

Adding constraints: MARK and linear models

[REDACTED]

6.1. A (brief) review of linear models

If you have a background in linear models, then much of this material will be familiar. If you're a statistician, obviously we're leaving out a lot of the 'details' (to say the least). Our purpose is to provide a minimum level of background, so even newcomers to linear models have a 'feel' for the approach. If you are new to linear models, we strongly suggest you supplement your reading of this chapter by having a look at one of the many good textbooks on this subject. McCullagh & Nelder (1989) and Dobson & Barnett (2008) are particularly good.

The basic idea underlying linear models can be stated quite simply: the response variable in many statistical analyses can be expressed as a linear regression function of 1 or more other factors. In fact, any ANOVA-type design can be analyzed using linear regression models (although interpretation of interactions is sometimes complex). In general, for data collected from marked individuals, the 'response variable' is often a probability or proportion (e.g., survival or recapture rate), which must be transformed prior to analysis using a linear models approach (we'll get to that in a moment). For the

moment, assume the response variable has been suitably transformed.

We'll start by demonstrating this relationship between 'regression' and 'ANOVA', by means of a simple example. Consider data from a study where the skull circumference of young pre-school children is measured, and we're interested in knowing if this structure is on average larger in males than in females (we'll assume for the moment that all of the children were the same chronological age). Let's suppose we measure 7 male and 7 female children, and analyze our data using a normal single-classification ANOVA. Here are the data:

```
male 7.2 7.1 9.1 7.2 7.3 7.2 7.5
female 9.8 8.5 8.7 8.6 8.4 7.7 8.2
```

First, the results from a 'standard ANOVA' (as you might generate using some statistical analysis software):

```
Source  df  SS    MS    F    P
SEX      1  3.806  3.806  8.33  0.0137
Error    12  5.485  0.457
Total    13  9.292
```

The results of this analysis indicate a marginally significant difference between male and female children.

However, what if our statistics package was limited only to a regression subroutine? Could we have analyzed our data using a linear regression model, instead of ANOVA, and arrived at the same result? The answer is, indeed, yes, we can. What we do is simply take the classification factor (SEX) and 'code' it as a '0' or '1' dummy variable (we'll see why in just a moment). For example, let '0' represent females, and '1' represent males. Thus, every individual in our data set is assigned a '0' or a '1', depending upon their gender. Let's call this dummy variable SEX. Now, all we need to do is regress our response variable (the skull circumference) on the dummy variable for SEX. Here are the results of the regression analysis:

```
Source  df  SS    MS    F    P
SEX      1  3.806  3.806  8.33  0.0137
Error    12  5.485  0.457
Total    13  9.292
```

No, it's not a typo – it is in fact the exact same table as above. The two approaches are entirely synonymous, yielding identical results. How can this be? The answer lies in the structure of the models actually being tested. So, let's step back to the beginning, and look at things a bit more formally.

In general, a linear model can be expressed in matrix form as

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

where \mathbf{y} is a vector of responses (i.e., a vector of the response variables), $\boldsymbol{\beta}$ is a vector of parameters (e.g., the intercept and 1 or more 'slopes'), \mathbf{X} is a matrix with either '0' or '1' elements, or values of 'independent' variables, and $\boldsymbol{\epsilon}$ is a vector of random error terms.

In cases of analysis of variation of the response variable among different levels of one or more classification (i.e., 'treatment' or 'factor') levels, there is a parameter β in the vector $\boldsymbol{\beta}$ to represent each level of a factor. The elements of \mathbf{X} (which is generally referred to as the *design matrix* – discussed below) are chosen to exclude or include the appropriate parameters for each observation. These elements are

often referred to as either ‘dummy’ or ‘indicator’ variables (‘indicator’ generally being used when only ‘1’ or ‘0’ are used as the coding variables).

The following simple example will make this clear, and will illustrate the underlying connection between a linear regression model and analysis of variation (ANOVA). Suppose you have collected data on the scutum width of male and female individuals of some insect species. You are interested in whether or not the difference in mean scutum width between the sexes differs more than would be expected by random chance. Normally, you might consider using a single-classification (Model I) ANOVA for this sort of analysis. Recall that for this sort of analysis, any single variate Y (in this case, Y = scutum width), can be decomposed as:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

In other words, each individual variate Y_{ij} is the sum of the global mean (μ), the deviation of the individual from that mean due to the ‘classification’ factor (sex; α_i), and the random error term (ϵ_{ij}). In this example, with 2 levels of the classification factor (i.e., males and females), we would be testing for differences of the type ($\alpha_1 - \alpha_2$). If ($\alpha_1 - \alpha_2$) = 0 (the null hypothesis), then we would conclude no significant group effect (i.e., no significant difference in group means between the sexes).

How could we use linear regression to approach the same analysis? In a regression analysis, each individual variate Y_i would be decomposed as:

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

In this case, each variate Y_i is the sum of the product of the slope (β_2) and the variable x , the intercept (β_1), and a random error term (ϵ). In this case, the hypothesis being tested is whether or not the estimate of the slope is significantly different from 0 ($H_0: \beta_2 = 0$).

However, what is the variable ‘ x ’? In fact, this is the key to understanding the connection between the regression model and the ANOVA analysis. In the regression formulation, x represents a coding (‘dummy’) variable specifying male or female (i.e., sex, the classification variable in the ANOVA analysis). The coding variable takes on the value of ‘0’ or ‘1’ (‘0’ for females, ‘1’ for males). We regress the response variable Y (scutum width) on the coding variable for sex. If the slope (β_1) is not different from 0, then we interpret this as evidence that the numerical value of the coding variable does not significantly influence variation in our data. Put another way, if the slope does not differ from 0, then this indicates no significant difference between the sexes. This is entirely analogous to test of the ($\alpha_1 - \alpha_2$) hypothesis in the ANOVA analysis.

Recall that we can express a linear model in matrix form as

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

where \mathbf{y} is a vector of responses (i.e., a vector of the response variables), $\boldsymbol{\beta}$ is a vector of parameters (e.g., the intercept and 1 or more ‘slopes’), \mathbf{X} is a matrix with either ‘0’ or ‘1’ elements, or values of ‘independent’ variables, and $\boldsymbol{\epsilon}$ is a vector of random error terms. For our present example, the design matrix \mathbf{X} consists of 2 columns of ‘0’ and ‘1’ dummy variables (the first column corresponding to the intercept, β_1 , and the second column corresponding to dummy variable coding for a given sex, β_2).

Given K individuals in each sex (although a balanced design is not required), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ can be written as

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1K} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2K} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1K} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2K} \end{bmatrix}$$

In fact, in this case, if we used '1' to code for males, and '0' to code for females, then the intercept (β_1) would represent the estimate for female survival (since if the dummy variable is '0', then all that remains in the model is the intercept, and the random error term). The β_2 term actually reflects (male survival - female survival), such that $\beta_1 + \beta_2 = (\text{female}) + (\text{male} - \text{female}) = \text{male survival}$. The structure of the design matrix is discussed in more detail in the next section.

It is perhaps worth noting that models of the form ' $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ' are called linear models because the non-error part of the expression $\mathbf{X}\boldsymbol{\beta}$ is a *linear* combination of the parameters (and not specifically because of the relationship of ANOVA to linear regression). **MARK** uses this general linear models approach as the basis for all of the analysis (data) types available.

begin sidebar

matrix approach to linear regression & ANOVA: simple introduction

Here, we provide a *very* simple example of a matrix approach to linear regression (and, by extension, to linear models in general). For deeper understanding, you are strongly urged to consult one of the several very good textbooks which give *much* fuller treatments of the subject.

Consider the linear model, say of individual (i) with mass (Y_i) relative to sex (X_i , where $X = 0$ or $X = 1$ for – say – female or male, respectively), measured with Gaussian (normally) distributed random variation (ϵ_i) about the mean. We'll assume the following 'fake' data:

	mass (Y)			
male ($X = 1$)	11	12	11	14
female ($X = 0$)	8	11	12	10

The mean mass for males ($\bar{x}_m = 12$) is larger than the mean mass for females ($\bar{x}_f = 10.25$) – the usual question being, is the difference between the two larger than expected due to random chance?

We could adopt a linear models approach to answering this question – first, we could write the relationship between mass and sex in linear model form as

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

The null hypothesis of 'no difference between sexes' can be expressed formally in terms of the β term for sex; i.e., $\mathbf{H}_0 : \beta_2 = 0$. The technical problem then is estimating the β_i coefficients in the linear model. To do this, first define a vector \mathbf{y} for all the Y_i , a matrix \mathbf{X} for a vector of 1s and all the X_i , a vector $\boldsymbol{\epsilon}$ for all the ϵ_i , and further define a vector $\boldsymbol{\beta}$ for the coefficients β_1 and β_2 .

Then (for our ‘fake’ data set) we get

$$\mathbf{Y} = \begin{bmatrix} 11 \\ 12 \\ 11 \\ 14 \\ 8 \\ 11 \\ 12 \\ 10 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Note that the matrix \mathbf{X} is referred to as the *design matrix* – the construction of the design matrix is fundamental to using linear models in **MARK**, as we will cover in considerable detail later in this chapter. So, to derive estimates of the β_i coefficients, we need to find a vector $\boldsymbol{\beta}$ such that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. Is this possible? The answer is clearly ‘no’, because that would require the points to lie exactly on a straight line. A more modest (and tractable) question is: can we find a vector $\hat{\boldsymbol{\beta}}$ such that $\mathbf{X}\hat{\boldsymbol{\beta}}$ is in a sense ‘as close to \mathbf{y} as possible?’. The answer is ‘yes’. The task is to find $\hat{\boldsymbol{\beta}}$ such that the length of the vector $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is as small as possible (i.e., $\boldsymbol{\epsilon} \rightarrow 0$).

How do we get there from here? Fairly easily. First, we note that what we’re trying to do is solve for $\boldsymbol{\beta}$ in the linear model. The first step is to let $\boldsymbol{\epsilon} = 0$ (such that it drops out of the equation – this should make sense, if you keep in mind that what we’re trying to do is to find $\hat{\boldsymbol{\beta}}$ such that the length of the vector $\boldsymbol{\epsilon}$ is, in effect, 0). This leaves us with

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

Then, a few steps of algebra to solve for the vector $\boldsymbol{\beta}$:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}^T \mathbf{y} &= \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \\ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \\ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} &= \boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

In words, we multiply both sides of the initial equation by the transpose of \mathbf{X} to get the crossproduct $\mathbf{X}^T \mathbf{X}$, which is a square matrix (*note*: the square matrix $(\mathbf{X}^T \mathbf{X})$ is called the *pseudo inverse* of \mathbf{X} . We cannot use the true matrix inverse of \mathbf{X} – that is, \mathbf{X}^{-1} – because it generally does not exist as \mathbf{X} is not a square matrix; $m \neq n$). We then find the inverse of this cross-product matrix and multiply both sides by that. This allows us to cancel out the term involving \mathbf{X} on the right-hand side of the equation, allowing us to find an estimate of $\boldsymbol{\beta}$, which we call $\hat{\boldsymbol{\beta}}$, in terms of the original data.

It is worth noting that we could also approach this problem using the more familiar method of *least squares*. Recall that least squares involves minimizing the sum of the squared residuals between the observed and expected values. More formally, we want to minimize the Euclidean norm squared of the residual $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, that is, the quantity

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = ([Y_1 - (\mathbf{X}\boldsymbol{\beta})_1])^2 + ([Y_2 - (\mathbf{X}\boldsymbol{\beta})_2])^2 + \cdots + ([Y_i - (\mathbf{X}\boldsymbol{\beta})_i])^2$$

where $(\mathbf{X}\boldsymbol{\beta})_i$ denotes the i th component of the vector $(\mathbf{X}\boldsymbol{\beta})$.

We could also rewrite this as

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= ([Y_1 - (\mathbf{X}\boldsymbol{\beta})_1])^2 + ([Y_2 - (\mathbf{X}\boldsymbol{\beta})_2])^2 + \cdots + ([Y_i - (\mathbf{X}\boldsymbol{\beta})_i])^2 \\ &= \sum_{i=1}^n (Y_i - (\beta_1 + \beta_2 x_i))^2\end{aligned}$$

You might recall (from some linear algebra class you might have taken) that for some vector $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^T \boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_n \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} = \theta_1^2 + \theta_2^2 + \cdots + \theta_n^2 = \sum_i^n \theta_i^2$$

Thus, if $\boldsymbol{\theta} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, then we can write

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}\end{aligned}$$

All that's left is to differentiate this expression with respect to $\boldsymbol{\beta}$, set to 0, and solve. Let

$$S = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Thus,

$$\begin{aligned}\frac{\partial S}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = 0 \\ \mathbf{X}^T \mathbf{y} &= \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}$$

Note the resulting solution is *identical* to that obtained from the preceding solution to the linear set of equations (i.e., the linear algebra approach).

In fact, we could show that both solutions are equivalent to the MLE estimates for $\boldsymbol{\beta}$ (the Gaussian linear model is nice in the sense that the parameter estimates – namely the solution to the linear set of equations, the least squares estimate, and the maximum likelihood estimate – are all the same).

So for our ‘fake’ data:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \begin{bmatrix} 10.25 \\ 1.75 \end{bmatrix}\end{aligned}$$

Thus, our estimates for the intercept and slope are $\hat{\beta}_1 = 10.25$ and $\hat{\beta}_2 = 1.75$, respectively. We would next estimate the error variance for $\hat{\beta}_1$ and $\hat{\beta}_2$. First, we derive an estimate of the variance-covariance matrix for the vector $\boldsymbol{\beta}$ estimates as

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma_e^2$$

We can estimate σ_e^2 from the residual sums of squares (RSS) as

$$RSS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

If the model estimates p parameters, then the estimate of σ_e^2 is simply $RSS/(N - p)$ where N is the number of data points. Thus,

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) / (N - p)$$

So, for our ‘fake’ data (where $N = 8$ and $p = 2$), and our vector $\hat{\beta}$,

$$\begin{aligned} \text{RSS} &= (\mathbf{y} - \mathbf{X}\hat{\beta})^T (\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= 14.75 \end{aligned}$$

and thus

$$\begin{aligned} \text{var}(\hat{\beta}) &= (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta})^T (\mathbf{y} - \mathbf{X}\hat{\beta}) / (N - p) \\ &= \begin{bmatrix} 0.6146 & -0.6146 \\ -0.6146 & 1.2292 \end{bmatrix} \end{aligned}$$

Thus $\widehat{\text{SE}}(\hat{\beta}_1) = \sqrt{0.6146} = 0.7840$, and $\widehat{\text{SE}}(\hat{\beta}_2) = \sqrt{1.2292} = 1.1087$. And, since a 95% CI for $\hat{\beta}_2$ (approximately $\hat{\beta}_2 \pm 2\text{SE}$; $[-0.4674, 3.9674]$) clearly bounds 0, we would conclude no significant sex effect at a nominal $\alpha = 0.05$ level.

end sidebar

6.2. Linear models and the ‘design matrix’: the basics

In program **MARK**, the default design matrix for a given model is determined by the parameter structure of the model you are trying to fit (number of groups, and the number and structure of the parameters; i.e., the PIMs). This design matrix is then modified in various ways to examine the relative fit of different models to the data. In order to understand this process, it is essential that you understand how the design matrix is constructed.

Perhaps the best way to introduce the concept of a design matrix is by means of an example. Suppose you are doing a ‘typical’ ANOVA on data with a single classification factor (say, ‘treatment’). Suppose that there are 4 levels for this factor (perhaps a control, and 3 different levels of the ‘treatment’). You want to test the hypothesis that there is no heterogeneity among ‘treatment’ levels ($\mathbf{H}_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$). Recall from the preceding discussion that this problem can be formulated as an applied linear regression problem using ‘0/1 dummy variable’ coding for the different levels of the ‘treatment’.

Recall the previous example (above) which had 1 ‘treatment’ or classification factor (sex), with 2 levels (male and female). The corresponding regression model was

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

where x represented a coding variable specifying male or female (i.e., sex, the classification variable in the ANOVA analysis). The coding variable took on the value of ‘0’ or ‘1’ (‘0’ for females, ‘1’ for males).

What would the regression model look like for our present example, with 4 levels of the treatment factor instead of 2? How can we use a simple ‘0’ or ‘1’ dummy variable coding scheme (which clearly has only 2 ‘levels’) to accommodate a treatment factor with 4 levels? The key is to consider the answer to the following question: if x_i can take on 1 of 2 values (0 or 1), then how many values of x_i do we need to specify k levels of the classification variable (i.e., the treatment variable)? If you think about it for a moment, you should realize that the answer is $k - 1$ (which, of course, corresponds to the degrees of freedom for a single-classification ANOVA). Thus, for the present example, x_1 , x_2 and x_3 could be:

$$x_1 = \begin{cases} 1 & \text{if trt 1} \\ 0 & \text{if other} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if trt 2} \\ 0 & \text{if other} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if trt 3} \\ 0 & \text{if other} \end{cases}$$

Clearly, when the coefficients for x_1 , x_2 and x_3 are all 0, then the treatment level must be 4 (‘other’).

Thus, our regression equation for this example would be:

$$Y_i = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \epsilon_i$$

In this case, β_1 is the intercept, while β_2 , β_3 and β_4 correspond to the slopes for each of the levels of the treatment factor. Since there are 4 levels of the treatment, 3 slopes are needed to code 4 levels of the treatment, because 1 of the levels of the treatment corresponds to the case where all 3 slopes are 0. Parameters β_2 , β_3 and β_4 refer to treatment levels 1, 2, and 3, respectively. If $x_1 = x_2 = x_3$, then β_1 refers to treatment level 4. In other words, the intercept corresponds to treatment level 4.

begin sidebar

why is level 4 the intercept?

Choosing the intercept to specify treatment 4 was entirely arbitrary – we could for example have used any other level of the treatment as the intercept, and adjusted the coding for the remaining levels according. For example, we could have used level 1 of the treatment as ‘other’ (i.e., the intercept), as follows:

$$x_1 = \begin{cases} 1 & \text{if trt 2} \\ 0 & \text{if other} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if trt 3} \\ 0 & \text{if other} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if trt 4} \\ 0 & \text{if other} \end{cases}$$

In this case, when the coefficients for x_1 , x_2 and x_3 are all 0, then the treatment level must be 1 (‘other’). Our regression equation would stay the same

$$Y_i = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \epsilon_i$$

but now, parameters β_2 , β_3 and β_4 refer to treatment levels 2, 3, and 4, respectively. If $x_1 = x_2 = x_3$, then β_1 refers to treatment level 1.

What is important to note here is that in either case, one of the levels is specified by the intercept (i.e., β_1). This level is referred to as the ‘control’ or ‘reference’ level. In this design, then, the other levels ($\beta_2 \rightarrow \beta_4$) are ‘offsets’ from this reference (control) level (i.e., the other β terms represent the magnitude that a particular level of the treatment differs from the control). We will discuss this and related issues in much more detail later.

end sidebar

From this step, it is fairly straightforward to derive the design matrix (so-called because it fully represents the design of the analysis). The design matrix is simply a matrix showing the structure of the ‘dummy’ coding variables in the analysis. Because there are 4 parameters being estimated in the equation (β_1 , β_2 , β_3 and β_4), each corresponding to the 4 levels of the main effect, then the design matrix will be a (4×4) square matrix.

To help construct the design matrix, we can decompose the general regression equation for this analysis (above) into n regression equations, where n is the number of parameters in the regression equation (i.e., the number of levels of the main effect; $n = 4$).

treatment	equation
1	$Y_i = \beta_1(1) + \beta_2(1) + \beta_3(0) + \beta_4(0)$
2	$Y_i = \beta_1(1) + \beta_2(0) + \beta_3(1) + \beta_4(0)$
3	$Y_i = \beta_1(1) + \beta_2(0) + \beta_3(0) + \beta_4(1)$
4	$Y_i = \beta_1(1) + \beta_2(0) + \beta_3(0) + \beta_4(0)$

The design matrix \mathbf{X} simply corresponds to the matrix of the coefficient multipliers (in bold) in these equations.

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

While this seems logical enough, there are, in fact, a number of alternative parameterizations of the design matrix, each of which yields the same ‘model fit’, but which have different interpretations.

For example, all 6 of the following design matrices (\mathbf{X}_1 , \mathbf{X}_2 and \mathbf{X}_3) give equivalent model fits for our example problem:

$$\begin{aligned} \mathbf{X}_1 &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} & \mathbf{X}_2 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} & \mathbf{X}_3 &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \\ \mathbf{X}_4 &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} & \mathbf{X}_5 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \mathbf{X}_6 &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \end{aligned}$$

\mathbf{X}_1 (above) is the design matrix we derived previously; we estimate an intercept term for the last ‘treatment’ level (4), and then an additional ‘treatment’ effect for ‘treatment’ levels 1, 2 and 3. Matrices $\mathbf{X}_2 \rightarrow \mathbf{X}_4$ are based on the same underlying idea, except that the intercept specifies a different ‘reference’ level in each case (see preceding –sidebar–). For example, in \mathbf{X}_2 , the intercept corresponds to treatment level 1. In \mathbf{X}_3 , the intercept corresponds to treatment level 2. And, in \mathbf{X}_4 , the intercept corresponds to treatment level 4.

In \mathbf{X}_5 is an *identity* design matrix. Here, each row corresponds to a parameter, and each column corresponds to a parameter. Thus, each parameter represents a treatment estimate directly, not as an ‘offset’ (deviation) from the ‘control’ or ‘reference’ (i.e., the intercept).

In \mathbf{X}_6 , we estimate a mean parameter among treatment levels, and then an ‘offset’ for each of the 4 levels; the first column corresponds to the mean treatment value, and the remaining columns provide the treatment effects.

We’ll consider these different design matrices later in the chapter. Note that the choice of the structure of the design matrix doesn’t affect the estimates of the parameters (ϕ , or p , for example) – but it does change how estimates of the individual slope parameters in the linear model are interpreted. We will see many examples of this later in the chapter.

Perhaps the most important thing to remember in considering design matrices is that the number of rows corresponds to the number of parameters in your PIMs, whereas the number of columns corresponds to the number of these parameters you are trying to individually estimate. As we will see in the next section, this distinction becomes important when fitting models where parameters are constrained to be functions of 1 or more effects.

Finally, a more complex example, using 2 groups (say, males and females), with multiple levels of a treatment within group (i.e., within sex). This example is clearly analogous to a 2-way ANOVA, with 2 main ‘effects’ (treatment, and sex). Again, assume there are 4 possible treatment levels. The response

variable Y can be decomposed as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where α_i is the sex (group) effect, β_j is the treatment effect, and $(\alpha\beta)_{ij}$ is the interaction of the two.

The corresponding regression equation would be:

$$Y_{ij} = \beta_1 + \beta_2(\text{SEX}) + \beta_3(t_1) + \beta_4(t_2) + \beta_5(t_3) \\ + \beta_6(\text{SEX} \cdot t_1) + \beta_7(\text{SEX} \cdot t_2) + \beta_8(\text{SEX} \cdot t_3) + \epsilon$$

If we derive the design matrix directly from this expression, then we see that we have 8 rows: 2 levels for SEX (male or female) multiplied by 4 treatment levels within sex (remember, $(n - 1) = 3$ columns). The design matrix \mathbf{X} (shown below) would also have 8 columns, corresponding to the intercept, the SEX (group effect), and the treatment and interaction terms, respectively

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

The first column represents the intercept, the second column the group (SEX) effect (1=male, 0=female; i.e., the additive effect of males-females), columns 3-5 represent the treatment effect ($t_1 \rightarrow t_3$), and columns 6-8 represent the interactions of SEX (male) and treatment. Why male, and not female? It depends on the coding – in this case, we're using '0' to represent females, and thus the interaction columns have non-zero elements for males only.

Suppose, for example, rather than the full model (with interactions), you wanted to fit the additive model consisting simply of the 2 main effects (no interaction term):

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

which, in regression form, is

$$Y_{ij} = \beta_1 + \beta_2(\text{SEX}) + \beta_3(t_1) + \beta_4(t_2) + \beta_5(t_3) + \epsilon$$

Using the design matrix \mathbf{X} (above), this is easily accomplished by simply deleting the columns corresponding to the interaction terms:

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

Got it? As we work through this chapter, we’ll come back to the concept of a ‘linear model’ and the ‘design matrix’ with considerable frequency, but hopefully you have the basic idea. In the examples we will explore in this chapter, you will learn the basic steps of creating these linear ‘dummy variable’ models, design matrices, and how to use them with **MARK** to test a variety of hypotheses.

The only thing we now need to consider is – how can we use ‘regression models’ for analysis of mark-recapture data, since both survival and recapture are not ‘normal’ response variables – normal in the sense that they are both constrained to be values from $0 \rightarrow 1$? If you simply regressed ‘live=1/dead=0’ or ‘seen=1, not seen=0’ on some set of explanatory variables x , it is quite conceivable that for some values of x the estimates value of survival or recapture would be > 1 or < 0 , which clearly can’t be correct! However, we clearly want to be able to bring the full power of ANOVA-type analyses to bear on capture-recapture studies.

As mentioned earlier in this chapter, the way around this problem is to transform the probability of survival or recapture, such that the transformed probabilities have been mapped from $[0, 1]$ to $[-\infty, +\infty]$, which is of course the ‘assumption’ for normal linear regression models. To accomplish this, **MARK** uses a link function (see the following ‘sidebar’ for more general background on link functions). In fact, **MARK** allows you to choose among a number of different link functions (some of which are more appropriate for certain types of analyses than others). The default the sin link, which has very good properties for analyses that use what is known as the ‘identity matrix’ (much more on this matrix in a minute. . .). For models which don’t use the identity matrix (such as constrained models), the logit link function is preferred (this is discussed later on in this chapter). Using these transformed probabilities, we can use linear regression models analogous to the one we just considered in the skull circumference example. We will now consider a simple example in detail, based on live encounter data from the European Dipper, to demonstrate how linear models are constructed using **MARK**.

begin sidebar

What is a link function?

In the context of analysis of data from marked individuals, a link function is a transformation of probability such that the transformed probability is mapped from $[0, 1]$ to $[-\infty, +\infty]$. For example, suppose you want to express a dichotomous (i.e., binary) response variable Y (e.g., survival or recapture) as a function of 1 or more explanatory variables. Let $Y = 1$ if alive or present; otherwise $Y = 0$. Let x be a vector of explanatory variables, and $p = \Pr(Y = 1 \mid x)$ is the probability of the response variable you want to model. We can construct a linear function of this probability by using a certain type of transform of the probability, p . For example, the logit transformation (one of several transformation or link functions you can use with **MARK**) is given as:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 x$$

where β_1 is the intercept, and β_2 is the vector of slope parameters. Since $\theta = \ln(p/(1-p))$ has inverse $p = e^\theta / (1 + e^\theta) = 1/(1 + e^{-\theta})$, then the back-transformed estimate of \hat{p} (i.e., back-transformed to the $[0, 1]$ probability scale) is

$$\hat{p} = \frac{e^{\hat{\beta}_1 + \hat{\beta}_2 x}}{1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x}} = \frac{1}{1 + e^{-\hat{\beta}_1 - \hat{\beta}_2 x}}$$

In other words, we can express the probability of the event (survival or recapture) as a linear function of a vector of explanatory variables. The logit (or logistic) model is a special case of a more general class of linear models where a function $f = f(m)$ of the mean of any arbitrary response variable is assumed to be linearly related to the vector of explanatory variables. The function f is the ‘link’ between the random component of the model (the response variable) and the fixed component (the

explanatory variables). For this reason, the function $f(m)$ is often referred to as a ‘link function’.

MARK allows you to choose among a number of different link functions (we will discuss the various link functions later in this chapter), some of which are more appropriate for certain types of analysis than others. **MARK** estimates the intercept and vector of the slope parameters, using the specified link, and then reconstitutes the values of the parameter from the values of the explanatory variables, x . **MARK** does this in 2 steps: (1) first, **MARK** reconstitutes estimates of the parameter from $\hat{\beta}_1, \hat{\beta}_2$ and x , and then (2) **MARK** computes values of the parameter from f using the back transform f^{-1} . There are several examples of this in the text.

end sidebar
