CHAPTER 1. INTRODUCTION AND REVIEW OF UNIVARIATE GENERAL LINEAR MODELS

Few data analytic techniques command a position of greater importance in the social, behavioral, and physical sciences than multiple regression analysis. Exemplary applications can be found in the full range of disciplines, including anthropology (Cardoso & Garcia, 2009), economics (Card, Dobkin, & Maestas, 2009), political science (Baek, 2009), sociology (Arthur, Van Buren, & Del Campo, 2009), and all branches of psychology (Ellis, MacDonald, Lincoln, & Cabral**,** 2008; Pekrun, Elliot, & Maier, 2009).

In each of these disciplines, the purpose of the investigator is to study the relationship between the variables. Fitting regression models to data allows the analyst the ability to account for or explain variation in a criterion variable as a function of one or more predictor variables. The general linear model is an extension of regression models to accommodate both qualitative and quantitative predictor variables. It is widely recognized that multiple regression analysis is a data analytic system that subsumes all linear models (Cohen, 1968), including those that are based on continuously distributed predictor variables (classic regression analysis), those that are based on schemes to accommodate categorical predictors (classic analysis of variance), and those models that are based on any combination of continuous and categorical predictors.[1] Together these models define the general linear model. The regression model is flexible enough to handle many different realizations of predictor variables, including interactions between continuous predictor variables, between categorical predictor variables, and between combinations of continuous and categorical predictor variables. The breadth of coverage of possible analyses afforded by these combinations explains why the technique is so widely used in all scientific disciplines from anthropology to zoology.

In this volume, our goal is to introduce the multivariate version of the general linear model and to illustrate several of its applications. Multivariate models are distinguished by the presence of more than one dependent

[1]Some authors prefer the terms *quantitative* and *qualitative* to describe predictor variables that are continuous or categorical. In this volume, we use the term *continuous* to denote variables whose underlying metric is continuous or discrete, and we use the term *categorical* to denote nominal group structure that has no meaningful underlying metric except to identify categories.

variable that are to be analyzed simultaneously by fitting a single model to the data. Much of the conceptual and statistical basis of multivariate linear model analysis is a direct generalization of univariate regression analysis, which we briefly review in this chapter. This review of univariate strategies for analyzing linear models is intended to set the stage for the remaining chapters. In Chapter 2, we introduce the example data sets to be used throughout along with a discussion of the first step in the general linear model (GLM) analysis of specifying the model. In Chapters 3, 4, and 5, we cover the estimation of parameters of the model, the assessment of goodness of fit of the model along with the related multivariate test statistics, and testing hypotheses on the model. Chapter 6 introduces the linear model solution to the multivariate analysis of variance, and Chapter 7 concludes the volume with an introduction to canonical correlation analysis, which is a linear model that subsumes all of the material of preceding chapters. The overriding goal of the text is to present an integrated view of all these various techniques under a single modeling framework.

Review of Univariate Linear Model Analysis

The main goal of the linear model is to evaluate relationships in order to explain variability in a response variable as a function of some specified model and an error of prediction:

$$\text{Response} = \text{Model} + \text{Error}.$$

In the *univariate* case, regression models are those models that are limited to a single criterion, response, dependent, or outcome variable.[2] Univariate regression models can be expressed mathematically as a regression function,

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \qquad [1.1]$$

[2]We use the terms *dependent*, *criterion*, *response*, and *outcome* interchangeably in this volume to describe the *Y* variable in models. The *X* variables in the model will be interchangeably referred to as predictor, explanatory, or independent variables. These terms appear throughout the literature on regression analysis. Some authors prefer to reserve the term *dependent variable* to experimental designs with manipulated conditions.

for a simple model with a single predictor variable. For a more complex model with multiple predictors, we may write[3]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \varepsilon. \qquad [1.2]$$

In Equations 1.1 and 1.2, Y represents a single column vector response variable that is intended to be explained by the weighted linear combination of regression coefficients, $\beta_0, \beta_1, \ldots, \beta_q$, and explanatory variables, $X_1, X_2, \ldots, X_q$, and includes a disturbance or error term ε, which captures all other sources of variability, both systematic and random, that are responsible for variation in Y. The X_j explanatory variables, $j = 1, 2, \ldots, q$, can be either continuous or categorical.[4] Many contemporary textbooks emphasize this integrative linear model approach to both regression analysis and the analysis of variance in the univariate case (see, e.g., Cohen, Cohen, West, & Aiken, 2003; Myers & Well, 2003).

Although we briefly review the basic ideas of univariate regression/linear model analysis in this chapter, our purpose is to set the stage for the analysis of *multivariate* multiple regression/general linear model analysis with continuous and categorical predictor variables—multivariate models can be conceptualized as generalizations of their univariate counterparts. Whereas univariate regression models are defined by their single column vector of Y scores, multivariate models are defined largely by the fact that *more than one dependent variable* is simultaneously included in the model specification. The collection of the explanatory variables, $X_1, X_2, \ldots, X_q$, can be identical for univariate and multivariate models; only the number of Y variables, the number of columns of regression coefficients, and the number of associated disturbance terms, ε, will differ.

As models become more complex, it will be convenient to express the models and their applications in matrix algebraic terms. Although we introduce the basic matrix notation to identify the linear models discussed in this volume, we do not present a full coverage of the topic. A chapter-length coverage of many of the details is given in Draper and Smith (1998,

---

[3]We do not identify the response and explanatory variables Y or X with a subscript to indicate the serial order of the 1st through the nth observations. In this volume, all models are based on the full set of n observations, and the index of summation or multiplication is assumed to be across all n participants.

[4]Coding schemes for categorical variables will be introduced at greater length in later sections.

Chap. 4); textbook-length coverage can be found in Namboodiri (1984) or Schott (1997).

The univariate multiple regression model of Equations 1.1 and 1.2 can be conveniently summarized in matrix notation as[5]

$$\mathbf{y}_{(n\times 1)} = \mathbf{X}_{(n\times q+1)}\ \boldsymbol{\beta}_{(q+1\times 1)} + \boldsymbol{\varepsilon}_{(n\times 1)} \qquad [1.3]$$

in which $\mathbf{y}_{(n\times 1)}$ is a single-column vector whose *dimensions* are noted in the row-by-column subscript. The X_j predictor variables, $j = 1, 2,\ldots, q$, collected in a design matrix, $\mathbf{X}_{(n\times q+1)}$, are the counterpart of the same predictor variables in the univariate model of Equation 1.2, now expressed as a matrix of *order* $(n \times q + 1)$ with n rows identifying each of the $i = 1, 2, \ldots, n$ cases and $q + 1$ columns that capture the predictor variables. The "+1" in the $q + 1$ dimension allows for a unit vector of $X_0 \equiv 1$ ($\equiv$ means "by definition equal to") to estimate the intercept of the model. The vector $\boldsymbol{\beta}$ of Equation 1.3 is a $(q + 1 \times 1)$ column vector of regression coefficients containing one row for each of the $q + 1$ explanatory variables. Expanding Equation 1.3 shows the elements contained in the matrices for a univariate multiple regression model with $q + 1$ predictor variables:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The multivariate multiple regression model is a generalization of Equation 1.3 and would be written as

$$\mathbf{Y}_{(n\times p)} = \mathbf{X}_{(n\times q+1)}\ \mathbf{B}_{(q+1\times p)} + \mathbf{E}_{(n\times p)}. \qquad [1.4]$$

The matrix $\mathbf{Y}_{(n\times p)}$ is a two-dimensional array of numbers in which the rows of the matrix represent all the n observations (subjects, cases) and the columns of the matrix contain the $p > 1$ response variables, Y_k, for $k = 1, 2, \ldots, p$. Hence, the *order* of the matrix $\mathbf{Y}$ is $(n \times p)$. The structure

[5]We use italics to represent scalars (e.g., X, Y, Z, β, ε), boldface lowercase letters to denote row or column vectors (e.g., **a, b, y, x, β, ε**), and boldface uppercase letters to denote matrices (e.g., **X, Y, B, E, Γ**). If a column or row vector is deliberately represented by a matrix symbol, its vector status will be made explicit by the order of the matrix, e.g., $(n \times 1)$ or $(1 \times p)$.

of the design matrix, $\mathbf{X}_{(n\times q+1)}$, does not differ from univariate to multivariate models and is identical to that of Equation 1.3. The matrix $\mathbf{B}_{(q+1\times p)}$ of Equation 1.4 is an augmented collection of regression coefficients, one row for each of the $q + 1$ explanatory variables and p columns to accommodate the multiple response variables. Finally, the matrix $\mathbf{E}_{(n\times p)}$ is a collection of vectors of disturbance terms, one row for each of the n cases on each of the p response variables in the model. Expanding Equation 1.4 reveals the matrix elements that would be contained in the multivariate model,

$$\begin{bmatrix} Y_{11} & Y_{12} & \cdots & X_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{n2} & \cdots & \beta_{qp} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{1} & \varepsilon_{n1} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{2} & \varepsilon_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n1} & \varepsilon_{n} & \varepsilon_{np} \end{bmatrix}.$$

In the succeeding chapters, we will pursue more of the details of structuring the design matrix to accommodate both continuous and categorical predictor variables. For the remainder of this chapter, we set the stage by focusing on a review of univariate linear models.

We assume that the reader has a reasonably good understanding of univariate multiple regression analysis at the level of Cohen et al. (2003) and a similarly good understanding of analysis of variance models at the level of Myers and Well (2003). We also assume an elementary grasp of matrix addition, subtraction, multiplication, and inverse (division). We hope to show that much of multivariate analysis can be seen as a generalization of univariate analysis. Toward that end, we turn now to a review of the univariate regression model in which we introduce four steps of general linear model analysis:

1. Specify the model.
2. Estimate the parameters of the model.
3. Define measures of goodness of fit of the model.
4. Develop methods for testing hypotheses about the model.

Because of space constraints, we do not undertake a discussion of diagnosis of the adequacy of the models that is covered in detail elsewhere (Cohen et al., 2003, Chap. 4).

Specifying the Univariate Regression Model

The dimensions of $\mathbf{Y}_{(n\times p)}$ define the initial distinction between univariate and multivariate models. If the designation of the model includes a single-column vector of scores, then $\mathbf{y}_{(n\times 1)}$ represents the dependent variable as noted in Equation 1.3. Consider a regression model in which $\mathbf{y}_{(n\times 1)}$ is hypothesized to be a function of three predictors—continuously distributed variables X_1 and X_2 and a dichotomous categorical variable X_3. Ultimately data must be collected that conform to the model specifications. To make matters more concrete, let *Y* represent the construct of executive functioning as measured by scores on the Trail Making Test–Part B (TMT-B, Tombaugh, 2004). Neuropsychologists consider the TMT-B to be a measure of higher-order brain function governing the activities of planning, organization, and anticipation. Since executive functioning is a critical cognitive skill, understanding how status on this dimension might vary with advancing age, increasing education, and differences in gender is important. A fictitious data set based on $n = 40$ observations with correlation structure nearly identical to that reported by Tombaugh (2004) specifies a three-predictor model defined in Equation 1.3. The prototypical matrices required to specify this linear model would include the following:

$$\mathbf{y}_{(40\times 1)} = \begin{bmatrix} 72 \\ 115 \\ 117 \\ \vdots \\ 111 \end{bmatrix}, \mathbf{X}_{(40\times 4)} = \begin{bmatrix} 1 & 41 & 13 & 0 \\ 1 & 51 & 18 & 1 \\ 1 & 80 & 14 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 59 & 10 & 0 \end{bmatrix}, \boldsymbol{\beta}_{(4\times 1)} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{40} \end{bmatrix}.$$

In this univariate model, the vector **y** is time to completion of the TMT-B task, X_1 is the participant's age and X_2 is the participant's education, both continuous predictor variables. The vector X_3 is a dummy-coded regressor, representing a categorical variable of gender coded as 1 = female and 0 = male. The vector $X_0 \equiv 1$ is included as the first column of the design matrix to accommodate the model intercept. The means, standard deviations, and correlations for these data are shown in Table 1.1.

Articulating this descriptive information along with writing out the regression model specified in Equation 1.2 or 1.3 are the statistical details required to specify the model.

Table 1.1 Means, Standard Deviations, and Correlations for the TMT-B Data

	TMT-B	*Age*	*Education*	*Gender*
TMT-B	1.000			
Age	.632	1.000		
Education	–.244	–.171	1.000	
Gender	–.046	.014	–.114	1.000
Mean	93.77	58.48	12.60	.45
Standard deviation	32.77	21.68	2.60	.50

Note: $n = 40$. TMT-B = Trail Making Test–Part B.

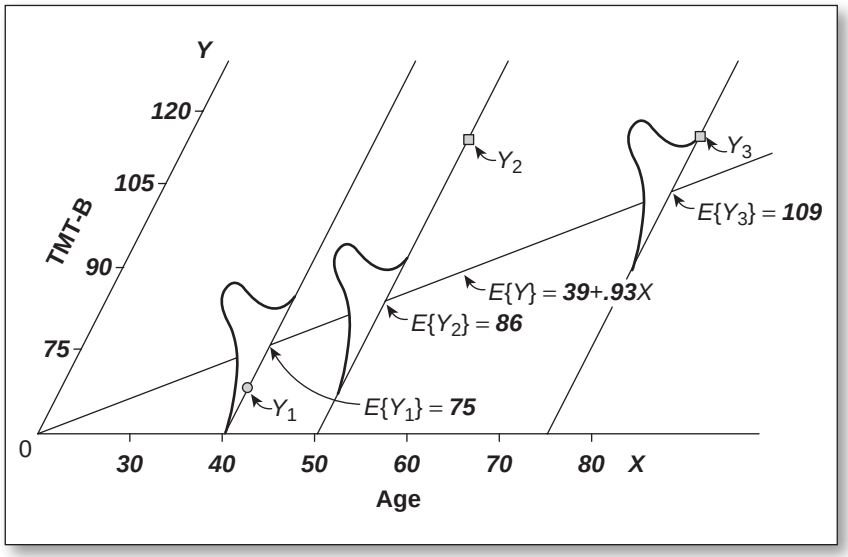
A second important aspect of linear model specification depends heavily on the theory that dictates the mathematical model and provides the substantive explanation of the hypothesized relationship between response and explanatory variables. The theoretical basis of the research often includes the logic used to explain the mechanism through which the Y and X variables are presumed to be associated. These very important details of model specification are context specific and will vary from study to study. While we will endeavor to provide the flavor of such arguments in the examples used to illustrate the procedures here, a full discussion of this aspect of model specification is beyond the scope of this volume. Extensive coverage of this topic is given in Jaccard and Jacoby (2010).

Estimating the Parameters of the Model

The models of Equations 1.1 to 1.4 are population regression functions with parameters of the model defined in the elements of $\boldsymbol{\beta}_{(q+1\times 1)} = (\beta_0, \beta_1, \ldots, \beta_q)$ for univariate models and of $\mathbf{B}_{(q+1\times p)}$ for the multivariate case. For the q-predictor univariate regression model of Equation 1.3, it is known that the long-run expected value of the function for a single criterion variable is given by

$$E(Y|\mathrm{X}) = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q. \qquad [1.5]$$

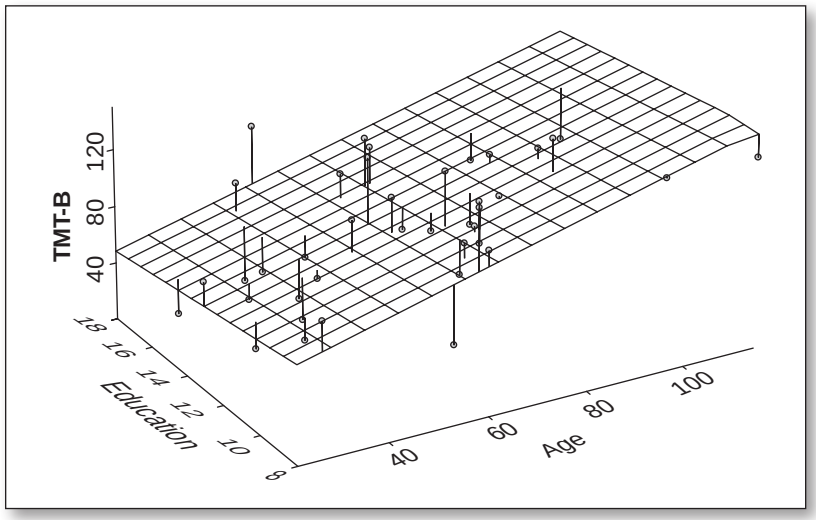
Figure 1.1 The Linear Regression Function With Expected Values (Means) of the Conditional Distributions of Y on X for the Data of Table 1.1



Note: Y_1, Y_2, and Y_3 are illustrative cases.

These expected values are the means of the conditional probability distributions of Y, say $\mu_{(Y|X_j)}$, for each of the values of X_j. The linear model specifying the relationship between Y and X requires that the conditional means of $Y|X$ fall precisely on a straight line defined by the model as illustrated in Figure 1.1 for a single predictor variable. Linear models with two predictors require that the regression surface defined by $\mathbf{X\beta}$ be a two-dimensional plane with partial slopes defining the X axes of the graph as shown in Figure 1.2. For the simple regression model of Equation 1.1, the parameter β_0 defines the expected value of $Y|X = 0$ and β_1 defines the expected rate of change in Y per unit change in X. From the example data of Table 1.1, the regression function of Y = TMT-B on X = Age would appear as in Figure 1.1, in which the conditional means of Y (time to completion of TMT-B) given three values of X = 40, 50, and 75, for example (i.e., $E\{Y_1\}$, $E\{Y_2\}$, $E\{Y_3\}$), lie precisely on the regression line to satisfy the assumption of linearity. Note that the values of the observations Y_1, Y_2, and Y_3 appear in the plane of their respective probability distribution but deviate from their conditional mean. The vector of deviations, $\boldsymbol{\varepsilon} = \boldsymbol{y} - \mathbf{X\beta}$, are the error terms of the regression model in Equation 1.3.

Figure 1.2 Regression of Trail Making Test–Part B on Age and Education



A similar example of a two-predictor model is illustrated in Figure 1.2 by the graph of the relationship between Y = TMT-B, X_1 = age, and X_2 = education for the $n = 40$ sample data descriptively summarized in Table 1.1. The scatterplot reveals a positive relationship between Y and X_1 and a negative relationship between Y and X_2. The population regression function $E(Y|X) = \mathbf{X\beta}$ is defined by the planar surface with partial slopes of β_{X_1} and β_{X2}. The discrepancies between the observations and the model (i.e., the distance between the circles and the plane) are indices of lack of model fit and are captured in the errors of the model, $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbf{X\beta}$.

Thus, all univariate linear models in which the observations are decomposed into model and error components can be written as

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}. \qquad [1.6]$$

The differences between Y and the expected values of Y are the errors of prediction of the model,

$$\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{X})^{6}, \qquad [1.7]$$

[6]The symbols $\widehat{Y}$, $\widehat{\mathbf{y}}$, $\widehat{\mathbf{Y}}$, and $\widehat{\mu}_{(\mathbf{y}|X_1X_2\cdots X_q)}$ will denote sample estimates of the population $E(Y|X)$.

which are illustrated by the distance from each point to the two-dimensional plane in Figure 1.2. The closer all the observed values are to the fitted regression plane, the better the fit of the model to the data.

The criterion of least squares is used to estimate optimal values of $\boldsymbol{\beta}$ such that the discrepancies between the observations and the value predicted by the model are as small as possible. Using the differential calculus, the values of $\boldsymbol{\beta}$ are chosen to minimize the sum of the squared errors of prediction:

$$\Sigma\boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad [1.8]$$

Substituting the sample estimates of the population parameters $\hat{\boldsymbol{\beta}}_{(q+1\times 1)} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_q)$ into Equation 1.8, it can be shown that taking the partial derivatives of $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, setting them to zero, and solving the resulting set of simultaneous equations lead to the optimal solution of the regression coefficients,[7]

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}). \quad [1.9]$$

Applying Equation 1.8 to the example data of Table 1.1 gives the unstandardized parameter estimates of the regression of TMT-B on age, education, and gender,[8]

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 65.69 \\ 0.92 \\ -1.87 \\ -4.68 \end{bmatrix}.$$

---

[7]We will use the diacritic ^ over the symbol to denote a sample estimate of its population parameter.

[8]$(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of the uncorrected raw score sum of squares and cross products matrix (SSCP) of $\mathbf{X}$ and $(\mathbf{X}'\mathbf{Y})$ is the uncorrected raw score sum of cross products (SCP) between $\mathbf{X}$ and $\mathbf{Y}$. The unstandardized regression coefficients of Equation 1.8 are identical to those obtained by mean corrected SSCP and SCP matrices. Details of the relationship between raw score and mean corrected SSCP and SCP matrices are given in Rencher (1998, pp. 269–271).

Interpretations follow the usual rules: Each one year increase in age is accompanied by an increase of approximately $\frac{9}{10}$ of a second to complete the TMT-B task; each additional year of education reduces the time-to-completion of about 2 seconds; and males and females differ by an average of about 4.7 seconds on the timed TMT-B where females show faster performance. The expected time to completion of the TMT-B for a 50-year-old woman with 12 years of education would be estimated at 85 seconds.

It is occasionally useful to reparameterize the regression model to mean zero and unit variance (e.g., $Z_Y, Z_{X_1}, Z_{X_2}, Z_{X3}$) in which the particulars of the regression model in standard score form[9] can be expressed in terms of correlation coefficients. The standard score regression model can be written in scalar and matrix form as

$$Z_Y = \beta_1^* Z_{X_1} + \beta_2^* Z_{X_2} + \cdots + \beta_q^* X_q + \epsilon$$

$$\mathbf{Z_y} = \mathbf{Z_X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \qquad [1.10]$$

with errors of prediction defined as

$$\boldsymbol{\epsilon} = \mathbf{Z_y} - \mathbf{Z_X}\boldsymbol{\beta}^*. \qquad [1.11]$$

The least squares estimates, $\hat{\boldsymbol{\beta}}^*$, of the standardized regression parameters chosen to minimize the sum of squared errors of Equation 1.11 are found by

$$\hat{\boldsymbol{\beta}}^* = \mathbf{R}_{\mathrm{XX}}^{-1}\mathbf{R}_{\mathrm{XY},} \qquad [1.12]$$

where $\mathbf{R}_{XX}$ and $\mathbf{R}_{XY}$ are, respectively, the correlation matrices between predictors and between predictors and criterion.[10] Estimating $\hat{\boldsymbol{\beta}}^*$ for the example data of Table 1.1 yields the fitted model,

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \hat{\beta}_3^* \end{bmatrix} = \begin{bmatrix} 0.61 \\ -0.15 \\ -0.07 \end{bmatrix}.$$

---

[9]The symbol β^* will be used to denote parameters in standard score form with the standardized estimates of the parameters denoted by $\hat{\beta}^*$.

[10] $\mathbf{R}_{XX}$ and $\mathbf{R}_{XY}$ are the sample size–adjusted SSCP and SCP matrices in standard score form.

The usual rules for interpreting standardized coefficients apply; each coefficient represents a $\hat{\beta}_j^*$ standard deviation change in Y per standard deviation change in X_j. There may be little to be gained by interpreting any single standardized regression coefficient in lieu of its unstandardized counterpart, but it is often recommended that standardized coefficients be used if comparative evaluation of the relative influence of predictors is a goal of the analysis (Bring, 1994; Darlington, 1990, pp. 217–218). These recommendations are based on the fact that the absolute value of the unstandardized regression coefficients ($\hat{\beta}_j$) are partly dependent on the scale of measurement, which can differ across predictors while the standardized coefficients ($\hat{\beta}_j^*$) are scale adjusted.[11] For the predictor variables of age and education, the raw regression coefficients suggest that age is a less important predictor than education (ignoring the differences in scale—$SD_{\text{age}} = 21.68, SD_{\text{education}} = 2.60$), whereas the standardized coefficients suggest the opposite relative importance with age being greater than education after adjusting for underlying scale differences. The issue of testing the significance of these differences (i.e., $\hat{\beta}_1$ vs. $\hat{\beta}_2$, and $\hat{\beta}_1^*$ vs. $\hat{\beta}_2^*$) will be shown in a later section to be tests of quite different conceptual hypotheses even if the raw scores are on equal scales, where $SD_{X_1} = SD_{X_2}$.

Assumptions Needed to Justify the Validity of the Least Squares Estimates

There are no assumptions required to justify the least squares estimation of the parameters—that process is purely descriptive. But several important assumptions about the linear model can be introduced at this point. If met, the assumptions provide a degree of confidence in the interpretation of the coefficients as well as justify the validity of the test statistics to be discussed in a later section of this chapter. The assumptions include the following:

- The model is linear; the $E(Y|X)$ lies precisely on a straight line.
- The model is correctly specified; no important variables are omitted from the analysis.

---

[11] Standardized regression coefficients have little meaning for categorical predictor variables. The standard deviation of the numbers used to designate categories of a nominal grouping variable has no meaningful interpretation beyond the ability of the numerals to distinguish categories. In a later section, we note that the standardized version of a dichotomous predictor may have a useful interpretation when involved in a test of relative importance when compared with other predictors in the model.

- The predictor variables X_j are measured without error.
- $E(\varepsilon) = 0$. The errors of the regression model are a random variable with mean zero.
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, for i ≠ j. The errors are assumed to be independent with covariance of zero.
- $V(\varepsilon) = \sigma^2 I_{(n \times n)}$. The variance of the errors is assumed to be a constant. The quantity σ^2 is a population parameter and is estimated in the sample by the mean square error,

$$\hat{\sigma}_2 = \frac{\sum (Y - X\hat{\mathrm{B}})^2}{n - q_f - 1},$$

  where $n - q_f - 1$ denotes the degrees of freedom for error based on q_f predictor variables in the full model.
- $\varepsilon_i \sim N(X\beta, \sigma^2 I)$. The errors of the model are assumed to be normally distributed with mean $X\beta$ and variance $\sigma^2 I$, which provides the connection to the probability distribution that underlies the test statistics applied to the regression coefficients.

More extensive accounts of the assumptions and the diagnosis of their violations can be found in Cohen et al. (2003, Sect. 4.3–4.5).

Partitioning the Sums of Squares and Defining Measures of Goodness of Fit

The strength of the relationship between the criterion and the predictor variables in a linear model is documented by two indices: the sum of squared errors $\left(SS_{ERROR} = \sum \left(Y - X\hat{\beta}\right)^2 = \sum \hat{\varepsilon}_2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\right)$ and the squared multiple correlation coefficient (R^2). To achieve each of these measures requires that the variability in the response variable be partitioned into its constituent parts related to Equation 1.3. The partitioned *SS* is

$$SS_{TOTAL} = SS_{MODEL} + SS_{ERROR}. \qquad [1.13]$$

The estimated vector of errors of the model is given by $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and the sum of squared errors of Equation 1.13 is defined by $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$. As a measure of goodness of fit, $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ has known lower and upper bounds, $0 \le \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \le SS_{TOTAL}$, defining a range from no relationship to perfect relationship. The measure $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ is ambiguous as a measure of strength of association unless SS_{TOTAL} is known. The *mean corrected* total sum of squares is $SS_{TOTAL} = \sum \left(Y - \bar{Y}\right)^2 =$

$\mathbf{y}'\mathbf{y} - \bar{\mathbf{y}}\bar{\mathbf{y}}n$, where $\bar{\mathbf{y}}$ is an $(n \times 1)$ vector of the mean of Y repeated n times. Redefining $\mathbf{y}'\mathbf{y} = (\mathbf{y}'\mathbf{y} - \bar{\mathbf{y}}\bar{\mathbf{y}}n)$ to be the *mean corrected* SS_{TOTAL}, and redefining $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = (\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \bar{\mathbf{y}}\bar{\mathbf{y}}n)$ to represent the *mean corrected* $\mathrm{SS}_{\mathrm{MODEL}}$, the partition of the sums of squares of Equation 1.13 is[12]

$$\mathbf{y}'\mathbf{y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}. \qquad [1.14]$$

It is common practice to rely on the value of R^2, which is scaled to take on values in the interval [0, 1], as an index of goodness of fit. SS_{TOTAL} is the maximum variability available in Y, SS_{ERROR} is the variability in Y that cannot be accounted for by the model, and SS_{MODEL} is that part of the variability in Y that is accounted for by the model. The proportion of the variability in Y that is accounted for by the model, R^2, is the scaled measure of goodness fit and is computed as

$$R^2_{Y \cdot X_1 X_2 \ldots X_q} = 1 - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\mathbf{y}'\mathbf{y}} \qquad [1.15]$$

or more commonly,

$$R^2_{Y \cdot X_1 X_2 \ldots X_q} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}. \qquad [1.16]$$

If the partitioning is done in terms of standard scores, it can be shown that a convenient definition of R^2 is given by

$$R^2_{Y \cdot X_1 X_2 \cdots X_q} = \hat{\beta}^*_1 r_{Y \cdot X1} + \hat{\beta}^*_2 r_{Y \cdot X2} + \cdots + \hat{\beta}^*_q r_{Y \cdot Xq}. \qquad [1.17]$$

For the TMT-B example data of Table 1.1, the mean corrected Total and Model *SS* are $\mathbf{y}'\mathbf{y}$ = 41875.33, $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = 17758.00$. The fit of the model is found to be

$$R^2_{Y \cdot X_1 X_2 X_3} = \frac{17758.00}{41875.33} = .424.$$

---

[12]The uncorrected sum of squares of Y, $\sum Y^2 = \mathbf{y}'\mathbf{y}$, contains both the *SS* associated with the predictor variables ($\beta_1, \beta_2, \ldots \beta_q$) and the *SS* associated with the intercept. The mean corrected *SS*, $\mathbf{y}'\mathbf{y} - \bar{\mathbf{y}}\bar{\mathbf{y}}n$ disaggregates these two quantities. Rencher (1998, Sect. 4.3–4.5) gives details of the relationships between uncorrected and mean-corrected SS.

About 42% of the variation in TMT-B performance is accounted for by age, education, and gender. About 58% of the variation in *Y* remains unexplained and is a function of other unknown sources, both systematic and random.

Full and Restricted Models and Squared Semipartial Correlations

In addition to the full model R^2 based on q_f predictors, it is often of interest to ascertain the proportion of variation in *Y* that is uniquely attributable to X_j adjusted for all remaining *X*-variables. These *squared semipartial correlation coefficients* ($r^2_{Y(X_1|X_2X_3\cdots X_{q_f})}$, $r^2_{Y(X_2|X_1X_3\cdots X_{q_f})}$,…,$r^2_{Y(X_{q_f}|X_1X_2\cdots X_{q_f-1})}$) can be computed from the extra sums of squares approach (Draper & Smith, 1998, pp. 149–160), which requires evaluating the difference between full and restricted model R^2s. Define the *full model* R^2_{full} as the proportion of variation in *Y* accounted for by all the q_f predictors in the model, $X_1, X_2, \ldots, X_{q_f}$. Define a *restricted model* $R^2_{restricted}$ as the proportion of variability in *Y* accounted for by a subset of $q_r < q_f$ predictors, say $X_1, X_2, \ldots, X_{q_r}$. Since the full model R^2_{full} documents the proportion of variability in *Y* accounted for by all the predictors and the restricted model $R^2_{restricted}$ represents the proportion of the variability in *Y* accounted for by the q_r predictors, the difference between the full and restricted model R^2s must represent the unique incremental variation in *Y* accounted for by those predictors that are not contained in the restricted model. The difference $R^2_{full} - R^2_{restricted}$ is the squared semipartial correlation coefficient. Examples of squared semipartial correlations for the TMT-B example data are

$$r^2_{Y(X_1|X_2X_3)} = R^2_{Y\cdot X_1X_2X_3} - R^2_{Y\cdot X_2X_3} = .424 - .065 = .359,$$

$$r^2_{Y(X_2|X_1X_3)} = R^2_{Y\cdot X_1X_2X_3} - R^2_{Y\cdot X_1X_3} = .424 - .403 = .021,$$

$$r^2_{Y(X_3|X_1X_2)} = R^2_{Y\cdot X_1X_2X_3} - R^2_{Y\cdot X_1X_2} = .424 - .419 = .005,$$

$$R^2_{Y(X_1X_2|X_3)} = R^2_{Y\cdot X_1X_2X_3} - R^2_{Y\cdot X_3} = .424 - .002 = .422.$$

About 36% of the variation in TMT-B performance is attributable to age after adjusting for the variance accounted for by education and gender; the variance in TMT-B that is uniquely attributable to education or gender is negligible. The multiple squared semipartial of age and education adjusted for gender appears to be the best prediction model, but it is unclear if this value is a significant improvement over age alone ($r^2_{YX_1} = .6324^2 = .400$) because we know little about the sampling variability that accompanies these models. Methods for assessing statistical

significance and testing hypotheses on contrasts between predictors are reviewed in the next section.

Testing Hypotheses on the Regression Coefficients and R^2s

The trustworthiness of $\hat{\beta}$ or R^2 depends on knowledge of the sampling variability of that statistic and test statistics for evaluating hypotheses on the parameters of the model. The two most common methods include the *F*-test on values of R^2 and the single degree of freedom *t*-test on the model regression coefficient where $t = \sqrt{F}$. A generic *F*-test on df_h and df_e degrees of freedom based on appropriately specified full and restricted models can be defined as

$$F_{(df_h,\, df_e)} = \frac{R^2_{full} - R^2_{restricted}}{1 - R^2_{full}} \cdot \frac{df_e}{df_h}. \qquad [1.18]$$

Let q_f be the number of predictors in the full model (exclusive of the unit vector X_0), let q_r be the number of predictors in the restricted model, and let $df_h = q_f - q_r$ and $df_e = n - q_f - 1$. If it is assumed that $\varepsilon_i \sim N(0,\sigma^2)$, then the test statistic in Equation 1.18 follows the *F* distribution with $q_f - q_r$ and $n - q_f - 1$ degrees of freedom. The numerator of the left-most ratio of the *F*-test is the definition of the squared semipartial correlation. The nature of $R^2_{restricted}$ will be dictated by the hypothesis to be tested since the hypothesis dictates the constraints to be placed on the full model. If the hypothesis on the whole model $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{q_f} = 0$ is desired,[13] the restricted model will contain only β_0 with $R^2_{restricted} = 0$, leading to the numerator $R^2_{full} - R^2_{restricted} = R^2_{full}$. A test of a hypothesis on a single regression coefficient, such as $H_0 : \beta_1 = 0$, would involve $R^2_{restricted} = R^2_{Y \cdot X_2 X_3 \cdots X_{q_f}}$. Hypotheses on any single coefficient, or set of coefficients, can be tested in this manner. Further examples of hypothesis tests involving restrictions placed on the linear model are given in Rindskopf (1984).

For single df_h tests, the *t*-test on the hypothesis $H_0 : \beta_j = k$ is, in common usage,[14]

$$t_{(df_e)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{MSE}{SS_{X_j}}\left(\frac{1}{1 - R^2_{X_j \cdot other}}\right)}}, \qquad [1.19]$$

[13]This is equivalent to the hypothesis $H_0 : \rho^2_{Y \cdot X_1 X_2 \cdots X_{q_f}} = 0$.

[14]The value of *k* need not be hypothesized to be 0; any theoretically defensible value of *k* is permissible.

where

$$MSE = \frac{SS_{ERROR}}{n - q_f - 1},$$

SS_{x_j} is the sum of squares of the predictor variable involved in the test, and

$$\frac{1}{1 - R^2_{X_j \cdot other}}$$

is the variance inflation factor (VIF) that adjusts for the multicollinearity among the predictor variables. For the TMT-B example, the *F*-test on the whole model $R^2 = .424$ is $F_{(3,36)} = 8.84,\ p < .001$. The test of the significance of each of the individual partial regression coefficients for age, education, and gender yielded, respectively, values of $t_{(36)} = 4.74,\ p < .001; t_{(36)} = -1.15,\ p = .258;$ and $t_{(36)} = .57,\ p = .575.$ Only the age variable is uniquely related to TMT-B performance. The *t*-test statistics on the individual coefficients are the $\sqrt{F}$ that would have been obtained by the full- versus restricted-model approach of Equation 1.18. The results of the test of hypotheses on the values of β_j and on the values of their respective partial and semipartial correlations are identical.

The General Linear Hypothesis Test

Although the two methods for testing hypotheses described above are in wide usage, they are special cases of a much more general approach to testing hypotheses in linear models—the general linear hypothesis test. The general linear test is a compact procedure that covers an astonishing array of common and specialized tests of hypotheses in both the univariate and the multivariate linear models.

Assume for the univariate model of Equation 1.3 that we wish to test the hypothesis that all the sample regression coefficients in a full model have been drawn from a population in which all the coefficients, with the exception of the intercept, are simultaneously equal to zero. We can formalize this hypothesis by a linear combination of the parameters specified by the matrix product $\mathbf{L\beta} = \mathbf{0}$. That is,

$$H_0 : \mathbf{L\beta} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{q_f} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{q_f} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \qquad [1.20]$$

The **L** matrix is of order $(c \times q_f + 1)$ whose role is to identify the coefficients of interest in any hypothesis. Other hypotheses might involve only a single-parameter estimate (e.g., $H_0 : \beta_1 = 0$), or some subset of the parameters $\left(\text{e.g., } H_0 : \begin{bmatrix} \beta_1 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right)$. In general, any desired hypothesis can be defined as a product of a vector (or matrix) of *contrast coefficients*, $\mathbf{L}_{(c \times q+1)}$, and the vector of parameters, $\boldsymbol{\beta}_{(q+1 \times 1)}$, from the full model analysis. A more general form of the contrast is possible where the vector **k** can contain zeros (the traditional null hypothesis) or any other vector of theoretically justified nonzero values:

$$\mathbf{L}_{(c \times q+1)} \boldsymbol{\beta}_{(q+1 \times 1)} = \mathbf{k}_{(c \times 1)}. \qquad [1.21]$$

The subscript *c* denotes the number of rows in **L** that will be equivalent to df_h in the associated test statistic. Once the desired hypothesis is specified, we can substitute the estimates of the parameters $\hat{\boldsymbol{\beta}}$ into Equation 1.22 to obtain the sums of squares for the hypothesis:

$$SS_{HYPOTHESIS} = (\mathbf{L}\hat{\boldsymbol{\beta}})' \left(\mathbf{L} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{L}' \right)^{-1} (\mathbf{L}\hat{\boldsymbol{\beta}}) \qquad [1.22]$$

and the $SS_{HYPOTHESIS}$ can be used as the numerator of a familiar version of the *F*-test,

$$F_{(df_h,\, df_e)} = \frac{SS_{HYPOTHESIS}}{SS_{ERROR}} \cdot \frac{df_e}{df_h}. \qquad [1.23]$$

Under the assumption that the errors of the model are normally distributed, *F* will follow the *F* distribution on $df_h = c$ and $df_e = n - q_f - 1$ degrees of freedom.

The Test of the Whole Model Hypothesis $\boldsymbol{\beta_1 = \beta_2 = \beta_3 = 0}$ and $\rho^2_{Y \cdot X_1 X_2 X_3}$

For the TMT-B example data, we found the estimated regression coefficients to be

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 65.69 \\ 0.92 \\ -1.87 \\ -4.68 \end{bmatrix}$$

and we desire a test the hypothesis that parameters for X_1, X_2, and X_3 are simultaneously equal to 0, $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. This statement is also a test of $H_0: \rho^2_{Y \cdot X_1 X_2 X_3} = 0$. The general linear test of the full model hypothesis is given in $\mathbf{L\beta}$

$$\mathbf{L\beta} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

which ignores the intercept. For the contrast matrix **L**, the inverse of the sum of squares and cross-products matrix among the three predictor variables, $\mathbf{X}'\mathbf{X}^{-1}$, and the estimates of the parameters $\hat{\boldsymbol{\beta}}$ the hypothesis SS of Equation 1.22 is

$$SS_{HYPOTHESIS} = \left(\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 65.69 \\ 0.92 \\ -1.87 \\ -4.68 \end{bmatrix} \right)' \left(\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 40 & 2,339 & 504 & 18 \\ 2,339 & 155,103 & 29,097 & 1,059 \\ 504 & 29,097 & 6,614 & 221 \\ 18 & 1059 & 221 & 18 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 65.69 \\ 0.92 \\ -1.87 \\ -4.68 \end{bmatrix} \right)$$

which yields $SS_{HYPOTHESIS} = 17758.00$. The $SS_{HYPOTHESIS}$ is identical to the SS_{MODEL} obtained from $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \bar{\mathbf{y}}'\bar{\mathbf{y}}n$. With $df_h = c = 3$, $df_e = n - q_f - 1 = 36$, and $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = 24117.33$, the *F*-test on the whole model association is found to be

$$F_{(3,16)} = \frac{17758}{24117.33} \cdot \frac{36}{3} = 8.84,\ p = .0002.$$

With $R^2_{Y \cdot X_1 X_2 X_3} = .424$, there is sufficient evidence to reject H_0.

Testing the Individual Contributions of X_1, X_2, and X_3 by the General Linear Test

Hypothesis tests on the individual partial regression coefficients β_1, β_2, and β_3 can be readily tested within the $\mathbf{L\beta = 0}$ framework. For testing a hypothesis on β_1, we specify

$$\mathbf{L\beta} = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \beta_1 = 0 \qquad [1.24]$$

and to specify the null hypothesis on β_2 and on β_3, we employ the vector $\boldsymbol{\beta}$ and the appropriate vectors $\mathbf{L} = [0 \quad 0 \quad 1 \quad 0]$ and $\mathbf{L} = [0 \quad 0 \quad 0 \quad 1]$, respectively. All these hypotheses are tested by substituting $\hat{\boldsymbol{\beta}}$ into Equations 1.22 and 1.23; the results are summarized in Table 1.2.

Table 1.2 General Linear Hypothesis Tests on Individual Partial Regression Coefficients

Hypothesis	$\hat{\beta}$	$\hat{\beta}^*$	$r_{semipartial}$	$F_{(1,16)}$	p
Age: $\beta_1 = 0$	0.919	0.608	.599	22.44	<.001
Education: $\beta_2 = 0$	−1.870	–0.148	–.145	1.32	.258
Gender: $\beta_3 = 0$	-4.683	–0.072	–.072	0.32	.575

In this model, age is the only significant contributor to the prediction of TMT-B. The test statistic on the any unstandardized $\hat{\beta}_j$ is also the test of the significance of the standardized $\hat{\beta}^*$ and the semipartial correlation $r_{Y(X_j|X_1X_2\cdots)}$. The test of hypotheses on sets of predictors is also identical for unstandardized and standardized partial regression coefficients and the multiple semipartial correlations associated with each set. These equivalences no longer hold when more complex hypotheses are tested by the general linear test.

Testing More Complex Hypotheses With the General Linear Test

The general linear hypothesis test is suitable for formulating and testing many complex hypotheses (Draper & Smith, 1998, pp. 217–221;

Rindskopf, 1984). Consider the question of whether age is a *better* predictor of TMT-B performance than is education, both adjusted for gender. This question asks if the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are *significantly different from one another* as expressed in the null hypothesis $H_0 : \beta_1 = \beta_2$. This hypothesis can be specified by the contrast matrix $\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}$ that deletes β_0 and β_3 from $\boldsymbol{\beta}$ and defines the difference between $\beta_1$and β_2. The symbolic contrast of the null hypothesis, $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ is

$$\mathbf{L}\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 - \beta_2 \end{bmatrix} = 0, \qquad [1.25]$$

which gives the basis for evaluating the $SS_{HYPOTHESIS}$ and the numerator of the *F*-test. Substituting the estimates $\hat{\beta}_j$ into Equation 1.22 we find,

$$\mathbf{L}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 65.69 \\ .92 \\ 1.87 \\ -4.68 \end{bmatrix} = \begin{bmatrix} -.95 \end{bmatrix}$$

and $\left(\mathbf{L}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{L}'\right)^{-1} = 259.53$ with $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = 24117.33$. The *F*-test of Equation 1.23 is

$$F_{(1,16)} = \frac{234.71}{24117.33} \cdot \frac{36}{1} = 0.32,\ p = .573.$$

The unstandardized partial slopes of age and education (reverse scored),[15] shown in Figure 1.3, do not differ significantly from one

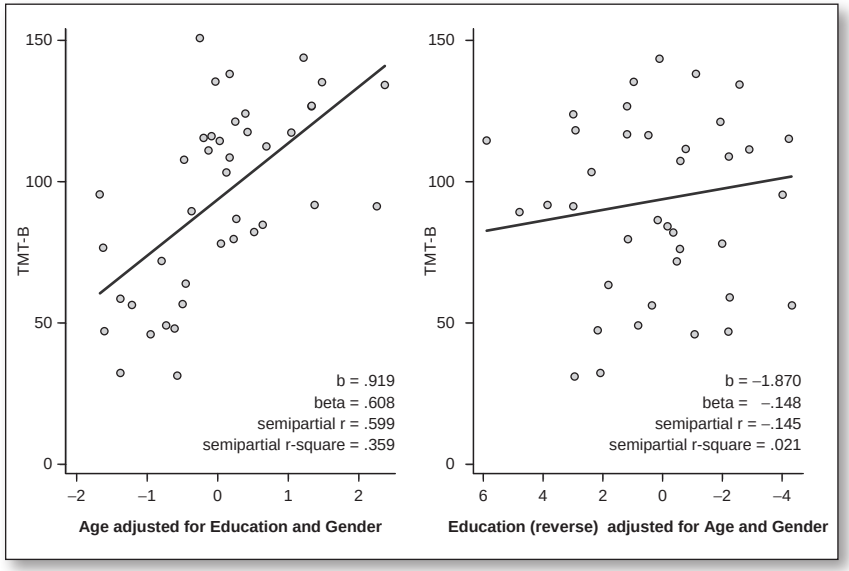
[15]Age has a positive relationship to TMT-B; performance deteriorates with increasing age. Conversely, TMT-B has a negative relationship with increasing years of education. Contrasts between regression coefficients are sensitive to both magnitude and direction and a choice must be made between testing differences in magnitude only, or testing differences in both magnitude and direction. Theoretical considerations based on substance knowledge should be brought to bear to make this choice. For the age versus education comparison illustrated here, only the magnitude of the effect is of interest. Reversing the scoring of the education variable equates the sign of both age and education coefficients; hence the contrast is one of magnitude and not direction. If there is theoretical justification to leave the signs of the regression coefficients in the original scoring of age and education, then a test of both magnitude and direction would result. The *F*-test on this contrast is $F_{(1,16)} = 3.01$, $p = .091$, still a nonsignificant result.

another. Interpreting this lack of a significant difference should be done cautiously. Many authors point out that such a contrast makes sense only if the two variables are measured on the same scale, which is not the case with age (range = 33–105, *SD* = 21.7) and education (range = 8–18, *SD* = 2.6).

The unstandardized partial slopes of Figure 1.3 do not differ significantly from one another, but the squared semipartial correlations suggest that the variance in TMT-B accounted for by age is substantially greater than the variance accounted for by education. The relative rank order of the two predictor variables is opposite when unstandardized slopes and semipartial correlations are used for the ranking, largely due to the differences of scale of the predictor variables.

An alternative test that avoids the issue of inequalities of scale is the general linear test applied to standardized coefficients by testing the hypothesis $H_0 : \beta_1^* - \beta_2^* = 0$.[16] Estimating the parameters by $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ and

Figure 1.3 Comparison of Unstandardized Partial Slopes, Standardized Partial Slopes, and Semipartial Correlations



Note: Education is reverse scored to guarantee a positive slope.

[16]The scoring of the education variable was also reversed in this analysis to constrain the sign of each standardized slope to a positive value. The contrast is therefore a test of the difference in magnitude of semipartial correlations.

performing the same sequence of computations on the standardized variables $Z_Y, Z_{X_1}, Z_{X_2},$ and Z_{X_3} leads to

$$\mathbf{L}\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 0.61 \\ 0.15 \\ -0.07 \end{bmatrix} = .460,$$

$SS_{HYPOTHESIS} = 3.397$, $SS_{ERROR} = 22.461$,[17] and $F_{(1,36)} = 5.44, p = .025$. The standardized parameter estimates differ significantly by the hypothesis test applied to standardized coefficients. The reason for the differing results is a consequence of the differences in the scales of measurement of the predictor variables; it can be shown that the *j*th standardized coefficient is a ratio of its semipartial correlation to the square root of the proportion of variation in X_j not accounted for by the remaining predictors X_j, (e.g., tolerance) in the full model, that is,

$$\hat{\beta}_j^* = \frac{r_{Y(X_j|X_j')}}{\sqrt{1 - R^2_{j.other}}}.$$

Unstandardized regression coefficients and their standard errors have absolute magnitude for two reasons: (1) scale of measurement and (2) the underlying relationship between the predictor and response variables. Conversely, the magnitude of the standardized coefficients is most heavily determined by the semipartial correlations and the tolerances of the predictors. Consequently, a difference between standardized coefficients constitutes a test of the differences between semipartial correlation coefficients[18]—it is a test of differences between correlations of *Y* and each predictor after adjustment for other predictors in the model. The test statistic of differences between raw regression coefficients and between semipartial correlations need not be equal. The two tests are numerically independent because they test conceptually different hypotheses—differences in rates of change versus differences in strength of association. The tests between coefficients for

---

[17]The error sum of squares in standard score form is $\mathbf{Z}_Y'\mathbf{Z}_Y - \hat{\mathbf{B}}^{*\prime}\,\mathbf{Z}_X'\mathbf{Z}_Y = (n-1)(1 - R^2_{Y.X_1X_2X_3})$.

[18]The test of the differences between two standardized regression coefficients from a regression analysis is defined as

$$t = \frac{\hat{\beta}_1^* - \hat{\beta}_2^*}{\sqrt{MSE\left(\mathbf{L}\mathbf{R}_{XX}^{-1}\mathbf{L}'\right)}}$$

unstandardized and standardized models will be identical only when $S_{X_1} = S_{X_2}$. Similar tests of differences between correlation coefficients are discussed in Olkin and Finn (1995). Draper and Smith (1998, pp. 218–219) and Rencher (1998, pp. 295–300) give examples of more complicated linear hypothesis tests in which the same principles apply.

Generalizing From Univariate to Multivariate General Linear Models

We have begun this volume with a review of the common strategies for modeling a single response variable as a function of one or more continuous and/or categorical explanatory variables. Such models have great flexibility and can accommodate any combination of predictor variable types, including their interactions and powers.

(Cohen et al., 2003, pp. 640–642), where $\mathbf{R}_{XX}^{-1}$ is the inverse of the correlation matrix among the predictors and

$$MSE = \frac{1 - R^2_{Y \cdot X_1 X_2 X_3}}{n - q_f - 1}.$$

Substituting the definitions

$$\hat{\beta}_1^* = \frac{r_{Y(X_1|X_2)}}{\sqrt{1 - R^2_{1.2}}}$$

and

$$\hat{\beta}_2^* = \frac{r_{Y(X_2|X_1)}}{\sqrt{1 - R^2_{1.2}}}$$

into t sets the numerator to

$$\frac{r_{Y(X_1|X_2)} - r_{Y(X_2|X_1)}}{\sqrt{1 - R^2_{1.2}}}.$$

Setting the contrast matrix to $\mathbf{L} = [1 \quad -1 \quad 0]$ and performing the symbolic multiplication of the quantity $\sqrt{MSE\ (\mathbf{L}\mathbf{R}_{XX}^{-1}\mathbf{L}')}$, the denominator of t reduces to

$$\frac{\sqrt{MSE\, 2\left(1 + r_{12}\right)}}{\sqrt{1 - R^2_{1.2}}}.$$

Recounting these details here sets the stage for the generalization of these same analytic concepts to those instances where *more than one* dependent variable is to be analyzed simultaneously. Models with $p > 1$ response variables are classified as multivariate models that can be treated with the same four-step process—the specification of the multivariate model, estimation of its parameters, identifying measures of strength association, and defining appropriate tests of significance. We pursue these topics in the chapters that follow.

The quantities $\sqrt{1-R^2_{1.2}}$ in the numerator and denominator cancel, leaving

$$t = \frac{r_{Y(X_1|X_2)} - r_{Y(X_2|X_1)}}{\sqrt{MSE\left(2(1+r_{12})\right)}} \cdot$$

Hence, the test of the hypothesis $\beta^*_1 - \beta^*_2 = 0$ is a test of the differences between semipartial correlation coefficients. In this interpretation, approximately 36% of the variance in TMT-B is accounted for by age while about 2% of the variance in TMT-B is accounted for by education. The absolute values of the two correlations are significantly different from one another, while the absolute values of the two unstandardized slopes do not differ significantly. The difference between unstandardized rates of change is being masked by differences in variance of the predictors.