# CHAPTER 7

## PROGRAM MARK: TECHNIQUES FOR MONITORING WILDLIFE POPULATIONS

### 7.1

## FIRST STEPS WITH PROGRAM MARK: LINEAR MODELS

EVAN COOCH
Department of Natural Resources, Cornell University, Ithaca, NY 14853, USA, evan.cooch@cornell.edu

***Abstract:*** Program MARK is the most comprehensive software application available for analyzing data from marked individuals, making full use of the linear models paradigm to provide a robust analytical and statistical framework. I briefly describe how to use MARK to implement some general linear models, focusing on the structure and modification of the design matrix. My intent is to provide a general understanding of the main concepts, as applied generally to analysis of variance and specifically to analysis of data from marked individuals. With any software application as complex and comprehensive as MARK, a great many details exist that are potentially important for a given analysis. Some details are theoretical (determined by the type of data being analyzed and the models being fit), and some mechanical (involving various means to use MARK to accomplish the desired task). A basic understanding of these concepts is essential for using MARK, for either simple or complex analyses.

***Key words:*** analysis of variance, design matrix, linear models, MARK, software

Increasingly, the management and conservation of natural systems requires the statistical analysis of data to draw inferences concerning the populations inhabiting those systems. This increasing use of statistical methods has in part been advanced by the growing availability of computer software, which has made application of various statistical models (including many theoretical ones) both tractable and efficient for wildlife and conservation scientists.

This relationship between software and application has been fundamental to recent increases in data analysis from marked individuals. Whereas the general concepts behind some typical data analyses (e.g., mark–recapture analysis of data from live encounters, recovery analysis of data from dead recoveries, analysis of known-fate telemetry data) are familiar, applying these approaches to data analysis has often lagged because of the difficulty in finding and using appropriate computer software. Further, the available software was often written by specialists, for specialists, and was often difficult to use without significant technical support.

Over the past 15 years, there has been a significant increase in the number of software applications available for analysis of data from marked individuals. However, many of the early efforts (e.g., JOLLY, JOLLYAGE; Pollock et al. 1990)—although significantly easier to use than previously available software (e.g., SURVIV; White 1983)—were extremely limited in their utility. Much software consisted of a set of predefined ("canned") models that the user then selected among; little if any flexibility existed in the software to modify the analysis to suit individual needs and purposes. While program SURVIV arguably provides infinite flexibility (such that SURVIV continues to be a major research tool), the program is not easy to use.

This situation changed significantly with the advent of program SURGE (*SUR*vival *G*eneralized *E*stimation; Pradel and Lebreton 1991, Cooch et al. 1997). SURGE represented a fundamental paradigm shift in both the principle and application of analyzing data from marked individuals (although SURGE is primarily intended for analysis of mark–recapture data, the underlying principles apply in general). SURGE provided the ability to model mark–recapture data within a generalized linear models framework. SURGE uses the concept of linear models (discussed in greater detail below) to allow the user to fit any model of arbitrary design to their data, significantly extending the capabilities of earlier, more restricted applications such as JOLLY and JOLLYAGE. For example, SURGE, and the linear models approach, made it easy to test models where various parameters were constrained to be linear functions of other variables, using 1 of several possible link functions (see Appendix), strictly analogous to familiar analysis of variance (ANOVA) and analysis of covariance (ANCOVA) approaches. The linear models approach, combined with the philosophy of sequential step-wise model fitting (sensu Lebreton et al. 1992), was an important event that precipitated a notable increase in the diversity and sophistication of analyses of mark–recapture datasets. Whereas the basic concepts of linear models and model selection

were not new, SURGE was the first software application to codify these approaches.

The release of SURGE was followed soon by the increasing incorporation of a general linear models approach in other software applications. However, 2 more recent advances have shifted the paradigm still further.

First, the concept of a graphical user interface (GUI) has radically changed the way that most users interact with the computer. SURGE represents a classic command-line application, echoing the days of DOS and UNIX. Although this is not a limitation for many individuals, the procedure creates some significant impediments to learning the software for users accustomed to GUI-based software. The first application of a GUI for data analysis from marked individuals was program SURPH (Smith et al. 1994). In many respects, SURPH represented an extension of SURGE, in terms of the user interface and some of the technical capabilities. Whereas SURGE retained greater flexibility in model fitting (owing to a stricter adherence to a general linear models approach), SURPH added several important diagnostic capabilities, including goodness of fit testing, the ability to model individual covariates, and a variety of options for model selection.

Second, the release of program MARK (White and Burnham 1999) represented a logical extension of the trends initiated by SURGE and SURPH, and MARK significantly surpassed the capabilities of both. These extended capabilities are important in 4 respects. First, while clearly owing a legacy to SURPH in terms of the GUI-based approach, MARK significantly extends the ease-of-use and flexibility of the graphical interface. Second, whereas both SURGE and SURPH were focused on analysis of mark–recapture data, MARK can handle numerous data types, including (but not limited to) mark–recapture, dead recovery, and telemetry data. Third, MARK implements many recent advances in the theory of model selection. Finally, for all data and analysis types, MARK implements a consistent linear models approach to model fitting.

Here, I introduce the linear models approach implemented in MARK. I review the concepts underlying general linear models, then provide a more detailed examination of the concept of a design matrix—which lies at the heart of how linear models are applied using MARK. I do not discuss the actual mechanics of using MARK or the broader considerations of model selection (in either an a priori or a posteriori context); these are explained in considerable detail elsewhere (see http://www.cnr.colostate.edu/~gwhite/mark/mark.htm and associated links). I focus on providing a general understanding of the linear models paradigm and how it is implemented in MARK.

## LINEAR MODELS: A BASIC REVIEW

For users with a background in linear models, much of this presentation may be oversimplified. Texts by Neter et al. (1996) and Kleinbaum et al. (1988) are good technical references.

The basic idea underlying linear models can be stated quite simply: the response variable in any statistical analysis 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 rate or proportion (e.g., survival or recapture rate) that must be transformed prior to analysis using a linear models approach (see Appendix). For the rest of this article, I assume the response variable has been suitably transformed.

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

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where $\mathbf{y}$ is a vector of responses (i.e., a vector of the response variables), $\beta$ 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 $\varepsilon$ is a vector of random error terms.

In cases of analysis of variation of the response variable among different levels of 1 or more classification (i.e., treatment or factor) levels, parameter $\beta$ in vector $\beta$ represents each level of a factor. The elements of $\mathbf{X}$ (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 is generally used only when "1" or "0" are used as the coding variables).

The following simple example illustrates the underlying connection between a linear regression model and ANOVA. Suppose you have collected data on the scutum width of male and female individuals of some insect species. You are interested in whether 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 type of analysis. Recall that for this sort of analysis, any single variate $Y$ (in this case, $Y$ = scutum width), can be decomposed as:

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

Each variate $Y_{ij}$ is the sum of the global mean ($\mu$), the deviation due to the classification factor (sex; $\alpha_i$), and the random error term ($\varepsilon_{ij}$). In this example, with 2 levels of the classification factor (i.e., males and females), we test 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 variate $Y$ would be decomposed as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

In this case, each variate $Y_i$ is the sum of the product of the slope ($\beta_1$) and the variable $x$, the intercept ($\beta_0$), and a random error term ($\varepsilon$). In this case, the hypothesis tested is whether the estimate of the slope is significantly different from 0 ($H_0$: $\beta_1 = 0$). However, what is the variable "$x$"? This is the key to understanding the connection between the regression model and the ANOVA analysis. In the regression formulation, $x$ represents a coding 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 ($\beta_1$) 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.

In matrix notation (above), the regression model for this analysis becomes:

$$\mathbf{y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ . \\ . \\ . \\ Y_{1k} \\ Y_{21} \\ Y_{22} \\ . \\ . \\ . \\ Y_{2k} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ . & . \\ . & . \\ . & . \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ . & . \\ . & . \\ . & . \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ . \\ . \\ . \\ \varepsilon_{1k} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ . \\ . \\ . \\ \varepsilon_{2k} \end{bmatrix} = \mathbf{X}\beta + \varepsilon$$

where there are $K$ individuals in each sex (although a balanced design is not required), and the design matrix $\mathbf{X}$ consists of 2 columns of 0 and 1 dummy variables (the first corresponding to the intercept $\beta_0$, and the second corresponding to the dummy variable coding for a given sex, $\beta_1$). In fact, in this example, if we use "1" to code for males, and "0" to code for females, then the intercept ($\beta_0$) would represent the estimate for female survival, while the $\beta_1$ term would reflect (male survival – female survival) such that $\beta_0 + \beta_1$ = (female) + (male-female) = male survival. The structure of the design matrix is discussed in more detail in the next section.

Models of the form $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ are called linear models because the nonerror part of the expression $\mathbf{X}\beta$ is a linear combination of the parameters (and not specifically because of the relationship of ANOVA to linear regression). Program MARK uses this general linear models approach as the basis for all analysis (data) types available.

## DESIGN MATRIX: THE BASICS

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

I introduce the concept of a design matrix by the following example. Suppose you are doing a typical ANOVA on data with a single classification factor (e.g., treatment). Suppose that 4 levels exist for this factor (e.g., a control and 3 levels of the treatment). You want to test the hypothesis that there is no heterogeneity among treatment levels ($H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$). Recall from the preceding discussion that this problem can be formulated as an applied linear regression problem; in fact, this is precisely how MARK treats the problem—using a linear regression of the appropriately transformed response variable (see Appendix). Also, recall that the regression approach to ANOVA involves using coding for the different levels of the treatment. One coding scheme uses 0/1 dummy variable coding.

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

$$Y_1 = \beta_0 + \beta_1 x_i + \varepsilon_1$$

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).

How would the regression model look 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)? The answer is $k - 1$ (which corresponds to the degrees of freedom for a single-classification ANOVA). Thus, for the present example, $x_1$, $x_2$, and $x_3$ would be:

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

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:

$$Y_1 = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_1$$

In this case, $\beta_0$ is the intercept, while $\beta_1$, $\beta_2$, and $\beta_3$ 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 treatment level corresponds to the case where all 3 slopes are 0. Parameters $\beta_1$, $\beta_2$, and $\beta_3$ refer to treatment levels 1, 2, and 3, respectively. If $x_1 = x_2 = x_3$, then $\beta_0$ refers to treatment level 4. In other words, the intercept corresponds to treatment level 4.

From this step, it is straightforward to derive the design matrix (so-called because it fully represents the design of the analysis). The design matrix is a matrix showing the structure of the dummy coding variables in the analysis. Because there are 4 parameters being estimated in the equation ($\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$)—each corresponding to the 4 levels of the main effect—then the design matrix will be a $(4 \times 4)$ 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 | $Y_i = \beta_0(\mathbf{1}) + \beta_1(\mathbf{1}) + \beta_2(\mathbf{0}) + \beta_3(\mathbf{0})$ |
| 2 | $Y_i = \beta_0(\mathbf{1}) + \beta_1(\mathbf{0}) + \beta_2(\mathbf{1}) + \beta_3(\mathbf{0})$ |
| 3 | $Y_i = \beta_0(\mathbf{1}) + \beta_1(\mathbf{0}) + \beta_2(\mathbf{0}) + \beta_3(\mathbf{1})$ |
| 4 | $Y_i = \beta_0(\mathbf{1}) + \beta_1(\mathbf{0}) + \beta_2(\mathbf{0}) + \beta_3(\mathbf{0})$ |

The design matrix **X** simply corresponds to the matrix of the coefficient multipliers (in bold) in these equations.

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

However, although this seems logical enough, there are a number of alternative parameterizations of the design matrix—each of which yields the same parameter estimates and model fit—but have different interpretations. For example, all 3 of the following design matrices ($\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$) give equivalent results for our example problem:

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{X}_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

$\mathbf{X}_1$ is the design matrix we derived previously; we estimate an intercept term for the last treatment level, and then an additional treatment effect for treatment levels 2, 3, and 4 (with the intercept corresponding to treatment 4). In $\mathbf{X}_2$, each row corresponds to a parameter and each column corresponds to a parameter. Thus, each parameter represents a treatment estimate. In $\mathbf{X}_3$, 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. If you are familiar with linear and matrix algebra, then you might recognize matrix $\mathbf{X}_2$ as an *identity* matrix (1's along the diagonal, 0's along the off-diagonal). Program MARK allows you to specify the type of default design matrix used in a given analysis.

An important fact in design matrices is that the number of rows corresponds to the number of levels of the main effect, whereas the number of columns corresponds to the number of these parameters you are trying to individually estimate. As you 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 uses 2 groups (e.g., males and females) with multiple levels of treatment within group (i.e., within sex). This example is 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:

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

where $\alpha_i$ is the sex (group) effect, $\beta_j$ is the treatment effect, and $(\alpha\beta)_{ij}$ is the interaction of the 2. The corresponding regression equation would be:

$$Y_i = \beta_0 + \beta_1(\text{sex}) + \beta_2(\text{T2}) + \beta_3(\text{T3}) + \beta_4(\text{T4}) + \beta_5(\text{sex.T2}) + \beta_6(\text{sex.T3}) + \beta_1(\text{sex}) + \beta_7(\text{sex.T4}) + \varepsilon$$

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. The design matrix **X** also would 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), columns 3–5 represent the treatment effect, and columns 6–8 represent the interactions of sex and treatment.

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

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

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

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

## FITTING CONSTRAINED MODELS: MODIFYING THE DESIGN MATRIX

As implied in the preceding section, the design matrix can be modified by simply modifying 1 or more columns. However, this is appropriate only in cases where you are eliminating 1 or more main effects, or interactions of main effects. In this section, we discuss the more general issue of modifying the design matrix, as a necessary step in fitting constrained models to allow testing various hypotheses.

Suppose you have conducted the following experiment. Each year, for 5 years, you capture, mark, and release a sample of both male and female individuals from some population of interest. You wish to estimate survival rate of these individuals over the 5 years of the experiment for both sexes. First, given that there are 5 years of the study, only 4 intervals (time between capture occasions) exist for which survival can (in theory) be estimated (year 1 to year 2, year 2 to year 3, year 3 to year 4, and year 4 to year 5); in essence, 4 levels of time. You want to know if survival varies among these time levels, which is analogous to asking if survival varies as a function of some treatment. Thus, time can (in effect) be considered a treatment effect. When considering data analysis from marked individuals, time (days, months, years) can be thought of variously as both a treatment (i.e., classification variable), or as a linear covariate. In this first example, we discuss the former, since it is crucial for understanding how to use linear models with MARK.

Consider first either sex alone. The design matrix **X** for an analysis where time is considered a fixed effect (or treatment) for this example of a 5-year study would look identical to the 1 for our previous example (note we use the default intercept-based design matrix):

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

Suppose you believe that survival is significantly lower over the second and third intervals (i.e., from year 2 to year 3 and from year 3 to year 4) because of some climatic event in those years. Specifically, you believe that

$$\text{survival}_{\text{year 2} \longrightarrow \text{year 3}} = \text{survival}_{\text{year 3} \longrightarrow \text{year 4}}$$

and

$$\text{survival}_{\text{year 1} \longrightarrow \text{year 2}} = \text{survival}_{\text{year 4} \longrightarrow \text{year 5}}$$

This is strictly analogous to an analysis where only 2 levels of the treatment (**A** and **B**, respectively) exist:

$$\text{Yr 1} \xrightarrow{\mathbf{A}} \text{Yr 2} \xrightarrow{\mathbf{B}} \text{Yr 3} \xrightarrow{\mathbf{B}} \text{Yr 4} \xrightarrow{\mathbf{A}} \text{Yr 5}$$

How would the design matrix corresponding to this analysis look? The answer depends on the starting regression model for your analysis. For example, given that there are only 2 possible survival rates to be estimated (**A** or **B**), you might use the expression:

$$Y_i = \beta_0 + \beta_1 xi + \varepsilon_i$$

such that the corresponding design matrix **X** would be:

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

Alternatively, you might use the expression

$$Y_i = \beta_0 + \beta_1 xi + \beta_2 xi + \beta_3 xi\ \varepsilon_i$$

In other words, you have full time-dependent parameterization; 1 parameter for each interval. Obviously, ignoring the **A** or **B** condition for the moment, the appropriate design matrix **X** would be:

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

How can this design matrix be modified (constrained) to reflect the **A** or **B** hypothesis? The key is to remember that in a design matrix, the number of rows corresponds to the number of parameters in the model, whereas the number of columns corresponds to the number of these parameters you are trying to estimate individually. In this example, we have 4 total parameters. However, because of the **A** or **B** constraint, only 2 of them are being estimated (i.e., we want a survival estimate for **A**, and another for **B**). Thus, the design matrix would be reduced to a matrix with 4 rows, but only 2 columns:

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

which is equivalent to:

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

Again, the distinction in this example depends on the number of parameters included in the linear regression model. However, note that in both cases, only 2 parameters are being estimated, even though the design matrices differ in the number of rows (4 vs. 2); the number of columns is the same in both cases.

Now, consider the additional complication of considering both sexes simultaneously. Clearly, this is equivalent to the 2-way ANOVA noted above, with 2 main effects (in this case the time treatment, and sex). There are 4 time levels, and 2 levels of sex. The response variable *Y* can be decomposed as:

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

where $\alpha_i$ is the sex (group) effect, $\beta_j$ is the time effect, and $(\alpha\beta)_{ij}$ is the interaction of the 2 effects. The corresponding regression equation would be:

$$Y_i = \beta_0 + \beta_1(\text{sex}) + \beta_2(\text{T1}) + \beta_3(\text{T2}) + \beta_4(\text{T3}) + \beta_5(\text{sex.T1}) + \beta_6(\text{sex.T2}) + \beta_7(\text{sex.T3}) + \varepsilon$$

where T*x* refers to the different time intervals. As above, 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 time levels (intervals between occasions) within sex. Because we are constraining survival to be a function of some climatic event (the A or B effect noted above), the design matrix **X** would have 4 columns, corresponding to the intercept, the sex (group effect), the climatic event term, and the interaction of the sex and climate terms, respectively:

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

For the final example, instead of constraining the parameter to be a function of some discrete climatic variable, we constrain the parameter to be a linear function of time. To code for a linear trend, you need to write a series of increasing (or decreasing) numbers, 1 through *n* (where *n* is the number of occasions to which you want to fit the "trend"). You don't have to start with the number 1, but you do need to use the sequence {starting value} + 1, {starting value} + 2, and so on. If we consider the case of 2 sexes, and 5 occasions (for time intervals), then the appropriate design matrix **X** would be:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 1 & 1 & 3 & 0 \\ 1 & 1 & 4 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 4 & 0 \end{bmatrix}$$

Clearly, this design matrix corresponds to ANCOVA, where variation in the response variable is modeled as a linear function of sex (the discrete classification variable) and a linear covariate (in this case, time).

## SUMMARY

The design matrix lies at the heart of understanding linear models and how they are applied using MARK. I have attempted to provide a basic review of the concepts underlying general linear models, focusing on the derivation of the design matrix. Using MARK to analyze data from marked individuals involves several steps: specifying the data type, the number of treatment or classification groups, and the parameter structure of the model. These in turn determine the structure of the design matrix. Virtually all analyses made using MARK involve application of the design matrix. The GUI-based interface to MARK makes it easy to modify the elements of the design matrix to build constrained or alternate models. For example, MARK has a number of ways to specify the starting structure of the design matrix. It also has a variety of matrix manipulation functions that are available to the user via a series of menu selections, essentially a spreadsheet paradigm. However, understanding the logic of linear models and the construction of the design matrix are essential to successfully using MARK as an analysis tool.

*Acknowledgments*. I thank G. C. White for inviting this article, and (obviously) for his superb efforts in developing MARK. I also thank W. L. Kendall and G. C. White for comments on a draft of this manuscript.

## LITERATURE CITED

COOCH, E. G., R. PRADEL, AND N. NUR. 1997. A practical guide to mark–recapture analysis using SURGE. Second edition. CEFE/CNRS, Montpelier, France.

KLEINBAUM, D. G., L. L. KUPPER, AND K. E. MÜLLER. 1988. Applied regression analysis and other multivariable methods. Second edition. P.W.S.-Kent, Boston, Massachusetts, USA.

LEBRETON, J-D., K. P. BURNHAM, J. CLOBERT, AND D. R. ANDERSON. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs 62:67–118.

NETER, J., W. WASSERMAN, AND M. H. KUTNER. 1996. Applied linear statistical models. Fourth edition. McGraw-Hill, New York, New York, USA.

POLLOCK, K. H., J. D. NICHOLS, C. BROWNIE, AND J. E. HINES. 1990. Statistical inference for capture–recapture experiments. Wildlife Monographs 107:1–97.

PRADEL, R., AND J.-D. LEBRETON. 1991. User's manual for program SURGE. Version 4.1. CEPE/CNRS, Montpelier, France.

SMITH, S. G., J. R. SKALSKI, W. SCHLECHTE, A. HOFFMAN, AND V. CASSEN. 1994. SURPH.1 manual. Bonneville Power Administration, Portland, Oregon, USA.

WHITE, G. C. 1983. Numerical estimation of survival rates from band-recovery and biotelemetry data. Journal of Wildlife Management 47:716–728.

———, AND K. P. BURNHAM. 1999. Program MARK: survival estimation from populations of marked individuals. Bird Study 46 (supplement):S120–139.

## APPENDIX

### Link Functions

Under a general linear models approach, variation in the response variable (which will reflect the data type used in the analysis) is modeled as a linear function of 1 or more explanatory variables. The only major distinction between typical regression analyses and analysis of data from marked individuals is that rates (survival, movement, resight) are not normal response variables, in the sense that they are constrained to be values from $0 \longrightarrow 1$. If you simply regressed "live = 1, dead = 0" or "seen = 1, not seen = 0" on some set of explanatory variables $\boldsymbol{x}$, then it is quite conceivable that for some values of $x$, the estimates of the particular rate could be >1 or <0, which are clearly impossible.

In general, the solution to this problem is to transform the 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 $\boldsymbol{x}$ be a vector of explanatory variables, and $p = \Pr(Y = 1|\boldsymbol{x})$ 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 (1 of a several transformation or link functions you can use with MARK; see below) is given as:

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

where $\alpha$ is the intercept, $\beta$ is the vector of slope parameters, and

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

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 function $f = f(\mu)$ 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 $f(\mu)$ is often referred to as a link function. 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 link is the sin link, which has good properties for analyses that use what is known as the identity matrix (see text). For models that do not use the identity matrix (such as constrained models), the logit link is preferred.

Program 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, $\boldsymbol{x}$. Program MARK does this in 2 steps: (1) first, MARK reconstitutes estimates of the parameter from $\alpha$, $\beta$, and $\boldsymbol{x}$, and then (2) MARK computes values of the parameter from $f$ using the back transform $f^{-1}$.