Chapter 6

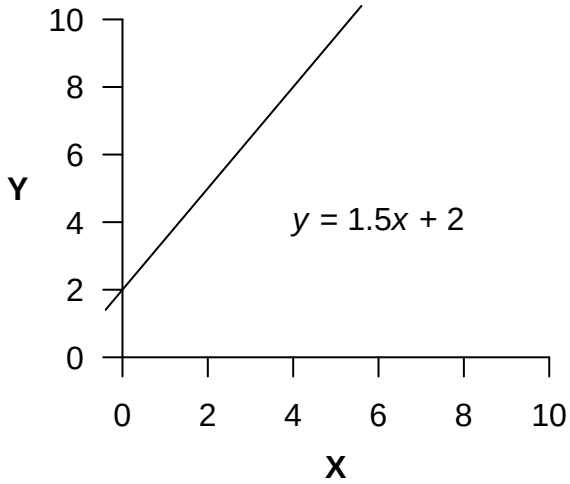
Introduction to Linear models

A **statistical model** is an expression that attempts to explain patterns in the observed values of a response variable by relating the response variable to a set of predictor variables and parameters. Consider the following familiar statistical model:

$y = mx + c$

This simple statistical model relates a response variable (y) to a single predictor variable (x) as a straight line according to the values of two constant parameters:

- m – the degree to which y changes per unit of change in x (gradient of line)
- c – the value of y when $x = 0$ (y-intercept)



The above statistical model represents a perfect fit, that is, 100% of the change (variation) in y is explained by a change in x. Rarely, however, would this be the case when modeling biological variables. In complex biological systems, variables are typically the result of many influential and interacting factors and therefore simple models usually fail to fully explain a response variable. Consequently, the statistical model also has an *error* component that represents the portion of the response variable that the model fails to explain. Hence, statistical models are of the form:

response variable = model + error

where the model component comprises of one or more categorical and/or continuous predictor variable(s) and their parameter(s) that together represent the effect of the predictors variable(s) on the mean the response variable. A parameter and its associated predictor variable(s) are referred to as a model *term*.

A statistical model is fitted to observed data so as to estimate the model parameters and test hypotheses about these parameters (coefficients).

6.1 Linear models

Linear models are those statistical models in which a series of parameters are arranged as a linear combination. That is, within the model, no parameter appears as either a multiplier, divisor or exponent to any other parameter. Importantly, the term 'linear' in this context does not pertain to the nature of the relationship between the response variable and the predictor variable(s), and thus linear

models are not restricted to 'linear' (straight-line) relationships.

An example of a very simple linear model, is the model used to investigate the linear relationship between a response variable (Y and a single continuous predictor variable, X):

$$\begin{array}{cccccccccc} y_i & = & \beta_0 & + & \beta_1 & \times & x_i & + & \varepsilon_i \\ \text{response variable} & = & \underbrace{\text{population intercept}}_{\text{intercept term}} & + & \underbrace{\text{population slope} \times \text{predictor variable}}_{\text{slope term}} & & & + & \text{error} \end{array}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{model}}$$

The above notation is typical of that used to represent the elements of a linear model. y denotes the response variable and x represents the predictor variable. The subscript (i) is used to represent a set of observations (usually from 1 to n where n is the total sample size) and thus y_i and x_i represent respectively the i^{th} observation of the Y and X variables. ε_i represents the deviation of the i^{th} observed Y from the value of Y expected by the model component. The parameters β_0 and β_1 represent population intercept and population slope (effect of X on Y per unit of x) respectively. Effect parameters are usually represented by Greek symbols[1]. The above linear model notation is therefore a condensed representation of a compilation of arithmetic relationships:

$$\begin{array}{cccccccccc} y_1 & = & \beta_0 & + & \beta_1 & \times & x_1 & + & \varepsilon_1 \\ y_2 & = & \beta_0 & + & \beta_1 & \times & x_2 & + & \varepsilon_2 \\ y_3 & = & \beta_0 & + & \beta_1 & \times & x_3 & + & \varepsilon_3 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{array}$$

the first y observation (y_1) is related to the first x observation (x_1) according to the values of the two constants (parameters β_0 and β_1) and ε_1 is the amount that the observed value of Y differs from the value expected according the model.

When there are multiple continuous predictor variables, in addition to the intercept parameter (β_0), the linear model includes a separate slope parameter for each of the predictor variables:
$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + ... + \varepsilon_i$

The model structure for linear models containing a single categorical predictor variable (known as a factor) with two or more treatment levels (groups) is similar in form to a multiple linear regression model with the overall mean (μ) replacing the y-intercept (β_0). The factor levels (groups) are represented in the model by binary (contain only of 0s and 1s, see Table 6.1) *indicator* (or 'dummy') variables and associated estimable parameters (β_1, β_2, ...).

For a data set comprising of p groups and n replicates within each group, the linear model is:
$y_{ij} = \mu + \beta_1 (dummy_1)_{ij} + \beta_2 (dummy_2)_{ij} + ... + \varepsilon_{ij}$
where i represents the treatment levels (from 1 to p) and j represents the set of replicates (from 1 to n) within the i^{th} group. Hence, y_{ij} represents the j^{th} observation of the response variable within the i^{th} group and $(dummy_1)_{ij}$ represents the dummy code for the j^{th} replicate within the i^{th} group of the first dummy variable (first treatment level).

The dummy variable for a particular treatment level contains all 0s except in the rows that correspond to observations that received that treatment level. Table 6.2 illustrates the dummy coding

[1]Typically, effect paramters associated with continuous variables are represented by β and those associated with categorical variables are represented by the symbols α, β, γ, ...

for a single factor within three levels ('G1', 'G2', 'G3') each with three replicates[2].

Table 6.1: Ficticious data set (consisting of three replicates for each of three groups:'G1','G1','G2') to illustrate the link between a) single factor dataset, and b) the indicator (dummy) variables

a)

Y	A
2	G1
3	G1
4	G1
6	G2
7	G2
8	G2
10	G3
11	G3
12	G3

b)

Y	$dummy_1$	$dummy_2$	$dummy_3$
2	1	0	0
3	1	0	0
4	1	0	0
6	0	1	0
7	0	1	0
8	0	1	0
10	0	0	1
11	0	0	1
12	0	0	1

More typically however, statistical models that include one or more factors are expressed as *effects models* in which the individual treatment levels (and their parameters) are represented by a single term (e.g. α_i) that denotes the effect of each of the levels of the factor on the overall mean. For a data set comprised of p groups and n replicates within each group, the linear effects model is:
$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
where i represents the set of treatments (from 1 to p) and j represents the set of replicates (from 1 to n) within the i^{th} group. Hence, y_{ij} represents the j^{th} observation of the response variable within the i^{th} group of the factor. μ is the overall population mean of the response variable (Y) and is equivalent to the intercept. α_i represents the effect of the i^{th} group calculated as the difference between each of the group means and the overall mean ($\alpha_i = \mu_i - \mu$).

6.2 Linear models in R

Statistical models in R are represented by a formula corresponding to the linear model (for continuous variables) or effects model (categorical variables):
`response~model`
where the tilde (`~`) defines a model formula and `model` represents a set of terms to include in the model. Terms are included in a model via their variable names and terms preceded by the `-` (negative sign) operator are explicitly excluded. The intercept term (denoted by a 1) is implicit in the model and need not be specified. Hence the following model formulae all model the effect of the variable `X` on the `Y` variable with the inclusion of the intercept:

```
Y~X
Y~1+X
Y~X+1
```

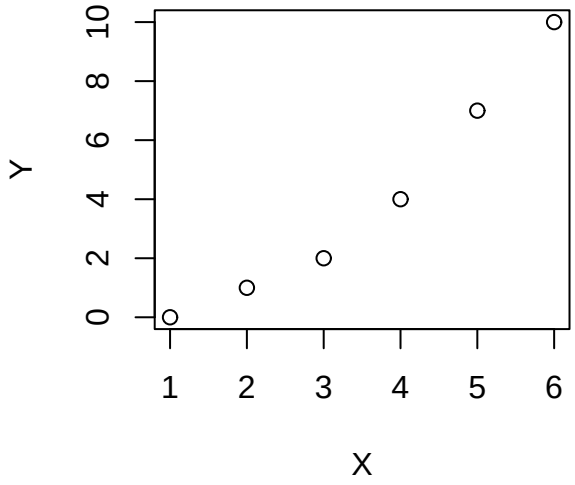
whereas the following exclude the intercept:

```
Y~-1+X
Y~X-1
```

[2]Note that linear model that this represents ($y_{ij} = \mu + \beta_1(dummy_1)_{ij} + \beta_2(dummy_2)_{ij} + \beta_3(dummy_3)_{ij} + \varepsilon_{ij}$) is over-parameterized, see section 6.3

Linear models are fitted by providing the model formula as an argument to the `lm()` *function*. To fit the simple linear regression model relating a fictitious response variable (Y) to fictitious continuous predictor variable (X):

```
> Y -c(0,1,2,4,7,10)
> X -1:6
> plot(Y~X)
```




```
> Fictitious.lm <- lm(Y~X)
```

To examine the estimated parameters (and hypothesis tests) from the fitted model, provide the name of the fitted model as an argument to the `summary()` *function*[3].

```
> summary(Fictitious.lm)

Call:
lm(formula = Y ~ X)

Residuals:
        1          2          3          4          5          6
1.000e+00  3.404e-16 -1.000e-00 -1.000e+00  6.280e-17  1.000e-00

Coefficients:
             Estimate  Std.  Error  t value  Pr(>|t|)
(Intercept)   -3.0000       0.9309   -3.223   0.03220  *
X              2.0000       0.2390    8.367   0.00112  **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 4 degrees of freedom
Multiple R-Squared: 0.9459,     Adjusted R-squared: 0.9324
F-statistic:    70 on 1 and 4 DF,  p-value: 0.001116
```

The summary output begins by specifying the nature of the call used to fit the model. Next is a summary of the residuals (differences between observed responses and expected responses for each value of the predictor variable). The estimated parameters are listed in the coefficients table. Each row of the table lists the value of an estimated parameter from the linear model along with the outcome of a hypothesis test for this parameter. The row labeled '`(Intercept)`' concerns the intercept (overall constant) and subsequent rows are labeled according to the model term that is associated with the estimated parameter. In this case, the row labeled 'X' concerns the population

[3]Actually, the `summary()` function is a generic wrapper that invokes different specific functions depending on the class of object provided as the first argument. In the `summary()` *function* invokes the `summary.lm()` *function*

slope (β_1). Finally a brief summary of the partitioning of total variation (ANOVA, see section 6.3.4) in the response variable is provided.

6.3 Estimating linear model parameters

During model fitting, parameters can be estimated using any of the estimation methods outlined in section 3.2, although ordinary least squares (OLS) and maximum likelihood (ML or REML) are most common.

The OLS approach estimates the value of one or more parameters such that they minimize the sum of squared deviations between the observed values and the parameter (typically the values predicted by the model) and will be illustrated in detail in the following sections.

Models that utilize OLS parameter estimates are referred to as 'general' linear models as they accommodate both continuous and categorical predictor variables. Broadly speaking, such models that incorporate purely continuous predictor variables are referred to as 'regression' models (see chapters 7 & 8) whereas models that purely incorporate categorical predictors are called 'ANOVA' models (see chapters 9 & **??**). ANCOVA models incorporate both categorical and continuous predictor variables (see chapter 14).

ML estimators estimates one or more population parameters such that the (log) likelihood of obtaining the observed values from such populations is maximized and are useful when there is evidence of a relationship between mean and variance. Models that utilize ML parameter estimates are referred to as 'generalized' linear models as they do not restrict the probability distribution of the response variable and residuals to normality. Rather, they accommodate any exponential probability distribution (including normal, binomial, Poisson, gamma and negative binomial), see chapter 15.

6.3.1 Simple linear regression

Consider again the simple linear regression model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
The predicted value of Y for a given value of X (denoted the symbol $\hat{y}_i$) from such a model is therefore equal to:
$\hat{y}_i = 1 \times \beta_0 + x_i \times \beta_1$

The difference between an observed value of Y and the expected value of Y for a given value of X is the residual ($\varepsilon_i = y_i - \hat{y}_i$) for that observation. Some of these differences will be positive and some will be negative. In fact, since the predicted values represent 'average' values along the length of the relationship between X and Y, the sum of these residuals must equal zero. Consequently, to get a measure of how much any model (with one of many possible combinations of parameters) deviates from the observed data, the residual values are first squared before being summed. The model that is considered to fit the observed data best (and therefore estimate the population parameters best) is the model with the smallest sums of squared residuals.

Each component of the statistical model is represented by a matrix; the *response variable* (y_i) by single column matrix with as many rows as there are observations (values of Y); the *error* (ε_i) component also by a single column matrix with as many rows as there are observations; and the *model* component ($\beta_0 + \beta_1 x_1$) by a *model matrix* that has as many rows as there are observations and as many columns as there are parameters to estimate. The first column of the model matrix is filled with 1s and represents the β_0 multiplier, and the second column is filled with the X values and is the

β_1 multiplier. Hence each column of a model matrix defines a null hypothesis test that a parameter equals zero. To illustrate OLS parameter estimation and linear model fitting, we will examine the small ficticious data set comprising of six pairs of X and Y observations.that was constructed in section 6.2.

$$y_i = 1\times\beta_0 + x_i\times\beta_1 + \varepsilon_i$$

$$\overbrace{\begin{matrix}\times\beta_0 & \times\beta_1\end{matrix}}^{\textit{model matrix}}$$

$$\begin{bmatrix}0\\1\\2\\4\\7\\10\end{bmatrix} = \begin{bmatrix}1 & 1\\1 & 2\\1 & 3\\1 & 4\\1 & 5\\1 & 6\end{bmatrix} + \begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\end{bmatrix}$$

In R, the model matrix associated with a linear model is obtained by providing the model formula as an argument to the `model.matrix()` *function*

```
> model.matrix(Y~X)
  (Intercept) X
1           1 1
2           1 2
3           1 3
4           1 4
5           1 5
6           1 6
attr(,"assign")
0 1
```

In the process of estimating the model parameters, the first column of the model matrix is multiplied by a possible estimate of the intercept (β_0) and the second column is multiplied by a possible estimate of the slope (β_1). The sum of these multiplications for each value of X yields the predicted values of Y. The final estimated parameters (β_0 & β_1) are the values that minimize the sum of square differences between observed and expected Y values. Fitting the above model[4] estimated β_0 as -3 and β_1 as 2. The following table shows the arithmetic for the final solution:

Model matrix		$\hat{y}$ (predict. y)		e_i (residual)	
	x_i	$1\times\beta_0 + x\times\beta_0 = \hat{y}$	y_i	$y_i - \hat{y}$	e_i^2
1	(observ. x)	when $\beta_0 = -3$ and $\beta_1 = 2$	(observ. y)	(obs-pred)	$(y_i - \hat{y})^2$
1	1	$1\times -3 + 1\times 2 = -1$	0	1	1
1	2	$1\times -3 + 2\times 2 = 1$	1	0	0
1	3	$1\times -3 + 3\times 2 = 3$	2	-1	1
1	4	$1\times -3 + 4\times 2 = 5$	4	-1	1
1	5	$1\times -3 + 5\times 2 = 7$	7	0	0
1	6	$1\times -3 + 6\times 2 = 9$	10	1	1
				Sum of $e^2 =$	4

No other combination of β_0 and β_1 parameter estimates will produce sum of squared deviations (e^2) as small. Hence the linear model relating y to x that fits the observed data best is estimated as: $y = -3 + 2x$

[4] Using the `lm()` *function*, see section 6.2

In reality, it is not necessary to try all the possible combinations of all the possible values of each of the parameters in order to determine the best combination. It turns out that the OLS estimation method uses calculus to derive a set of simultaneous (normal) equations that can be solved to produce the set of parameters that minimize the sum of square deviations.

6.3.2 Multiple linear regression

For multiple regression, the model matrix includes a column of ones followed by a separate column for each of the predictor variables:

$$\begin{array}{ccccccccc} y_i & = & \beta_0 & + & \beta_1 x1_i & + & \beta_2 x2_i & + & \varepsilon_i \\ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ .. \end{bmatrix} & = & \begin{bmatrix} 1 & & x1_1 & & x2_1 \\ 1 & & x1_2 & & x2_2 \\ 1 & & x1_3 & & x2_3 \\ .. & & .. & & .. \end{bmatrix} & + & \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \ldots \end{bmatrix} \end{array}$$

6.3.3 Analysis of variance

Recall from section 6.1 that linear models comprising of a single factor are expressed as an effects model:

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

where μ estimates the overall constant (mean response for the groups) and α_i estimates the effect of each treatment group on the overall mean of groups ($\alpha_i = \mu_i - \mu$), and that the factor levels are represented by a set of dummy variable codings. The model component is therefore represented by a matrix that containings a column for the μ multiplier and a column for each of the dummy variable codings and has as many rows as there are observations. This is known as a *design matrix*. Table 6.2 illustrates the design matrix for the linear model appropriate for the fictitious data set containing a single factor within three levels ('G1', 'G2', 'G3') each with three replicates that was introduced in table 6.1.

Table 6.2: Fictitious data set (consisting of three replicates for each of three groups: 'G1','G2', 'G3') to illustrate the link between a) the indicator (dummy) variables and b) design matrix for the fictitious data set introduced in Table 6.1

a)

Y	$dummy_1$	$dummy_2$	$dummy_3$
2	1	0	0
3	1	0	0
4	1	0	0
6	0	1	0
7	0	1	0
8	0	1	0
10	0	0	1
11	0	0	1
12	0	0	1

b)

$$\begin{array}{c} \mu \quad \alpha_1 \quad \alpha_2 \quad \alpha_3 \\ design\ matrix = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Since the structure of the design matrix is repetitive, it can be summarized as:

$$design\ matrix = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

However, the effects model for a factor with p groups, will have $p+1$ parameters (the overall mean μ plus the p α parameters). There are more parameters than there are groups and consequently, the normal equations sets cannot yield single parameter estimates. Furthermore, recall from section **??** that all estimates must be independent of other estimates. Given that $\alpha_i = \mu_i - \mu$, it is only possible to estimate $p-1$ orthogonal (independent) parameters. For example, once μ and $p-1$ of the effects parameters have been estimated, the final effects parameter is no longer 'free to vary' and therefore cannot be independently estimated. Therefore the linear effects model is considered to be 'over-parameterized'[5]. In order to obtain parameter estimates, the model (and thus model matrix) must be reduced to a total of p parameters.

Over-parameterization can be resolved by removing one of the parameters from the effects model (either the overall mean (μ) or one of the treatment effects (α_i) parameters) by setting the value of the parameter to zero (this is known as the set-to-zero constraint). Hence a *model matrix* is created from a reduced form of the *design matrix*.

set μ to zero - in which case each of the treatment effects are equal to the means of their respective groups ($\alpha_i = \mu_i - \mu = \mu_i - 0 = \mu_i$). The model matrix is generated by removing the column of 1s (the first column) representing the intercept from the design matrix, and the fitted effects model becomes:

$y_{ij} = \alpha_i + \varepsilon_{ij}$

$$\underset{(three\ groups)}{model\ matrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Parameter	Estimates	Null hypothesis
α_1	mean of group 1 (μ_1)	H_0: $\mu_1 = 0$
α_2	mean of group 2 (μ_2)	H_0: $\mu_2 = 0$
α_3	mean of group 3 (μ_3)	H_0: $\mu_3 = 0$
...		

```
> Y1 <- c(2,3,4,6,7,8,10,11,12)
> A <- gl(3,3,9,c('G1','G2','G3'))
> summary(lm(Y1~-1+A))

Call:
lm(formula = Y1 ~ A - 1)

Residuals:
       Min          1Q      Median          3Q         Max
-1.000e+00  -1.000e-00  2.323e-16  1.000e+00  1.000e+00

Coefficients:
     Estimate  Std.  Error  t value  Pr(>|t|)
AG1    3.0000       0.5774    5.196   0.00202  **
AG2    7.0000       0.5774   12.124  1.91e-05  ***
AG3   11.0000       0.5774   19.053  1.35e-06  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 6 degrees of freedom
Multiple R-Squared:  0.989,      Adjusted R-squared:  0.9834
F-statistic:   179 on 3 and 6 DF,  p-value:  2.939e-06
```

[5]If the full linear model contains as many dummy variables as there are treatment groups, then it too is over-parameterized.

The effects parameters (α_1, α_2, ...) are labeled in the output by appending the names of each of the treatment groups to the name of the factor variable. Since the effect parameters estimate the group means (μ_i), null hypotheses about these individual parameters (for example that the mean of group 1 is equal zero; H_0: $\mu_1 = 0$, etc) as well as the null hypothesis tested for the overall linear model (that all group means equal zero; H_0: $\mu_1 = \mu_2 = ... = 0$) are of little biological relevance or interest.

Alternatively, parameter reduction can be achieved by generating a new set $(p-1)$ of effects parameters (α_q^*, where p represents the set of orthogonal parameters from 1 to $p-1$) each of which represent a linear combination of groups rather than a single group effect. That is, each α^* can include varying contributions from any number of the groups and are not restricted to a single contrast of $(= \mu_i - \mu)$.

The new set of effects parameters (α_q^*) are defined to represent $p-1$ orthogonal contrasts, each of which consists of a linear combination of one or more groups.

$\alpha_1^* = \bar{y_1}(C_{11}) + \bar{y_2}(C_{12}) + ...$

This is achieved by generating a new reduced model matrix via matrix algebra between the original over-parameterized model matrix and a matrix of contrast coefficients (*contrast matrix*). A contrast matrix (C_a) is a matrix of p rows and $p-1$ columns that defines the linear combinations of groups represented by each α^* parameter.

$$\begin{array}{c} \\ G1 \\ G2 \\ G3 \end{array} \begin{array}{c} \begin{array}{cc} \alpha_1^* & \alpha_2^* \end{array} \\ \begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \\ C_{13} & C_{23} \end{bmatrix} \end{array}$$

where the coefficients C_{11}, C_{12} and C_{13} define the degree and polarity of contributions from each of the groups ('G1', 'G2' and 'G3') towards α_1^* and so on.

To ensure orthogonality of the new parameters, the contrast matrix should be chosen such that the resulting reduced model matrix has a rank (sum of all values in matrix) equal to the number of columns in the reduced model matrix. This condition is satisfied when the sum of the products between each pair of contrast coefficients is equal to zero. The link between the original over-parameterized design matrix, the contrast matrix and the model matrix is depicted below for a factor with three groups ('G1', 'G2', 'G3'):

$$\begin{array}{c} \text{over-parameterized design matrix} \\ \begin{array}{ccc} \text{Intercept} & \alpha_1 & \alpha_2 \quad \alpha_3 \\ (\mu) & (G1) & (G2) \quad (G3) \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array} * \begin{array}{c} \\ G1 \\ G2 \\ G3 \end{array} \begin{array}{c} \begin{array}{cc} \alpha_1^* & \alpha_2^* \end{array} \\ \begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \\ C_{13} & C_{23} \end{bmatrix} \end{array} \Rightarrow \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \text{Intercept} & \alpha_1^* & \alpha_2^* \end{array} \\ \begin{bmatrix} 1 & C_{11} & C_{21} \\ 1 & C_{12} & C_{22} \\ 1 & C_{13} & C_{23} \end{bmatrix} \end{array}$$

The model matrix therefore defines a new linear model ($y_{qj} = \mu + \alpha_q^* + \varepsilon_{qj}$) such that $\alpha_q = C_a \alpha_q^*$. However, unlike the over-parameterized linear model, an algebraic representation of this new linear model (that includes the individual groups as terms) may be very difficult to write down and its parameters are often not directly interpretable. Nevertheless, since α is related to α^* according to the contrast matrix ($\alpha_i = C_a \alpha_q^*$), it is possible to back convert the estimated α^* into α estimates in order to reconstruct the original linear model.

A number of 'prefabricated' contrast matrices exist, each of which estimate a different set of specific comparisons between treatment combinations. The most common contrasts types include:

Treatment contrasts - in which each of the treatment groups means are compared to the mean of a 'control' group. The control treatment group effect (usually α_1^*) is set to zero and since $\alpha_1^* = \mu_1 - \mu$ (and thus $\mu_1 = \mu$), μ simply estimates the mean of the control treatment group (μ_1).

over-parameterized design matrix					contrast matrix				model matrix		
Intercept (μ)	α_1 ($G1$)	α_2 ($G2$)	α_3 ($G3$)			α_2^*	α_3^*		Intercept	α_2^*	α_3^*
1	1	0	0		$G1$	0	0		1	0	0
1	0	1	0	$*$	$G2$	1	0	$\Rightarrow$	1	1	0
1	0	0	1		$G3$	0	1		1	0	1

Parameter	Estimates	Null hypothesis
Intercept	mean of 'control' group (μ_1)	H_0: $\mu = \mu_1 = 0$
α_2^*	mean of group 2 minus mean of 'control' group ($\mu_2 - \mu_1$)	H_0: $\alpha_2^* = \mu_2 - \mu_1 = 0$
α_3^*	mean of group 3 minus mean of 'control' group ($\mu_3 - \mu_1$)	H_0: $\alpha_3^* = \mu_3 - \mu_1 = 0$
...		

```
> # specify that treatment contrasts should be used
> contrasts(A) <-contr.treatment
> summary(lm(Y~A))
Call:
lm(formula = Y1 ~ A)

Residuals:
       Min          1Q     Median         3Q        Max
-1.000e+00  -1.000e-00  6.939e-17  1.000e+00  1.000e+00

Coefficients:
            Estimate  Std.  Error  t value  Pr(>|t|)
(Intercept)   3.0000       0.5774    5.196   0.00202  **
A2            4.0000       0.8165    4.899   0.00271  **
A3            8.0000       0.8165    9.798   6.5e-05  ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  1 on 6 degrees of freedom
Multiple R-Squared:  0.9412,     Adjusted R-squared:  0.9216
F-statistic:    48 on 2 and 6 DF,  p-value:  0.0002035
```

Row labels in the table of estimated parameters (`Coefficients:`) reflect the α^* parameter that is estimated and tested on each line in that they comprise of the name of the factor vector followed by α^* number. The label `A1` literally represents α_1^*. The labels do not have a reference to any specific groups, since the α^* parameters represent linear combinations of multiple groups.

This approach to over-parameterization essentially applies the sum-to-zero constraint to one of the effects parameters and is computationally identical to fitting $p-1$ dummy variables

via multiple linear regression. However, due to the interpretation of the parameters (groups compared to a control) and the fact that treatment effects are not orthogonal to the intercept, the use of treatment contrasts (and thus dummy regression) is restricted to situations where there is clearly a single control group to which the other groups can be compared.

Sum to zero contrasts - this technique constrains the sum of the unconstrained treatment effects (α) to zero. In this model, the intercept estimates the average treatment effect and the remaining (α^*) estimate the differences between each of the treatment means and the average treatment mean.

$$\begin{array}{cccc} \text{over-parameterized design matrix} & & \text{contrast matrix} & & \text{model matrix} \\ \begin{array}{c}\text{Intercept}\\(\mu)\end{array} \quad \begin{array}{c}\alpha_1\\(G1)\end{array} \quad \begin{array}{c}\alpha_2\\(G2)\end{array} \quad \begin{array}{c}\alpha_3\\(G3)\end{array} & & \begin{array}{cc}\alpha_1^* & \alpha_2^*\end{array} & & \begin{array}{ccc}\text{Intercept} & \alpha_1^* & \alpha_2^*\end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} & * & \begin{array}{c} G1 \\ G2 \\ G3 \end{array} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} & \Rightarrow & \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix} \end{array}$$

Parameter	Estimates	Null hypothesis
Intercept	mean of group means (μ_{i^*}/p)	H_0: $\mu = \mu_q/p = 0$
α_1^*	mean of group 1 minus mean of group means ($\mu_1 - (\mu_q/p)$)	H_0: $\alpha_1 = \mu_1 - (\mu_q/p) = 0$
α_1^*	mean of group 2 minus mean of group means ($\mu_2 - (\mu_q/p)$)	H_0: $\alpha_2 = \mu_2 - (\mu_q/p) = 0$
...		

```
> # specify that sum-to-zero contrast should be used
> contrasts(A) <-contr.sum
> summary(lm(Y1~A))
Call:
lm(formula = Y1 ~ A)

Residuals:
       Min          1Q     Median         3Q        Max
-1.000e+00  -1.000e+00  1.388e-17  1.000e+00  1.000e+00

Coefficients:
              Estimate  Std.  Error   t value  Pr(>|t|)
(Intercept)   7.000e+00    3.333e-01    21.000   7.6e-07  ***
A1           -4.000e+00    4.714e-01    -8.485  0.000147  ***
A2            1.228e-16    4.714e-01  2.60e-16  1.000000
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  1 on 6 degrees of freedom
Multiple R-Squared:  0.9412,     Adjusted R-squared:  0.9216
F-statistic:    48 on 2 and 6 DF,  p-value:  0.0002035
```

Helmert contrasts - the intercept estimates the average treatment effect and the remaining (α^*) estimate the differences between each of the treatment means and the mean of the group before it.

over-parameterized design matrix | contrast matrix | model matrix

$$\begin{array}{c} \text{Intercept} \\ (\mu) \end{array} \begin{array}{c} \alpha_1 \\ (G1) \end{array} \begin{array}{c} \alpha_2 \\ (G2) \end{array} \begin{array}{c} \alpha_3 \\ (G3) \end{array} \quad \begin{array}{c} \\ G1 \\ G2 \\ G3 \end{array} \begin{array}{c} \alpha_1^* \\ \\ \end{array} \begin{array}{c} \alpha_2^* \\ \\ \end{array} \quad \begin{array}{c} \text{Intercept} \\ \\ \end{array} \begin{array}{c} \alpha_1^* \\ \\ \end{array} \begin{array}{c} \alpha_2^* \\ \\ \end{array}$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & -1 \\ 1 & 0 & 2 \end{bmatrix}$$

Parameter	Estimates	Null hypothesis
Intercept	mean of group means (μ_q/p)	H$_0$: $\mu = \mu_q/p = 0$
α_1^*	mean of group 2 minus mean of (group means and mean of group1) ($\mu_2 - (\mu_q/p + \mu_1)/2$)	H$_0$: $\alpha_1^* = \mu_2 - (\mu_q/p + \mu_1)/2 = 0$
α_2^*	mean of group 3 minus mean of (group means, mean of group1 and mean of group2) ($\mu_3 - (\mu_q/p + \mu_1 + \mu_2)/3$)	H$_0$: $\alpha_2^* = \mu_3 - (\mu_q/p + \mu_1 + \mu_2)/3 = 0$
...		

```
> # specify that Helmert contrasts should be used
> contrasts(A) <-contr.helmert
> summary(lm(Y1~A))
Call:
lm(formula = Y1 ~ A)

Residuals:
       Min          1Q      Median          3Q         Max
-1.000e+00  -1.000e-00  -7.865e-17  1.000e-00  1.000e+00

Coefficients:
             Estimate  Std.  Error  t value  Pr(>|t|)
(Intercept)    7.0000       0.3333   21.000   7.6e-07  ***
A1             2.0000       0.4082    4.899  0.002714  **
A2             2.0000       0.2357    8.485  0.000147  ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  1 on 6 degrees of freedom
Multiple R-Squared:  0.9412,     Adjusted R-squared:  0.9216
F-statistic:    48 on 2 and 6 DF,  p-value:  0.0002035
```

Polynomial contrasts - generate orthogonal polynomial trends. This is equivalent to fitting a multiple linear regression (or polynomial regression) with orthogonal parameters.
$y_{qj} = \beta_0^* + \beta_1^* x_{qj} + \beta_2^* x_{qj} + ... + \varepsilon_{qj}$

over-parameterized design matrix | contrast matrix | model matrix

$$\begin{array}{c} \text{Intercept} \\ (\mu) \end{array} \begin{array}{c} \alpha_1 \\ (G1) \end{array} \begin{array}{c} \alpha_2 \\ (G2) \end{array} \begin{array}{c} \alpha_3 \\ (G3) \end{array} \quad \begin{array}{c} \\ G1 \\ G2 \\ G3 \end{array} \begin{array}{c} \alpha_1^* \\ \\ \end{array} \begin{array}{c} \alpha_2^* \\ \\ \end{array} \quad \begin{array}{c} \text{Intercept} \\ \\ \end{array} \begin{array}{c} \alpha_1^* \\ \\ \end{array} \begin{array}{c} \alpha_2^* \\ \\ \end{array}$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} -7.071 & 0.408 \\ 0.00 & -0.816 \\ 7.071 & 0.408 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -7.071 & 0.408 \\ 1 & 0.00 & -0.816 \\ 1 & 7.071 & 0.408 \end{bmatrix}$$

Parameter	Estimates	Null hypothesis
Intercept	y-intercept	H_0: $\beta_0^* = 0$
β_1^*	partial slope for linear term	H_0: $\beta_1^* = 0$
β_2^*	partial slope for quadratic term	H_0: $\beta_2^* = 0$
...		

```
> # specify that orthogonal polynomial contrasts should be
   used
> contrasts(A) <-contr.poly
> summary(lm(Y1~A))
Call:
lm(formula = Y1 ~ A)

Residuals:
       Min          1Q      Median          3Q         Max
-1.000e+00  -1.000e-00  -1.712e-16  1.000e-00  1.000e+00

Coefficients:
              Estimate  Std.  Error    t value  Pr(>|t|)
(Intercept)   7.000e+00    3.333e-01     21.000   7.6e-07  ***
A.L           5.657e+00    5.774e-01      9.798   6.5e-05  ***
A.Q          -9.890e-16    5.774e-01  -1.71e-15         1
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  1 on 6 degrees of freedom
Multiple R-Squared:  0.9412,     Adjusted R-squared:  0.9216
F-statistic:    48 on 2 and 6 DF,  p-value:  0.0002035
```

user defined contrasts - In addition to the 'prefabricated' sets of comparisons illustrated above, it is possible to define other contrast combinations that are specifically suited to a particular experimental design. Contrasts are defined by constructing a contrast matrix (such as the one depicted below for a single factor with four levels: 'L1', 'L2', 'L3', 'L4') according to the following:

$$\begin{array}{c} \\ L1 \\ L2 \\ L3 \\ L4 \end{array} \begin{array}{c} \begin{array}{ccc} \alpha_1^* & \alpha_2^* & \alpha_3^* \end{array} \\ \begin{bmatrix} 1 & 0.5 & 0 \\ -1 & 0.5 & 0 \\ 0 & -0.5 & -1 \\ 0 & -0.5 & 1 \end{bmatrix} \end{array}$$

Parameter	Estimates	Null hypothesis
Intercept	mean of group means (μ_q/p)	H_0: $\mu = \mu_q/p = 0$
α_1^*	mean of group 1 minus mean of group 2 ($\mu_1 - \mu_2$)	H_0: $\alpha_1^* = \mu_1 - \mu_2 = 0$
α_2^*	partial mean of groups 1 and 2 minus mean of groups 3 and 4 ($(\mu_1+\mu_2)/2 - (\mu_3+\mu_4)/2$)	H_0: $\alpha_2^* = (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2 = 0$
α_3^*	mean of group 3 minus mean of group 4 ($\mu_3 - \mu_4$)	H_0: $\alpha_3^* = \mu_3 - \mu_4 = 0$

1. since each contrast (α^*) is a linear combination of one or more groups ($\alpha_1^* = \bar{y_1}(C_{11}) + \bar{y_2}(C_{12}) + ...$), groups to be included and excluded are represented by non-zero and zero coefficients respectively. Hence, the set of contrast coefficients (1, -1, 0, 0) would indicate that groups 1 and 2 should be included and groups 3 and 4 should be excluded.
2. groups to be apposed (contrasted) to one another should have apposing signs.Therefore, the contrast coefficients appropriate for contrasting group 1 against group 2 (when four groups) would be: 1, -1, 0, 0. Note that -1, 1, 0, 0 defines the same contrast, except that the estimated parameter will have the opposite sign. A consequence of this is that the sum of the contrast coefficients for a contrast should equal zero. That is if the coefficients within a column of the contrast matrix are added up they should equal zero.

Provided the following rules are adhered to (essentially to ensure orthogonality), any linear combinations of group comparisons can be defined:

1. the number of contrasts must not exceed $p-1$, where p is the number of groups. In fact, for the original effects parameters (and thus fitted linear model: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$) to be fully encapsulated by the orthogonal effects parameters (α_p^*), it is essential that exactly $p-1$ parameters be included. In practice, if fewer than $p-1$ contrasts are defined, the deficit will be automatically filled with additional contrasts that are orthogonal to existing contrasts.
2. each contrast (comparison) should be independent of (orthogonal to) each other comparisons. This restriction means that while it is possible to estimate $p-1$ parameters, in reality, only a maximum of $p-2$ are free to represent any contrast. Once $p-2$ contrasts have been defined, the final contrast must represent a specific linear combination of groups so as to satisfy the requirement that the sum of α_i equal zero. Often it is difficult to determine whether two contrasts are independent, however provided the sum of the products between their pairs of contrast coefficients equals zero, the pairs of contrasts are orthogonal to one another.
3. if the exact value of the estimated parameters is important, then within a contrast (column) the sum of the positive contrasts should equal 1, as should the sum of the negative coefficients. Note that in the example contrast matrix above, the contrast coefficients for α_2^* where issued as 0.5, 0.5, -0.5, -0.5 rather than 1, 1, -1, -1. Although this rule is not important for hypothesis testing (as it does not impact on t-values/F-ratios and p-values) it does effect the magnitude of the parameter estimates.

Some of these points are illustrated in table 6.3.

Consider again the data introduced in table 6.1 (single factor called A, comprising of four levels each with three replicates). Supposing group 1 (G1) represented a control treatment and groups 2 (G2) and 3 (G3) represented two different experimental treatments. We may wish to contrast the mean of group 1 to the mean of groups 2 and 3, thereby examining whether the treatments are different to the controls.

We begin by confirming that the categorical variable is considered to be a factor (see section **??**) and determining the ordering of the factor.

```
> A
[1] G1 G1 G1 G2 G2 G2 G3 G3 G3
Levels:  G1 G2 G3
```

This informs us that the first group is called `G1`, the second is `G2`, etc. Therefore, contrast coefficients corresponding to an examination of $\alpha_1^* = \mu_1 - (\mu_2 + \mu_3)$, could be 1, -0.5, -0.5.

```
> # define potential contrast matrix
```

Table 6.3: Estimations of α_1^* by user-defined contrasts applied when μ (mean of group means) was 8 and μ_i (means of the groups 'L1' to 'L4') are 5, 7, 10 and 10 respectively.

'L1'	'L2'	'L3'	'L4'	Estimates	α^*
1	-1	0	0	$\alpha^* = (1)\times\mu_1 + (-1)\times\mu_2 + (0)\times\mu_3 + (0)\times\mu_4$	
				$= (1)\times5 + (-1)\times7 + (0)\times10 + (0)\times10$	-2
-1	1	0	0	$\alpha^* = (-1)\times\mu_1 + (1)\times\mu_2 + (0)\times\mu_3 + (0)\times\mu_4$	
				$= (-1)\times5 + (1)\times7 + (0)\times10 + (0)\times10$	2
2	-2	0	0	$\alpha^* = (2)\times\mu_1 + (-2)\times\mu_2 + (0)\times\mu_3 + (0)\times\mu_4$	
				$= (2)\times5 + (-2)\times7 + (0)\times10 + (0)\times10$	-4
1	1	-1	-1	$\alpha^* = (1)\times\mu_1 + (1)\times\mu_2 + (-1)\times\mu_3 + (-1)\times\mu_4$	
				$= (1)\times5 + (1)\times7 + (-1)\times10 + (-1)\times10$	-8
0.5	0.5	-0.5	-0.5	$\alpha^* = (0.5)\times\mu_1 + (0.5)\times\mu_2 + (-0.5)\times\mu_3 + (-0.5)\times\mu_4$	
				$= (0.5)\times5 + (0.5)\times7 + (-0.5)\times10 + (-0.5)\times10$	-4

```
> contrasts(A) <- cbind(c(1, -0.5, -0.5))
> # examine the resulting contrast matrix
> contrasts(A)
    [,1]          [,2]
G1   1.0 -6.407635e-17
G2  -0.5 -7.071068e-01
G3  -0.5  7.071068e-01
```

Note that because we only defined a single contrast (and yet two are required), an additional contrast was added that is orthogonal to the first and ensures that $\alpha_i = 0$. To confirm orthogonality, examine the lower left hand corner of a matrix of cross products generated using the `crossprod()` *function*. The matrix is also rounded to a maximum of 2 decimal places to improve readability.

```
> round(crossprod(contrasts(A)),2)
      [,1] [,2]
 [1,]  1.5    0
 [2,]  0.0    1
```

Since the cross product of contrast 1 and 2 equals zero, orthogonality is satisfied. Fit linear effects model with defined orthogonal contrasts.

```
> l <- lm(Y1~A)
> # summarize the model fitting
> summary(l)
....
Coefficients:
            Estimate  Std.  Error  t value  Pr(>|t|)
(Intercept)   7.0000       0.3333   21.000   7.6e-07  ***
A1           -4.0000       0.4714   -8.485  0.000147  ***
A2            2.8284       0.5774    4.899  0.002714  **
---
....
```

Although it is difficult to construct an algebraic representation of the orthogonal linear model, and it is difficult to interpret exactly what the α_2^* parameter is actually estimating[6], it is possible to estimate the original linear model and its parameters.

```
> dummy.coef(l)
Full coefficients are

(Intercept):                7
A:                         G1            G2            G3
-4.000000e+00  7.987317e-16  4.000000e+00
```

After rounding, the original linear (dummy) model is therefore:
$y_{ij} = \mu + \alpha_i + _{ij}$
$y_{ij} = 7 + -4(G1)_{ij} + 0(G2)_{ij} + 4(G3)_{ij} + _{ij}$
Or more generally:
$y = 7 + -4(G1) + 0(G2) + 4(G3)$

Note that if we had used the contrast coefficients 2, -1, -1 to define α_1^* instead, the estimate of α_1^* would have been different (-2 instead of -4), but all other aspects relating to hypothesis testing would have been unaffected.

Supposing instead we had considered two contrasts ($\alpha_1^* = \mu_1 - (\mu_2 + \mu_3)/2$ and $\alpha_2^* = \mu_1 - \mu_3$), where the second contrast examines the difference between the control group ('G1') and just one of the experimental groups ('G3'):

```
> # define potential contrast matrix
> contrasts(A) <- cbind(c(1, -0.5, -0.5), c(1, 0, -1))
> # examine the resulting contrast matrix
> contrasts(A)
    [,1]  [,2]
G1   1.0     1
G2  -0.5     0
G3  -0.5    -1
```

We now need to determine whether the contrasts are orthogonal.

```
> round(crossprod(contrasts(A)),2)
      [,1]  [,2]
 [1,]  1.5   1.5
 [2,]  1.5   2.0
```

The sum of the product between pairs of contrasts for contrasts 1 and 2 is 1.5 (ie not zero) and thus the two contrasts are not orthogonal. We should not proceed to fit the model with these contrasts.

By default, R[7] employs treatment contrasts for unordered factors[8] and orthogonal polynomial contrasts for ordered factors, although this behavior can be altered to an alternative (such as `contr.sum` for unordered factors) using the `options(contrasts=c("contr.sum", "contr.poly"))` *function.*

[6] While the contrasts indicated that α_2^* was contrasting the means of group 2 and 3, the weightings of the contrasts are make biological interpretation of the estimated parameter difficult, i.e $\alpha_2^* = \mu - (-7.078 \times \mu_1 + 7.078 \times \mu_2)$

[7] Note that the default behaviour of S-PLUS is to employ sum to zero contrasts for unordered factors

[8] Unordered factors are factors that have not specifically defined as 'ordered', see section 2.6.1. The order of groups in an ordered factor is usually important - for example when examining polynomial trends across groups

Note that while the estimates and interpretations of individual model parameters differ between the alternative approaches, in all but the $\mu = 0$ (set-to-zero) case, the overall effects model is identical ($y_{qj} = \mu + \alpha^*_q + \varepsilon_{qj}$). Hence, the overall null hypothesis tested from the effects model (H_0: $\alpha^*_1 = \alpha^*_2 = ... = 0$) is the same irrespective of the contrasts chosen.

When the model contains more than one factor, a separate term is assigned for each factor and possibly the interactions between factors (e.g. $\alpha_i + \beta_j + \alpha\beta_{ij}$). Alternatively, statistical models containing factors can be expressed as *cell means models* in which the overall mean and treatment effects ($\mu + \alpha_i$) are replaced by the treatment (cell) means (μ_i). In the cell means model, there are as many cell means as there are unique treatment levels. These differences are thus summarized:

Linear model	$y_{ij} = \mu + \beta_1(dummy_1)_{ij} + \beta_2(dummy_2)_{ij} + ... + \varepsilon_{ij}$
Linear effects model	$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
Orthogonal linear effects model	$y_{i^*j} = \mu + \alpha^*_{i^*} + \varepsilon_{i^*j}$
Cell means model	$y_{ij} = \mu_i + \varepsilon_{ij}$

For simple model fitting the choice of model type makes no difference, however for complex factorial models in which entire treatment levels (cells) are missing, full effects models cannot be fitted and therefore cell means models must be used.

6.3.4 Linear model hypothesis testing

Hypothesis testing is usually concerned with evaluating whether a population parameter is (or set of parameters are) equal to zero, as this signifies no 'relationship' or 'effect'.

Null hypotheses about individual model parameters

In a linear model, there is a null hypothesis associated with each of the individual model parameters (typically that the parameter is equal to zero), although not all the testable null hypotheses are necessarily biologically meaningful. Consider again the simple linear regression model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
This linear model includes two parameters (β_0 and β_1), and thus there are two individual testable null hypotheses - that the population y-intercept is equal to zero (H_0: $\beta_0 = 0$) and the slope is equal to zero (H_0: $\beta_1 = 0$). While rejecting a null hypothesis that the slope parameter equals zero indicates the presence of a 'relationship', discovering that the value of the response variable when the predictor variable is equal to zero is usually of little biological relevance.

Null hypotheses about individual model parameters are usually tested using a t-test (see chapter **??**), or equivelently via a single factor ANOVA (see chapter 9) with a single degree of freedom. The latter approach is often employed when user-defined contrasts are involved as it enables the results to be expressed in the context of the overall linear model (see below and section **??**).

Null hypotheses about fit of overal model

Recall that in hypothesis testing, a null hypothesis (H_0) is formulated to represent all possibilities except the hypothesized prediction and that disproving the null hypothesis provides evidence that some alternative hypothesis (H_A) is true. Consequently, there are typically at least two models fitted. The *reduced model*, in which the parameter of interest (and its associated predictor variable) is absent (or equivalently set to zero) represents the model predicted by null hypothesis. The *full model* represents the alternative hypothesis and includes the term of interest. For example, to test the null hypothesis that there is no relationship between populations *x* and *y* (and thus that the population slope (β_1)= 0):

$$\begin{aligned} \textit{full model } (H_A) \text{ -} \quad & y_i = \beta_0 + \beta_1 x_i + \textit{error}_i \\ \textit{reduced model } (H_0) \text{ -} \quad & y_i = \beta_0 + 0 x_i + \textit{error}_i \\ & = \beta_0 + \textit{error}_i \end{aligned}$$

The degree to which a model 'fits' the observed data is determined by the amount of variation that the model fails to explain, and is measured as the sum of the squared differences (termed *SS* or sums-of-squares) between the observed values of the response variable and the values predicted by the model. A model that fits the observed data perfectly will have a SS of 0.

The *reduced model* measures the amount of variation left unexplained by the statistical model when the contribution of the parameter and predictor variable (term) of interest is removed (SS_{Total}). The *full model* measures the amount of variation left unexplained by the statistical model when the contribution of the term is included ($SS_{Residual}$). The difference between the reduced and full models (SS_{Model}) is the amount of explained variation attributed to the term of interest. When the null hypothesis is true, the term of interest should not explain any of the variability in the observed data and thus the full model will not fit the observed data any better than the reduced model. That is, the proposed model would not be expected to explain any more of the total variation than it leaves unexplained. If however, the full model fits the data 'significantly' better (unexplained variability is substantially less in the full model compared to the reduced model) than the reduced model, there is evidence to reject the null hypothesis in favour of the alternative hypothesis.

Hypothesis testing formally evaluates this proposition by comparing a ratio of explained to unexplained variation to a probability distribution representing all possible ratios theoretically obtainable when the null hypothesis is true. The total variability in the observed data is partitioned into at least three sources.

1. the variation that is explained by the model (SS_{Model})
   $SS_{Model} = \quad SS_{Total}$ (*reduced model*) $- SS_{Residual}$ (*full model*)

2. the variation that is unexplained by the model
   $SS_{Residual}$ (*full model*)

3. the total variation in the observed data
   SS_{Total} (*reduced model*)

The number of degrees of freedom associated with estimates of each source of variation reflect the number of observations involved in the estimate minus the number of other parameters that must have been estimated previously. Just like *SS*, df are additive and therefore:

$$df_{Model} = df_{Total} \text{ (\textit{reduced model})} - df_{Residual} \text{ (\textit{full model})}$$

Each of the sources of variation are based on a different number of contributing observations. Therefore more comparable, standardized versions are generated by dividing by the appropriate number of (degrees of freedom). These new measures of variation (known as mean squares or *MS*) are thus conservative mean measures of variation and importantly, they have known probability distributions.

The partitioned sources of variation are tabulated in the form of an analysis of variance (ANOVA) table (see table 6.4), which also includes a ratio (F-ratio) of MS_{Model} to $MS_{Residual}$. When the null hypothesis is true MS_{Model} and $MS_{Residual}$ are expected to be the same, and thus their ratio (F-ratio) should be approximately 1. An F-ratio based on observed data is thus compared to an appropriate F-distribution (theoretical distribution of all possible F-ratios for the set of degrees of freedom) when the null hypothesis is true. If the probability of obtaining such an F-ratio (or one more extreme) is

less than a critical value, the null hypothesis is rejected.

Table 6.4: Analysis of variance (ANOVA) table for a simple linear model. n is the number of observations, f_p is the number of parameters in the full model and r_p is the number of parameters in the reduced model

Source of variation	SS	df	MS	F-ratio
Model	SS_{Model}	$f_p - 1$	$\frac{SS_{Model}}{df_{Model}}$	$\frac{MS_{Model}}{MS_{Residual}}$
Residual	$SS_{Residual}$	$n - f_p$	$\frac{SS_{Residual}}{df_{Residual}}$	
Total	SS_{Total}	$n - r_p$	$\frac{SS_{Residual}}{df_{Residual}}$	

When there are multiple predictor variables, as well as assessing the fit of the overall model, we usually want to determine the effect of individual factors. This is done by comparing the fit of models with and without the specific term(s) associated with that variable.

6.4 Comments about the importance of understanding the structure and parameterization of linear models

An understanding of how to formulate the correct statistical model from a design and set of null hypotheses is crucial to ensure that the correct R syntax (and thus analysis) is employed. This is particularly important for more complex designs which incorporate multiple error strata (such as partly nested ANOVA). Table 6.5 briefly illustrates the ways in which statistical models are represented in R. Moreover, in each of the remaining chapters, the statistical models as well as the appropriate R model formulae for each major form of modeling will be highlighted and demonstrated, thereby providing greater details about use of R in statistical modeling.

Table 6.5: Statistical models in R. Lower case letters denote continuous numeric variables and upper-case letters denote factors. Note that the error term is always implicit.

Effects model	R Model formular	Description
$y_i = \beta_0 + \beta_1 x_i$	`y ~ 1 + x` `y ~ x`	Simple linear regression model of `y` on `x` with intercept term included

$y_i = \beta_1 x_i$	`y ~ 0 + x`
	`y ~ -1 + x`
	`y ~ x - 1`
		Simple linear regression model of `y` on `x` with intercept term excluded

$y_i = \beta_0$	`y ~ 1` `y ~ 1 - x`	Simple linear regression model of `y` against the intercept term
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$	`y ~ x1 + x2`	Multiple linear regression model of `y` on `x1` and `x2` with the intercept term included implicitly
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2$	`y ~ 1 + x + I(x^2)` `y ~ poly(x, 2)`	Second order polynomial regression of `y` on `x` As above, but using orthogonal polynomials
$y_{ij} = \mu + \alpha_i$	`y ~ A`	Analysis of variance of `y` against a single factor `A`
$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$	`y ~ A + B + A:B` `y ~ A*B`	Fully factorial analysis of variance of `y` against `A` and `B`
$y_{ijk} = \mu + \alpha_i + \beta_j$	`y ~ A*B - A:B`	Fully factorial analysis of variance of `y` against `A` and `B` without the interaction term (equivalent to `A + B`)
$y_{ijk} = \mu + \alpha_i + \beta_{j(i)}$	`y ~ B %in% A` `y ~ A/B`	Nested analysis of variance of `y` against `A` and `B` nested within `A`
$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \overline{x})$	`y ~ A*x` `y ~ A/x`	Analysis of covariance of `y` on `x` at each level of `A`
$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{j(i)k}$	`y ~ A + Error(B) + C + A:C + B:C`	Partly nested ANOVA of `y` against a single between block factor (`A`), a single within block factor (`C`) and a single random blocking factor (`B`).