Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} 2 & 4 \\ 1 & 6 \end{bmatrix}$$

Find $\mathbf{y}'\mathbf{A}\mathbf{y}$. Why do you think $\mathbf{y}'\mathbf{A}\mathbf{y}$ is called a quadratic form?

III.3 An investigation was conducted to examine differences between 3 brands of tyre in their braking distances (ft.) from a speed of 30 miles per hour. Altogether 9 tests were conducted with the particular brand tested at each test being chosen at random. The results are as follows:

Brand		
A	B	C
28	27	27
26	25	32
30	26	31

a) Give the matrix $\mathbf{X}_{\mathrm{T}}$ in terms of both the individual elements and a partitioned matrix of **1** and **0** column vectors. Also give an expression for $\mathbf{M}_{\mathrm{T}}$ in terms of **J** and **0** matrices.

b) Compute the column vectors $\bar{\mathbf{g}} = \mathbf{M}_{\mathrm{G}}\mathbf{y}$, $\mathbf{e}_{\mathrm{G}} = \mathbf{Q}_{\mathrm{G}}\mathbf{y}$, $\bar{\mathbf{t}} = \mathbf{M}_{\mathrm{T}}\mathbf{y}$, $\mathbf{t}_e = \mathbf{Q}_{\mathrm{T}}\mathbf{y}$ and $\mathbf{e}_{\mathrm{T}} = \mathbf{Q}_{\mathrm{U}_{\mathrm{Res}}}\mathbf{y}$. From these compute the sums of squares for the analysis of variance.

c) Construct the analysis of variance table. Use the R function `pf` to obtain the p-value — the call should be of the form `1 - pf(F,` ν_1`,` ν_2`)`.

III.4 Use R to produce a randomized layout for an experiment involving 21 runs to which 7 treatments are to be allocated so that each is replicated 3 times. Use the `seed` 911 in producing the layout. The effect of this is that everyone should get the same answer — in practice, one should randomly choose the seed.

III.5 An advertising agency conducts an experiment to assess the effectiveness of various formats of TV commercials. Forty regular viewers are randomly assigned to one of four formats of a TV commercial for a wine. Each viewer is interviewed and a score measuring impact recorded. The results are as follows:

FORMAT			
A	B	C	D
20	28	33	33
23	27	34	29
21	22	25	31
23	28	26	29
26	23	27	27
24	29	33	25
26	27	25	26
23	25	32	26
20	28	25	33
24	21	34	32

The response variable Impact has been saved in *CRDForm.dat.rda* and is available from the Statistical Modelling resources web site. It has been entered from the above table by columns. Add the 40-level factor Viewer and the factor Format with levels A–D to this data frame. This can be done by assigning factors to `CRDForm.dat$Viewer` and `CRDForm.dat$Format`, using the `rep` function for Format to save on typing. However, you will have to be careful about your use of `each` and `times` to ensure that the values that you generate line up with the values of Impact.

Obtain boxplots for each format as an initial exploration of the data. (Note: Appendix C, *Analysis of designed experiments in R*, contains a summary of the instructions for analysing a completely randomized design.)

Use the R function `aov` to perform an analysis of variance on this data to determine if there is any measurable effect of format on the impact score.

Use R to obtain a residuals-versus-fitted-values plot and a normal probability plot as a check on the assumptions underlying the analysis. (Note that you will need to use the functions `resid.error` and `fitted.error` from the `dae` package. For instructions on the installation and use of `dae` see the Statistical Modelling resources web site.

Also, use Tukey's procedure to determine exactly which Formats differed.

III.6 An experiment was conducted to test the effect of adding a small percentage of coal dust to the sand used for making concrete. Twenty batches of concrete were mixed under as identical conditions as possible and one of the five chosen additions of coal selected at random for each batch so that each addition occurred four times. From each batch a cylinder of concrete was made and tested for the breaking strength in pounds per square inch (lb/in^2).

PERCENTAGE COAL				
0	0.05	0.1	0.5	1.0
1690	1550	1625	1725	1530
1580	1445	1450	1550	1545
1745	1645	1510	1430	1565
1685	1545	1527	1445	1520

This response variable Strength has been saved in *CRDConcr.dat.rda* and is available from the Statistical Modelling resources web site. As in exercise III.5, you will need to add the factor Batch and Cola.Percent (or some similar names). Obtain boxplots for each format as an initial exploration of the data.

Use R to perform an analysis of variance on this data, including the checking of assumptions. Investigate the fitting of a quadratic response to the treatment means and obtain a plot of the means together with the fitted quadratic. (Hint: the `ylim` argument of `plot` will be useful in changing the range on the y-axis.)