CHAPTER 2. SPECIFYING THE STRUCTURE OF MULTIVARIATE GENERAL LINEAR MODELS

The transition from the scalar version of the univariate linear model to the univariate model expressed in matrix algebraic terms is given in Chapter 1 (see Equations 1.2 and 1.3). The univariate linear model is readily generalized to the multivariate model with $p > 1$ response variables by augmenting the orders of $\mathbf{Y}$, $\mathbf{B}$, and $\mathbf{E}$ to accommodate the additional columns of dependent variables, the added columns of regression coefficients associated with each dependent variable, and the additional columns of the disturbances associated with each Y variable in the matrix of errors. To specify the multivariate model, we write

$$\mathbf{Y}_{(n \times p)} = \mathbf{X}_{(n \times q+1)}\mathbf{B}_{(q+1 \times p)} + \mathbf{E}_{(n \times p)}. \qquad [2.1]$$

In this chapter, we will define the elements of these matrices and discuss both the statistical and the substantive ideas necessary to specify the multivariate model that must accommodate multiple columns of $\mathbf{Y}$, $\mathbf{B}$ and $\mathbf{E}$. The order of these three matrices is one key feature of the specification that differentiates multivariate from univariate models. Conversely, the design matrix $\mathbf{X}$, in all of its possible variations, will be identical to the comparable univariate design matrix—we need only develop the mechanism for coping with the multiplicities of dependent variables, parameter estimates, and disturbances that characterize multivariate linear models.

Specifying the multivariate linear model involves at least two discrete, but related, activities:

- Choosing reliable and valid criterion and predictor variables based on theoretical explanations of their hypothesized relationships, including their direction, magnitude, and conceptual mechanism (see Jaccard & Jacoby, 2010, for a discussion of building conceptual theoretical models), and
- Specifying the mathematical model that is consistent with these theoretical arguments.

In this chapter, we introduce methods for specifying the mathematical form of the multivariate model, discuss several different specifications of the design matrix $\mathbf{X}$, and introduce the numerical examples that will be used to illustrate these developments in subsequent chapters.

The Mathematical Specification of the Model

The mathematical specification of the multivariate linear model begins with the definitions of the four matrices of Equation 2.1 that denote a truly multivariate problem if the number of criterion variables (p) is greater than 1. Multivariate models are written in matrix terms, and following the usual conventions[1] we denote the *order* of the matrix by designating its dimensions by reference to the number of rows and columns in the matrix. The intersection of any row and any column defines a specific *element* of the matrix; Y_{23}, for example, denotes the observation of the second row and the third column of **Y**. Letting n denote the number of observations and p denote the number of dependent variables in a model, then the $(n \times p)$ dependent variable matrix $\mathbf{Y}_{(n \times p)}$ denotes a matrix of n rows and p columns. An expanded version of all such **Y** matrices will therefore have a similar general form in which the order of **Y** and all of its elements can be readily identified,

$$\mathbf{Y}_{(n \times p)} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{12} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix}.$$

Similarly, the explanatory variables of the model are contained in the *design matrix*, $\mathbf{X}_{(n \times q+1)}$ in which the order of the matrix is defined by the n rows and the $q+1$ column vectors consisting of the q predictor measures $(X_1, X_2, \ldots X_q)$ and the unit column vector $X_0 \equiv 1$ as previously defined for estimating the model intercept. The design matrix will have a general form of

$$\mathbf{X}_{(n \times \mathrm{q}+1)} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{12} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix}.$$

The matrix of model parameters **B** of Equation 2.1 differs substantially from the univariate model of Equation 1.3. Multiple dependent variables are accompanied by multiple columns of **B** to accommodate all of the *Y*-*X* relationships. The order of **B** is governed by the $q+1$ columns of **X** and the

[1]We assume some familiarity with matrix terminology and matrix algebraic procedures. Detailed coverage is given in Namboodiri (1984) and Schott (1997); succinct coverage relevant to regression analysis is given in Draper and Smith (1998, Chap. 4).

p columns of **Y**; $\mathbf{B}_{(q+1\times p)}$ defines the matrix of parameters in the population model that must be estimated as part of the analysis. The rows of **B** correspond to the predictor variables $X_0, X_1, X_2, \ldots X_q$ and the columns represent the response variables $Y_1, Y_2, \ldots Y_p$,

$$\mathbf{B}_{(q+1\times p)} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix}.$$

In Equation 2.1, the matrix product $\mathbf{X}_{(n\times q+1)}\mathbf{B}_{(q+1\times p)}$ conforms with respect to multiplication, and the order of the product $\mathbf{XB}_{(n\times p)}$ is determined by the number of rows of **X** and the number of columns of **B**. Following the row-by-column rules for matrix multiplication results in a product matrix that contains the weighted linear combinations of **XB** for each of the variables in **Y** across all of the *n* observations:

$$\mathbf{XB}_{(n\times p)} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{12} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix}.$$

The additive equality expressed in Equation 2.1 is satisfied since the order of $\mathbf{XB}_{(n\times p)}$ conforms to the order of $\mathbf{E}_{(n\times p)}$, which in turn conforms to the order of $\mathbf{Y}_{(n\times p)}$. Using these results, the expanded matrix version of the full multivariate linear model for the variables **Y**, **X**, **B**, and **E** would appear as

$$\begin{bmatrix} Y_{11} & Y_{12} & \cdots & X_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{01} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{01} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{01} & \cdots & \beta_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{01} & \cdots & \beta_{qp} \end{bmatrix}$$

$$+ \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix}.$$

All the multivariate models to be covered in this volume will be specified by mathematical models consistent with Equation 2.1. The number of units of observation (cases, participants) and the number of the variables in the response matrix $\mathbf{Y}_{(n \times p)}$ and design matrix $\mathbf{X}_{(n \times q+1)}$ determine the initial specification of the model. The remaining aspects of the model specification rest on theoretical and conceptual arguments and will also depend on design considerations (e.g., multivariate multiple regression [MMR] or multivariate analysis of variance [MANOVA]) that will dictate the nature of the vectors of the design matrix $\mathbf{X}_{(n \times q+1)}$.

Defining the Substantive Roles of Criterion and Predictor Variables

The specification of the model in multivariate analysis is partly nonmathematical, and it is best that there be clear reasons and careful definitions for inclusion of both dependent and explanatory variables. Theoretical considerations are paramount in this endeavor. Since theoretical arguments are project specific, we attempt to lay out briefly the conceptual arguments that underlie each of the examples introduced at the end of this chapter. More extensive advice on this important aspect of model specification is given in Jaccard and Jacoby (2010). Beyond the theoretical and conceptual arguments that dictate the choice of response and explanatory variables, there are four general considerations and decision points that apply across all projects that require attention prior to data collection and analysis. They include the following:

- Measurement level of the *Y* variables
- Measurement of the *X*-variables; either continuously distributed, categorical, or both
- Experimental status of the *X*-variables; either manipulated or observed
- Purpose of the *X*-variables; theoretical substance or control of confounding

The first consideration is the nature of the dependent variables. In this volume, we deal exclusively with continuously distributed dependent variables.[2] Most traditional multivariate analyses have been developed around interval data that can be assumed to be multivariate normal in distribution. Although multivariate models that deal with limited dependent variables such as rank

[2]We refer here to truly continuous variables (an infinite number of possible gradations on the real number line) and discrete variables (a quantitative scale whose integer values are ordinal but on which the gradations between integers is suspect; i.e., 2.5 children). In this volume, we follow the looser tradition of treating both variables as continuous. The difference will be evident by the context of the examples.

transformation analysis (Puri & Sen, 1971), multivariate logistic regression (Glonek & McCullagh, 1995), and cross-classified frequency counts (Zwick & Cramer, 1986) have been proposed, we do not cover them here.

On the predictor variable side of the model several features of the *X*-variables must be considered. These decisions determine how the design matrix will be formulated and how the data are (or have been) collected, how inferences are made from the analysis, and what inferences are justified. The first of these decision points is to decide if the *X*-variable is continuously distributed, discrete, or categorical in nature.[3] A model containing only continuous or discrete explanatory variables is typically classified as a traditional regression model while those models containing only categorical predictor variables are often classified as analysis of variance models. Models with both continuous and categorical predictors in the design matrix have no special designation but are equally possible in the linear model analysis. We present example data sets below that contain continuous explanatory variables, categorical variables (requiring one or more vectors), and combinations of both types of variables.

The second decision that must be considered about the predictor variables is their intended role in inference: Are they theoretically important and require tests of hypotheses, or are they to be treated as covariates for purposes of controlling extraneous variance and potential confounding? A variable's role will usually be clear from a carefully argued theoretical context and is part of the process of specifying the model. The same is true for control variables—their inclusion is based on whether they serve one of two purposes—either they are (1) included because they are known to be substantially correlated with the dependent variables, but not to the theoretically important predictors in the model, such variables included in the design matrix $\mathbf{X}$ can reduce error variance, or (2) they are substantially correlated with both an explanatory variable and one or more response variables and therefore are serious candidates for common-cause, third-variable confounders (Rothman, Greenland, & Lash, 2008). In both instances, their inclusion is intended to be one of control and may or may not require a hypothesis test on the variable.

---

[3]It is necessary to keep in mind the distinction between a variable (say X) and a vector (say $\mathbf{x}_1$). Continuously distributed variables require only a single vector to represent their variability. Categorical variables, such as group membership in multiple groups, require multiple vectors to represent their variability. The *variable* of "treatment" that compares two different treatments with a single control contains three groups and requires two *vectors* to fully represent its variability. Discussions of categorical or qualitative variable coding schemes in linear model analysis can be found in Cohen et al. (2003, Chap. 8).

A final judgment that must be made in the selection of predictor variables in a linear model is related to the experimental versus observational origin of the *X*-variables in the model, namely, what is the underlying source of the variability in the predictor? Is the variability of the explanatory variable under the control of the experimenter or does its variability derive from unknown sources? The first of these sources of variability characterizes the manipulated experiment and the second describes ex post facto observational studies. While it matters little to the mathematical specification of the model, this characteristic of the specification plays an important role in the permissible conclusions that can be drawn from the analysis; the permissible strength of causal conclusions that can be attributed to the results of an analysis often hinge on this distinction (Morgan & Winship, 2007).

The Example Data and Specification of the Models

Throughout the remainder of this volume, we use several numerical examples to illustrate a variety of multivariate linear model analyses. All the examples use continuously distributed interval level–dependent variables. The first and second data sets are used in Chapter 3 to introduce the estimation of the parameters in the multivariate general linear model. They will also be used as running examples to illustrate results on multivariate measures of strength of association (Chapter 4), multivariate test statistics (Chapter 4), and the multivariate general linear hypothesis testing procedure (Chapter 5). The third data set is used to illustrate MANOVA models, including a single-classification MANOVA and a 3×2 factorial MANOVA with two main effects and their interaction (Chapter 6). The first and second data sets are also used to illustrate the recovery of two of the four multivariate test statistics from only univariate quantities (Chapter 4) and to illustrate the details of canonical correlation analysis (CCA) that subsumes all the models dealt with in this volume (Chapter 7). The examples are drawn from several disciplines, including personnel psychology, anthropology, environmental epidemiology, and neuropsychology. To set the stage of model specification, the conceptual basis of each example data set is described below along with summary descriptive statistics. The specification of the analytic models appropriate for each of the examples will be a central part of subsequent chapters. The models to be specified will include MMR, MANOVA, and CCA.

Example 1: Personality and Success in the Job Application Process (MMR, CCA)

Caldwell and Burger (1998) conducted an observational study of 99 college students nearing the completion of their studies and who were anticipating entering the employment market. Individual differences on personality

dimensions are thought to be among the many factors that are important in achieving a successful outcome to the job application and interviewing process. Three dimensions of personality drawn from the Five-Factor model of personality (Costa & McRae, 2000)[4] are used here to illustrate the estimation of the parameters and tests of hypotheses of an MMR model with four response variables: background preparation for the interviews, social preparation for the interviews, the number of follow-up interviews achieved, and the number of offers of employment received. For three predictor variables of Neuroticism, Extraversion, and Conscientiousness, their defining characteristics (facets) provide the conceptual bases for the predictions. The personality variable of Neuroticism is characterized by anxiety, hostility, depression, self-consciousness, impulsiveness, and vulnerability. It is easy to see how these characteristics might impede both preparation for, and success in, the job-seeking process. On the other hand, Extraversion is characterized by warmth, gregariousness, assertiveness, activity, excitement seeking, and positive emotions—all of which would predict success in the interpersonal aspects of seeking employment. The personality dimension of Conscientiousness is defined by features of competence, order, dutifulness, achievement striving, self-discipline, and deliberation—facets that may well predict variation in the careful preparation for the job interview process that should also be related to success. It can be hypothesized that a significant proportion of the joint variation in the successful outcome variables would be predictable from these personality variables. Caldwell and Burger (1998) give further details of the underlying rationale. The means, standard deviations, and correlations of the Personality–Job Application data are presented in Table 2.1.

---

[4]Caldwell and Burger (1998) present means, standard deviations, and correlations for all five of the Five-Factor personality dimensions. Neuroticism, Extraversion, and Conscientiousness were selected for predictor variables due to their theoretical relevance to the dependent variables. For this example, we generated a set of $n = 99$ fictitious data cases based on the descriptive statistics of Caldwell and Burger (1998, p. 128, Table 2), which exactly reproduced the mean, variance, and correlational structure reported in their manuscript. These fictitious data were used for the illustrative analyses presented here. The individual cases, per se, are not absolutely necessary for the analyses presented in this text. The multivariate analyses reported in this volume can be computed from intermediate statistics (e.g., means, variances, and correlations) by matrix language programs (e.g., SAS IML, SPSS MATRIX, STATA MATRIX) or by some available software packages. Harris (2001, pp. 305–307) gives instructions for multivariate analysis based on means, standard deviations, and correlations using the SPSS MANOVA procedure.

Table 2.1 Means, Standard Deviations, and Correlations for the Personality–Job Application Data

	1	*2*	*3*	*4*	*5*	*6*	*7*
1. Neuroticism	1.000						
2. Extraversion	–.100	1.000					
3. Conscientiousness	–.200	.330	1.000				
4. Background preparation	–.140	–.040	.270	1.000			
5. Social preparation	–.090	.380	.220	.420	1.000		
6. Follow-up interview	–.050	.270	.380	.200	.350	1.000	
7. Offers	–.210	.340	.050	–.140	.240	.410	1.000
Mean	25.62	38.03	39.99	13.34	11.89	0.49	0.38
SD	7.10	6.00	5.98	4.12	4.98	0.45	0.35

Source: From Caldwell & Berger's "Personality Characteristics of Job Applicants and Success in Screening Interviews" (1998), Table 1, p. 128.

Note: $n = 99$, critical values of $r = .195$ (at $\alpha = .05$) and .254 (at $\alpha = .01$).

Example 2: PCB Exposure, Age, Gender, Cardiovascular Disease Risk Factors, and Cognitive Functioning (MMR, CCA)

In certain areas of the United States, there is concern over the possible adverse effects of industrially produced environmental contaminants (e.g., polychlorinated biphenyls [PCBs]) on public health—both physical and psychological (Carpenter, 2006). Exposure to PCBs has been hypothesized to adversely affect measures of two related, but conceptually distinct, sets of outcome variables: two major risk factors of cardiovascular disease (physical) and three measures of cognitive functioning (neuropsychological). Because the liver is heavily involved in the body's attempt to remove toxic substances from the bloodstream (PCBs in this example), it has been hypothesized that overactivation of the liver concomitantly leads to an overproduction of cholesterol and triglycerides, which are two known major risk factors for cardiovascular disease (Goncharov et al., 2008). There is also speculation that exposure to PCBs may also have adverse effects on cognitive functioning—such as memory and cognitive flexibility (Lin, Guo, Tsai, Yang, & Guo, 2008). The data of Example 2 consist of six response variables: cholesterol, triglycerides, immediate memory, delayed memory, and two measures of cognitive flexibility (Stroop Color and Stroop Word tests), which are hypothesized to be adversely affected by exposure to PCBs. The multivariate linear model fitted to these data also includes age and gender. It is well known that liver function, memory, and cognitive flexibility are declining functions of age; assessing the effect of age on these dependent variables can provide control of inevitable confounding—since body burden of PCBs is a function of time, age is an obvious confound for any effect of exposure (e.g., $r_{PCBs.age} = .73$). We include gender as an explanatory variable in these models insofar as gender is known to be modestly related to both physical and psychological classes of dependent variable. The descriptive statistics for these example data, based on $n = 262$ cases, is shown in Table 2.2 and will be used to illustrate both MMR analysis and the related CCA.

Example 3: Stature Differences of Indigenous North American Populations (MANOVA)

Auerbach and Ruff (2010) present data on measurements of stature, relative lower limb length, and crural index[5] of skeletal pre-European indigenous

[5]A crural index is the ratio of the length of the tibia to the length of the femur bone. Since the data used in this example are summary statistics, the data are in the aggregate and will show less within group variability than would data based on the original 967 observations. There are both pros and cons (Lubinski & Humphreys, 1996; Robinson, 1950) surrounding the use of aggregate data; such data are more than adequate for our purposes here.

Table 2.2 Means, Standard Deviations, and Correlations for the PCB-CVD-NPSY Data

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1. Age	1.000								
2. Gender	.047	1.000							
3. PCBs	.731	–.130	1.000						
4. Visual memory–I	–.387	–.043	–.364	1.000					
5. Visual memory-D	–.374	.033	–.373	.779	1.000				
6. Stroop word	–.199	.145	–.169	.202	.209	1.000			
7. Stroop color	–.260	.137	–.207	.172	.193	.733	1.000		
8. Cholesterol	.359	.001	.378	–.114	–.142	–.104	–.143	1.000	
9. Triglycerides	.327	–.102	.386	–.070	–.100	–.044	–.080	.561	1.000
Mean	37.89	0.67	0.37	9.93	8.58	91.98	70.30	2.27	2.08
SD	13.47	0.47	0.37	3.29	3.60	23.98	19.72	0.09	0.25

Source: Data has appeared in Goncharov et al. (2008). *Environmental Research, 106,* 226–239; and Haase et al. (2009). *Environmental Research, 109, 73–85.*

Note: PCB = polychlorinated biphenyls; CVD = cardiovascular disease; NPSY = neuropsychological functioning. $n = 262$, critical values of $r = .10$ (at $\alpha = .05$) and .18 (at $\alpha = .01$). Visual memory-I = immediate recall, Visual memory-D = delayed recall. PCBs are log transformed.

populations of North America. Stature information is important in the study of the origin and distribution of pre-European indigenous populations in North America. From 75 different sites in North America, the authors evaluated the three variables on the skeletal remains of 535 males and 432 females. The means of the three dependent variables for males and females at each of the 75 sites provide a total sample of $n = 145$ cases as the data used in this example (see Auerbach & Ruff, 2010, Tables 1 and 2). Auerbach and Ruff have clustered these archeological sites into 11 regions based on natural (geographic) and cultural designations, and further clustered the sites into four geographically distinct groupings: (1) High Latitude Arctic Group, (2) Temperate: West Group, (3) Great Plains Group, and (4) Temperate: East Group. This clustering leads naturally to the specification of a four-group, one-way MANOVA model with three dependent variables.[6] There are at least two ways to describe the research question, formulate hypotheses, and specify the model for MANOVA designs. One common approach is to ask whether the vectors of the three dependent variable means differ simultaneously across the four clusters of sites. Specifying this hypothesis in vector notation with columns defined by $g = 4$ groups, and rows defined $p = 3$ response variables, the null hypothesis of equality of the group mean vectors is written as

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4$$

or in expanded form as

$$\mathrm{H}_0 : \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{31} \end{bmatrix} = \begin{bmatrix} \mu_{12} \\ \mu_{22} \\ \mu_{32} \end{bmatrix} = \begin{bmatrix} \mu_{13} \\ \mu_{23} \\ \mu_{33} \end{bmatrix} = \begin{bmatrix} \mu_{14} \\ \mu_{24} \\ \mu_{24} \end{bmatrix}.$$

An alternate way of characterizing the one-way MANOVA is to ask if there is a significant amount of joint variation in the three dependent variables that can be accounted for by group membership. The linear model of Equation 2.1 can be specified to address this question by adopting one of several methods for coding the design matrix $\mathbf{X}$ to identify levels of a MANOVA factor contained in a categorical (qualitative) variable of group membership. The coding method can be chosen such that the parameter estimates identify differences between the means as reflected in the hypothesis above. This method of solving MANOVA problems is instructive in that the

[6]Auerbach and Ruff combine the temperate groups into a single cluster in their manuscript. We preserve the four-group clustering of regions for this one-way MANOVA example for pedagogical reasons.

output from the linear model most frequently associated with regression analysis (e.g., R^2) is integrated with the information most frequently associated with the classical solution to the analysis of variance (i.e., mean differences). We will undertake a more careful discussion of these equivalences in Chapter 6 and illustrate different methods of coding the design matrix to capture group differences. The means and standard deviations of the three stature response variables classified by the four site clusters of the Auerbach and Ruff data are displayed in Table 2.3. The correlations among the response variables and the grand means are given in Table 2.4.

Table 2.3 Means and (Standard Deviations) for the Four Group, One-Way MANOVA on Stature

	Mean Stature	*Mean Lower Limb Length*	*Mean Crural Index*
Group 1	153.07	48.32	81.61
High Latitude Arctic	(4.90)	(0.73)	(1.37)
Group 2	157.20	48.64	84.87
Temperate: West	(7.43)	(0.82)	(1.28)
Group 3	161.20	49.21	85.64
Great Plains Group	(6.91)	(0.60)	(1.27)
Group 4	161.60	49.12	84.59
Temperate: East	(5.84)	(0.71)	(0.96)

Source: From Auerbach & Ruff (2010). "Stature Estimation Formulae for Indigenous North American Populations", Table 1, pp. 193–194.

Note: $n_1 = 26$, $n_2 = 54$, $n_3 = 14$, $n_4 = 51$.

Table 2.4 Correlations Among the Three Response Variables for the Stature Estimation Data

	Mean Stature	*Mean Lower Limb Length*	*Mean Crural Index*
Mean stature	1.000		
Mean lower limb length	.570	1.000	
Mean crural index	.354	.265	1.000
Mean	158.39	48.81	84.27
SD	7.12	0.80	1.74

Source: From Auerbach & Ruff (2010). "Stature Estimation Formulae for Indigenous North American Populations", Table 2, pp. 195–197.

Note: Means, standard deviations, and correlations are based on the full sample of $n = 145$.

Table 2.5 Means and Standard Deviations for the 2 × 3 Factorial MANOVA

		Factor B								
		b_1			b_2			b_3		
		Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
Factor A	a_1	156.74	45.54	81.71	164.03	49.11	.01	167.65	49.62	85.833
		(3.41)	(0.60)	(1.13)	(4.55)	(0.66)	(1.10)	(2.10)	(0.43)	(1.60)
	a_2	149.39	48.11	81.51	154.17	48.62	84.44	154.74	48.79	85.45
		(3.02)	(0.80)	(1.61)	(5.48)	(0.86)	(1.12)	(1.34)	(0.44)	(9.92)

Note: Y_1 = stature, Y_2 = lower limb length, Y_3 = crural index. Factor A = Sex, Factor B = Geographic Cluster. Standard deviations are in parentheses.

Example 4: A 2 × 3 Factorial MANOVA—Sex by Geographic Group of the Stature Data

In addition to the regional identification of each case in the 75 North American sites, Auerbach and Ruff (2010) also catalogued their data as male or female according to the sex of the skeletal remains. Thus, the 70 sites with complete data (five sites had no females) can be partitioned into male (n = 75) and female (n = 70) groups. When the factor for sex is crossed with a factor of geographic organization—11 regions sorted into three clusters—the data can be organized into a 2 × 3 factorial analysis of variance design. In a factorial design with multiple dependent variables (i.e., stature, lower limb length, and crural index), the primary focus of the MANOVA is on the three sources of influence in the model—the main effects of sex and geographic region and the interaction between the two. While differences in the dependent variables across geographic groups (Factor A) as well as mean differences between genders (Factor B) can be important, the interpretation of the analysis may depend on the A × B interaction. Assessing if vectors of mean differences between levels of one factor are constant across the levels of the second factor is usually a major goal of factorial MANOVA. This 2 × 3 classification of the Auerbach and Ruff data provide the basis of the factorial MANOVA illustration presented in Chapter 6. As was the case with the one-way MANOVA, the design can be characterized in the classical way as tests of differences between mean vectors or as a linear model with predictor vectors designed to contrast vectors of group mean differences. The means and standard deviations of the six cells of this 2 (Sex) × 3 (Geographic Cluster) MANOVA design are displayed in Tables 2.5 and 2.6.

Table 2.6 Cell Sample Sizes for the 2 × 3 Factorial MANOVA

	b_1	b_2	b_3
a_1	$n_{11} = 13$	$n_{11} = 55$	$n_{11} = 7$
a_2	$n_{11} = 13$	$n_{11} = 50$	$n_{11} = 7$