TABLE 6.4.1
Exponents, Logarithms, and Their Relationships

A. Some rules for exponents
(1) $X^a X^b = X^{(a+b)}$
(2) $X^a / X^b = X^{(a-b)}$
(3) $X^{-n} = 1/X^n$
(4) $X^{1/2} = \sqrt{X}$
(5) $X^0 = 1$

B. Some rules for logarithms
(6) $\log(bX) = \log b + \log X$
(7) $\log(b/X) = \log b - \log X$
(8) $\log(X^b) = b \log X$

C. Relationship between exponents and logarithms
(9) $\log_m m^x = X$, so $\log_{10} 10^X = X$ for base 10 logs
and $\ln_e e^X = X$ for natural (base e) logs

Note: This presentation is drawn from presentations by Hagle (1995) and Hamilton (1992).

Many presentations of transformations use *natural logarithms*, noted *ln*, with a base (specially noted e instead of m) of $e = 2.71878$ (approximately). For example, ln 1000 = 6.907755279, since $2.71878^{6.907755279} = 1000$. A third form of logarithms are base 2 logarithms, for example, $\log_2 8 = 3$, since $2^3 = 8$.

The computations for base 10 and natural logarithms can be accomplished on a simple statistical calculator that has the following functions: log, ln, 10^X, and e^X. Enter 1000, press **log** to get $\log_{10} 1000$. Enter 3, press $\mathbf{10^X}$ to get $10^3 = 1000$. Enter 1000, press **ln**, to get ln 1000 = 6.907755279. Enter 6.907755279, press e^X to get $2.71878^{6.907755279} = 1000$. From the perspective of transforming variables using logarithms, it actually does not matter whether $\log_{10}$ or ln is used—the two logarithms are linearly related to one another. In fact, 2.302585 ln = $\log_{10}$. We will use the general notation "log" throughout this section, to indicate that either ln or $\log_{10}$ can be employed.

In the numerical examples, we first took the logarithm of the number 1000. Then we took the result and raised the base of the logarithm to the log (e.g. 10^3); the latter manipulation is called taking the *antilog* of a logarithm. Having found the logarithm of a number, taking the antilog returns the original number. In general, raising any number to a power is called *exponentiation*, and the power to which a number is raised is called the *exponent*. Logarithms and exponents are inverse functions of one another. In Table 6.4.1, we present rules for exponents, for logarithms, and for the relationship between exponents and logarithms.

Logarithms and Proportional Change

When, as X changes by a constant proportion, Y changes by a constant additive amount, then Y is a logarithmic function of X; hence Y is a linear function of $\log X$. Following are a series of values of X in which each value is 1.5 times the prior value, a *constant proportionate increase*; for example, 12 = 1.5(8). In the corresponding series of Y, each value of Y is 3 points higher than the prior value (e.g., 8 = 5 + 3); Y exhibits *constant additive increase*. When a variable like X increases by a proportionate amount, $\log X$ (either $\log_{10}$ or ln) increases by a constant additive amount. Within rounding error, $\log_{10} X$ increases by .18 through the series; $\ln X$ increases by .40 through the series; $\log_2 X$ increases by .585 through the series. The increases

in $\log X$ are constant additive increases, as with Y. Thus, the relationship between $\log X$ and Y is linear and can be estimated in linear OLS regression, as in Eq. (6.4.7).

X	8	12	18	27	40.5	where $X_{(i+1)} = 1.5X_i$
$\log_{10} X$	.90	1.08	1.26	1.43	1.61	
$\ln X$	2.08	2.48	2.89	3.29	3.70	
$\log_2 X$	3	3.58	4.17	4.75	5.34	
Y	5	8	11	14	17	where $Y_{(i+1)} = Y_i + 3$

Conversely, if constant additive changes in X are associated with proportional changes in Y, then $\log Y$ is a linear function of X, and again the linear regression model correctly represents the relationship.

In some circumstances, we may transform Y rather than X. When only Y is log transformed, our basic regression equation for transformed Y becomes $\hat{Y}' = \log Y = B_1 X + B_0$. In this equation, B_1 is the amount of change that occurs in Y' given a 1-unit change in X. Note that now the change in Y is in $\log Y$ units. A 1-unit increase in $\log_{10} Y$ is associated with a 10-fold increase in raw Y; a 2-unit increase in $\log_{10} Y$ is associated with a 100-fold increase in raw Y. Similarly a 1-unit increase in $\log_2 Y$ is associated with a twofold increase (doubling of raw Y), and a 2-unit increase in $\log_2 Y$ is associated with a fourfold increase in raw Y.

Finally, proportionate changes in X may be associated with proportionate changes in Y, for example:

X	8	12	18	27	40.5	where $X_{(i+1)} = 1.5\ X_i$
Y	2	4	8	16	32	where $Y_{(i+1)} = 2\ Y_i$

If logarithms of both variables are taken, then

$\log_{10} X$	.90	1.08	1.26	1.43	1.61
$\log_{10} Y$	.30	.60	.90	1.20	1.51

Each proceeds by constant additive changes and again $\log Y$ is a linear function of $\log X$, this time after logarithmic transformation of both X and Y, as in Eq. (6.4.11) below.

6.4.4 Linearizing Relationships

Given that some nonlinear relationship exists, how does one determine which, if any, of a number of transformations is appropriate to linearize the relationship? For some relationships—for example, psychophysical relationships between stimulus intensity and subjective magnitude—there are strong theoretical models underlying the data that specify the form of the relationship; the task is one of transforming the variables into a form amenable to linear MRC analysis. Weaker models imply certain features or aspects of variables that are likely to linearize relationships. In the absence of any model to guide selection of a transformation, empirically driven approaches, based on the data themselves, suggest appropriate transformation. These include procedures presented here, including the ladder of re-expression and bulge rules of Tukey (1977) and Mosteller and Tukey (1977), and more formal mathematical approaches like the Box-Cox and Box-Tidwell procedures.

Intrinsically Linear Versus Intrinsically Nonlinear Relationships

Whether a strong theoretical model can be linearized for treatment in linear MR depends upon the way that random error is built into the model, as a familiar *additive* function or as a *multiplicative* function. Multiplicative error in a model signifies that the amount of error in the

dependent variable Y increases as the value of Y increases. Suppose we have a multiplicative theoretical model with multiplicative error.

(6.4.1) $$Y = B_0 X_1^{B_1} X_2^{B_2} e^{\varepsilon},$$

where ε refers to random error, and e is the base of the natural logarithm.

This form of regression equation, with regression coefficients as exponents, is not the familiar form of an OLS regression; it signals a nonlinear relationship of the predictors to Y.[9] The question before us is whether we can somehow transform the equation into an equation that can be analyzed with OLS regression.

Using rule (8) from Table 6.4.1, we take the logarithms of both sides of the equation. This yields a transformed equation that is linear in the coefficients (Section 6.1) and that thus can be analyzed using OLS regression:

(6.4.2) $$\log Y = \log B_0 + B_1 \log X_1 + B_2 \log X_2 + \varepsilon$$

Eq. (6.4.1) has been linearized by taking the logarithms of both sides. As shown in Eq. (6.4.2) the regression coefficients B_1 and B_2 can be estimated by regressing $\log Y$ on $\log X_1$ and $\log X_2$; the resulting B_1 and B_2 values are the values of B_1 and B_2 in Eq. (6.4.1). The value of B_0 in Eq. (6.4.1) can be found by taking the antilog of the resulting regression constant from Eq. (6.4.2). In other words, we started with a nonlinear equation (Eq. 6.4.1), transformed the equation into an equation (Eq. 6.4.2) that could be solved through OLS regression, and were able to recover the regression coefficients of the original nonlinear equation (Eq. 6.4.1). The errors are assumed to be normally distributed with constant variance in the *transformed* equation (Eq. 6.4.2). Because Eq. (6.4.1) can be linearized into a form that can be analyzed in OLS regression, it is said to be *intrinsically linearizable*. In Section 6.4.5 we illustrate four nonlinear models used in psychology and other social and biological sciences. All four are intrinsically linearizable; we show how the linearization can be accomplished.

Now we modify the equation slightly to

(6.4.3) $$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} + \varepsilon,$$

where the error ε is additive, not multiplicative; that is, the variance due to error in predicting Y is constant across the range of the predictors in the form of regression equation Eq. (6.4.3). Additive error is our standard assumption in linear MR. If we try to linearize Eq. (6.4.3) by taking the logarithms of both sides, we discover that in the resulting expression the error variance is a function of the value of Y. Heteroscedasticity would be introduced by the logarithmic transformation of the equation (see Myers, 1986, for a complete demonstration). The equation is *intrinsically nonlinear. Nonlinear regression*, introduced in Section 6.5, must be employed.

Whether we specify a model with multiplicative or additive error is a matter for theory. As Draper and Smith (1998) point out, a strategy that is often used is to begin with transformation of variable(s) to linearize the relationship (implicitly assuming that the error is multiplicative in the original scale); OLS (ordinary least squares) regression is then employed on the transformed variables. Then the residuals from the OLS regression with the transformed variable(s) are examined to see if they approximately meet the assumptions of homoscedasticity and normality. If not, then a nonlinear regression approach may be considered. (From now on, we will refer to linear MR as OLS regression in order to clearly distinguish this model from alternative regression models.)

---

[9]Logistic regression, covered in Chapter 13, is a form of nonlinear regression with regression coefficients as exponents.

Equation (6.4.2) illustrates that transformations to linearize relationships may involve both the predictors $X_1, X_2, \ldots, X_k$ and the dependent variable Y. As indicated later, the choice of transformation of both X and Y may be driven by strong theory.

6.4.5 Linearizing Relationships Based on Strong Theoretical Models

In such fields as mathematical biology, psychology and sociology, neuropsychology, and econometrics, relatively strong theories have been developed that result in postulation of (generally nonlinear) relationships between dependent and independent variables. The adequacy of these models is assessed by observing how well the equations specifying the relationships fit suitably gathered data. We emphasize that the equations are not arbitrary but are hypothetically descriptive of "how things work." The independent and dependent variables are observables, the form of the equation is a statement about a process, and the values of the constants of the equation are estimates of parameters that are constrained or even predicted by the model. In our presentations of models, we assume *multiplicative error* in the nonlinearized form, omit the error term, and show the expression with the predicted score $\hat{Y}$ in place of the observed Y. (We discuss the treatment of the same models but with additive error assumed in Section 6.5 on nonlinear regression.) Here we illustrate four different nonlinear relationships that appear as *formal models* in the biological or social sciences, including psychology and economics.

Logarithmic Relationships

Psychophysics is a branch of perceptual psychology that addresses the growth of subjective magnitude of sensation (e.g., how bright, how loud a stimulus seems) as a function of the physical intensity of a stimulus. A common psychophysical model of the relationship of energy X of a physical stimulus to the perceived magnitude Y of the stimulus is given in Eq. (6.4.4),

$$c^{\hat{Y}} = dX_1, \tag{6.4.4}$$

where c and d are constants. The equation asserts that changes in stimulus strength X are associated with changes in subjective response Y as a power of a constant. The relationship between X and Y is clearly nonlinear. Figure 6.4.1(A) illustrates an example of this relationship for the specific equation $8^Y = 6X$, where $c = 8$ and $d = 6$. Suppose we wish to analyze data that are proposed to follow the model in Eq. (6.4.5) and to estimate the coefficients c and d. We transform Eq. (6.4.5) into a form that can be analyzed in OLS regression. We take logarithms of both sides of the equation, yielding

$$\hat{Y} \log c = \log d + \log X_1, \tag{6.4.5}$$

Solving for Y we find

$$\hat{Y} = \frac{\log d}{\log c} + \frac{1}{\log c} \log X_1. \tag{6.4.6}$$

If we let $(\log d)/(\log c) = B_0$ and $1/(\log c) = B_1$, we see that the psychophysical model in Eq. (6.4.4) postulates a logarithmic relationship between stimulus strength (X), and subjective response (Y), which is, in fact, a form of Fechner's psychophysical law (Fechner, 1860), given in Eq. (6.4.7).

$$\hat{Y} = B_1 \log X_1 + B_0. \tag{6.4.7}$$

We can apply Eq. (6.4.7) to suitably generated data using OLS regression by regressing Y (e.g., judgments of brightness of lights) on $\log X_1$ (e.g., the logarithm of a measure of light intensity). This yields estimates of B_1 and B_0. From these estimates, we solve for the constant c in Eq. (6.4.4) from the relationship $1/(\log c) = B_1$, or the reciprocal, yielding $\log c = 1/B_1$. Then

$$c = \text{antilog}\,\frac{1}{B_1}. \tag{6.4.8}$$

To solve for the constant d, we use $(\log d)/(\log c) = B_0$, which yields

$$d = \text{antilog}\,\frac{B_0}{B_1}. \tag{6.4.9}$$

The values of c and d will be of interest because they estimate parameters in the process being modeled (e.g., the relationship of light intensity to perceived brightness), as will R^2 as a measure of the fit of the model (see Section 6.4.16). Finally, the shape of the function in Fig. 6.4.1(A) is typical of logarithmic relationships between X and Y in which Y varies linearly as a function of $\log X$.

Power Relationships

Now consider an alternative formulation of the psychophysical relationship of stimulus to subjective magnitude expressed by the equation

$$\hat{Y} = cX^d \tag{6.4.10}$$

where c and d are constants, and Y is a power function of X, such that proportional growth in Y relates to proportional growth in X. This theoretical model has been offered by Stevens (1961) as the psychophysical law that relates X, the energy of the physical stimulus, to the perceived magnitude of sensation Y. In Stevens' model, the exponent or power d estimates a parameter that characterizes the specific sensory function and is not dependent on the units of measurement, whereas c does depend on the units in which X and Y are measured. We stress that Stevens' law is not merely one of finding an equation that fits data—it is rather an attempt at a parsimonious description of how human discrimination proceeds. It challenges Fechner's law, which posits a different fundamental equation, one of the form of Eq. (6.4.4), in which proportional growth in X relates to additive growth in Y. Two specific examples of Eq. (6.4.5) are given in Fig. 6.4.1(B). The left hand panel of Fig. 6.4.1(B) illustrates a power function with an exponent $d > 1$, specifically $Y = .07X^{1.7}$, where $c = .07$ and $d = 1.7$. The right-hand panel of Fig. 6.4.1(B) illustrates a power function with $d < 1$, specifically $Y = .07X^{.2}$, where $c = .07$ and $d = .20$. Values of d are a critical component of Stevens' law applied to different sensory continua. For example, the exponent d for perceived brightness of short duration lights is $d = .33$; for perceived saltiness of sips of sodium chloride (salt) solution, $d = 1.3$ (Marks, 1974).

To linearize the relationship in Eq. (6.4.10), we take the logarithms of both sides, yielding

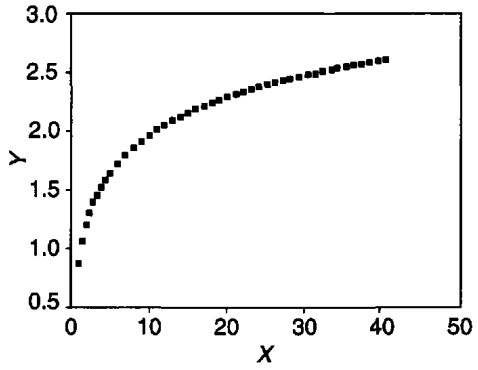
$$\log \hat{Y} = d \log X + \log c \quad \text{or} \quad \log \hat{Y} = B_1 \log X_1 + B_0. \tag{6.4.11}$$

To analyze the relationship between X and Y in Eq. (6.4.10) using OLS regression, we would compute the logarithms of X and Y and predict $\log Y$ from $\log X$. In Eq. (6.4.11) $B_0 = \log c$, so that

$$c = \text{antilog}\, B_0. \tag{6.4.12}$$

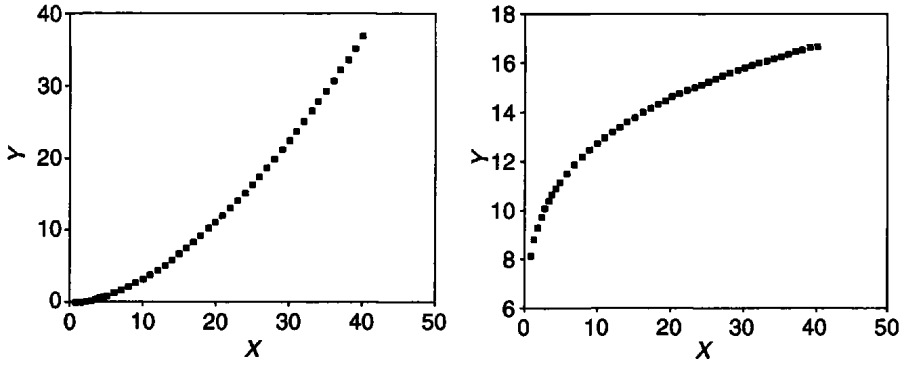
$$d = B_1. \tag{6.4.13}$$

(A) Logarithmic relationship.



$8^Y = 6X$

(B) Power relationships.



$Y = .07X^{1.7}$ $Y = .07X^{.2}$

FIGURE 6.4.1 Some functions used to characterize the relationship of X to Y in theoretical models.

Note that Eq. (6.4.11) informs us that the predicted scores will be in the log metric; to convert the predicted scores to the raw metric, we would take the antilog of each predicted score in the log metric, that is, antilog $\hat{Y}_{\log} = \hat{Y}_{\text{original units}}$.

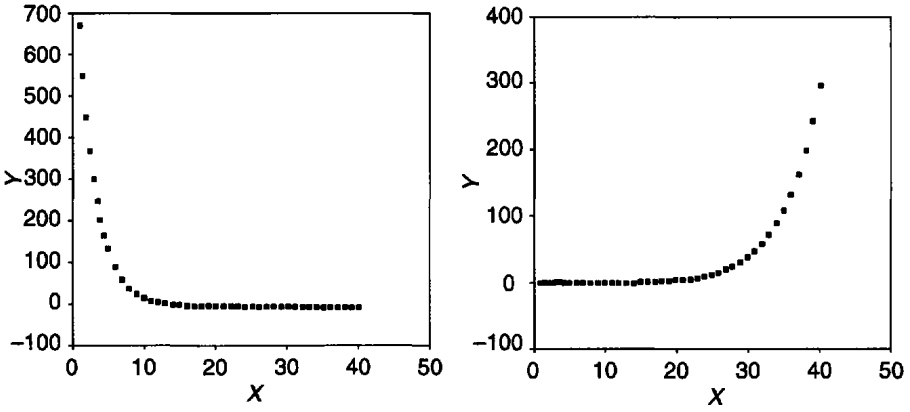
Exponential Growth Model Relationships

There is great interest in psychology in the trajectories of growth of various phenomena over time (e.g., drug use, learning, intellectual growth and decline). An exponential relationship between X and Y used to model growth or decay of Y as a function of X is given by

(6.4.14) $$\hat{Y} = ce^{dX}.$$

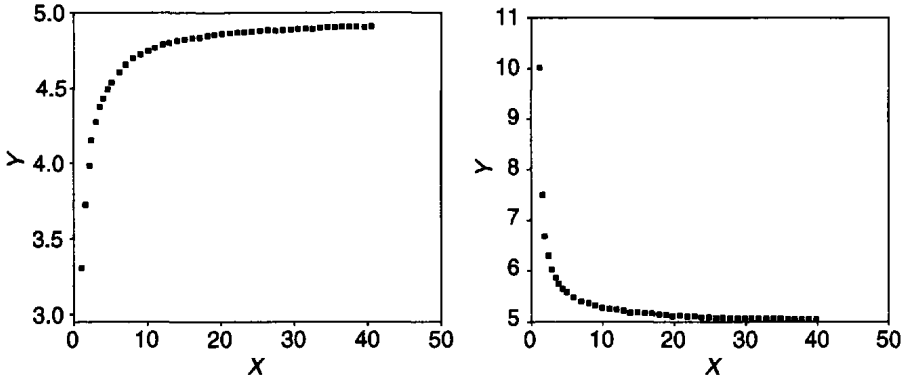
In this model, the change in Y at any point depends on the level of Y. If $d > 0$, Y grows from a starting value of c when $X = 0$, with Y rising in ever increasing amounts, for example, as in college tuition over time, referred to as *exponential growth*. Exponential growth is illustrated in

(C) Exponential growth relationships.



$Y = 1000e^{-.4x}$ $\qquad$ $Y = .10e^{.2X}$

(D) Hyperbolic (inverse polynomial) relationships.



$$Y = \frac{X}{.1 + .2X} \qquad Y = \frac{X}{-.1 + .2X}$$

Note: The scaling of the y-axis changes from panel to panel.

FIGURE 6.4.1 Continued.

the right-hand panel of Fig. 6.4.1(C), specifically in the equation $Y = .10e^{.2X}$, where $c = .10$ and $d = .2$. If $d < 0$, we have *exponential decay*. Y declines from an initial value of c when $X = 0$, illustrated in the left-hand panel of Fig. 6.4.1(C), specifically in the equation $Y = 1000e^{-.4X}$, where $c = 1000$ and $d = -.4$. If c were the amount of knowledge of statistics one had on the day of a statistics final exam, and the amount one forgot each day following the exam were proportional to the amount retained by the early morning of the day, we would have *exponential decay*. Eq. (6.4.14) is linearized by taking the logarithms of both sides, yielding

(6.4.15) $$\log \hat{Y} = dX + \log c \quad \text{or} \quad \log \hat{Y} = B_1 X_1 + B_0$$

so that

(6.4.16) $$B_0 = \log c$$

and

(6.4.17) $$B_1 = d.$$

If the model $\hat{Y} = ce^{dX}$ is expanded to include an asymptote a, as in the expression $\hat{Y} = a + ce^{dX}$, and the coefficients c and d are negative, then the resulting form of the equation will be a curve that rises and levels off at the value a—as, for example, the additional amount of statistics learned for each additional hour of studying before an exam, up to an asymptote representing all the material in the statistics course. If the model is applied in an experiment with a treated and a control group, another dichotomous predictor can be added to code the groups; the result will be curves of different elevations with the difference in height representing the effect of treatment (Neter, Kutner, Nachtsheim, & Wasserman, 1996, Chapter 13).

Hyperbolic Relationship (Inverse Polynomial Model)

A second form of growth model used in economics and biology (Myers, 1986) is a *diminishing returns model*, which characterizes growth to some asymptote (upper limit or lower limit). This is the hyperbolic (inverse polynomial) function:

(6.4.18) $$\hat{Y} = \frac{X}{c + dX}.$$

In this model the value $1/d$ is the asymptote; the increase in Y is inversely related to distance from the asymptote, hence the term *diminishing return*. Figure 6.4.1(D) illlustrates two such curves. The left hand panel shows the equation $Y = X/(.1 + .2X)$, where $c = .1$ and $d = .2$; it rises to an asymptote of 5, since $d = .2$, and $1/d = 5$. The right-hand figure shows the equation $Y = X/(-.1 + .2X)$, where $c = -.1$ and $d = .2$; it falls to an asymptote of 5, again because $d = .2$.

Unlike the use of logarithms to linearize the previous equations, linearizing Eq. (6.4.18) involves the use of reciprocals. By algebraic manipulation, Eq. (6.4.18) can be written as a linear function of the reciprocals of X and Y:

(6.4.19) $$\frac{1}{\hat{Y}} = c\frac{1}{X} + d \quad \text{or} \quad \frac{1}{\hat{Y}} = B_1\frac{1}{X} + B_0.$$

To estimate this equation using OLS regression, we would predict the reciprocal of Y from the reciprocal of X. The coefficient B_1 from the OLS regression equals c from Eq. (6.4.18), and $B_0 = d$.

Assessing Model Fit

Even when we accomplish the transformation to linear form, as has been shown for four different theoretical models, a problem exists that is worth mentioning. When the dependent variable analyzed in the transformed equation is not itself Y, but is rather some function of Y—for example, $\log Y$ or $1/Y$—the B_0 and B_1 coefficients from the transformed equation are the coefficients that minimize the sum of squared residuals (the least squares estimates) for predicting the transformed Y. They are not the least squares estimates that would result if the untransformed Y were predicted. There is no direct function for converting the coefficients from the transformed equation to corresponding coefficients from the untransformed equation. The issue arises as to whether there is better fit in a variance accounted for sense (R^2_Y) in the transformed over the untransformed equation. The R^2s associated with models predicting different forms of Y (e.g., $Y, \log Y, \sqrt{Y}$) are not directly comparable. *In general, one cannot directly compare the fit of two models with different dependent variables.* Comparing fit across models employing different transformations is explored in Section 6.4.16.

6.4.6 Linearizing Relationships Based on Weak Theoretical Models

We may employ the same transformations from strong theoretical models in linearizing relationships between variables that are well below the level of exact mathematical specification of the strong theoretical models discussed previously. However, our present "weak theory" framework is more modest; we are here not generally interested in estimating model parameters c and d, as we are in Section 6.4.5, because we do not have a theory that generated the equations in the first place.

Logarithmic Transformations

Logarithmic transformations often prove useful in biological, psychological, social science, and economics applications. All we might have, for example, is a notion that when we measure a certain construct X by means of a scale X, it changes proportionally in association with additive changes in other variables. As we discussed earlier, if we expect that proportionate changes in X are associated with additive changes in Y, we might well transform X to $\log X$. If we expect proportionate changes in Y to be associated with proportionate changes in X, we might well transform Y to $\log Y$ and X to $\log X$. Variables such as age or time-related ordinal predictors such as learning trials or blocks of trials are frequently effectively log-transformed to linearize relationships. This is also frequently the case for physical variables, as for example energy (intensity) measures of light, sound, chemical concentration of stimuli in psychophysical or neuropsychological studies, or physical measures of biological response. At the other end of the behavioral science spectrum, variables such as family size and counts of populations as occur in vital statistics or census data are frequently made more tractable by taking logarithms. So, often, are variables expressed in units of money, for example, annual income or gross national product.

By logarithmic transformation, we intend to convey not only $\log X$ but such functions as $\log(X-K)$ or $\log(K-X)$, where K is a nonarbitrary constant. Note that such functions are not linearly related to $\log X$, so that when they are appropriate, $\log X$ will not be. K, for example, may be a sensory threshold or some asymptotic value. (In Section 6.4.8 we discuss the use of small arbitrary additive constants for handling the transformation of Y scores of zero, yielding *started logs and powers.*)

Reciprocal Transformation

Reciprocals arise quite naturally in the consideration of rate data. Imagine a perceptual-motor or learning task presented in time limit form—all subjects are given a constant amount of time (T), during which they complete a varying number of units (u). One might express the scores in the form of rates at u/T, but because T is a constant, we may ignore T and simply use u as the score. Now, consider the same task, but presented in work limit form—subjects are given a constant number of units to complete (U) and are scored as to the varying amounts of time (t) they take. Now if we express their performance as *rates*, it is U/t and, if we ignore the constant U, we are left with $1/t$, not t. If rate is linearly related to some other variable X, then for the time limit task, X will be linearly related to u, but for the work limit task, X will be linearly related not to t, but to $1/t$. There are other advantages to working with $1/t$. Often, as a practical matter in a work limit task, a time cutoff is used that a few subjects reach without completing the task. Their exact t scores are not known, but they are known to be very large. This embarrassment is avoided by taking reciprocals, because the reciprocals of very large numbers are all very close to zero and the variance due to the error of using the cutoff $1/t$ rather than the unknown true value of $1/t$ is negligible relative to the total variance of the observations.

6.4.7 Empirically Driven Transformations in the Absence of Strong or Weak Models

Suppose that through our use of diagnostic approaches suggested in Chapter 4, we discover in our data characteristics that suggest the need for transformation. Graphical displays of X–Y relationships (e.g., lowess plots, Section 4.2.2) may suggest nonlinear relationships. Scatterplots of residuals around the lowess line (Section 4.4.4 and Fig. 4.4.5) may suggest heteroscedasticity. Quantile-quantile (q-q, Section 4.4.6) plots of residuals against a normal variate may uncover their nonnormality.

Suppose, however, that we have neither strong nor weak models to suggest specific transformations to ameliorate these conditions in our data. We can nonetheless draw on a rich collection of strategies for linearizing relationships and for improving the characteristics of residuals. Sections 6.4.8 through 6.4.14 describe these strategies. Our use of these strategies is empirically driven by our data rather than by theory. This approach is usual in statistics but to date has had less impact in the behavioral sciences. The approach is certainly appropriate and potentially useful for behavioral science data. The purpose of undertaking empirically driven transformations is to produce a regression equation that both characterizes the data and meets the conditions required for accurate statistical inference. Section 4.5 provides a discussion of the form of relationships and conditions in the data that lead to particular strategies for transformation.

6.4.8 Empirically Driven Transformation for Linearization: The Ladder of Re-expression and the Bulging Rule

Let us assume that a lowess plot of Y against X has revealed a curvilinear relationship that is monotonic with one bend, as in all the illustrations of Fig. 6.4.1. Also assume that we have no theoretical rationale for declaring that a particular mathematical function generated the curve. How should we approach linearizing (straightening) the relationship? Both informal (by inspection) and formal (numerical) approaches have been developed to guide transformation for linearization. If the relationship we observe is monotonic and has a single bend, one strong possibility for transformation is the use of *power transformations*, in which a variable is transformed by raising it to some power. In general, the power function is

$$Y' = Y^{\lambda}, \tag{6.4.20}$$

where Y is the original variable, Y' is the transformed variable, and λ is the exponent, (i.e., the power to which Y is raised). The transformed variable then replaces the original variable in regression analysis.

The Ladder of Re-expression

Actually, we have already encountered and will continue to encounter examples of power transformations, which include reciprocals, logarithms, powers in polynomial regression, square roots, and other roots. In their classic work, Mosteller and Tukey (1977) described a *ladder of re-expression* (*re-expression* is another term for *transformation*) that organizes these seemingly disparate transformations under a single umbrella. This ladder of re-expression was proposed to guide the selection of transformations of X and Y to linearize relationships. The ladder can also be used to transform skewed variables prior to analysis.

The ladder is a series of *power functions* of the form $Y' = Y^{\lambda}$, which transform Y into Y' (or, equivalently, X into X'). Again, power functions are useful for straightening a relationship between X and Y that is monotonic and has a single bend; hence power functions are characterized as *one-bend transformations*.

The problem is to find an appropriate value of λ to use in transforming a variable that makes its distribution more normal or that eliminates nonlinearity of relationship between that variable and another variable. Some values of λ, shown below, lead to familiar transformations (Neter, Kutner, Nachtsheim, & Wasserman, 1996, p. 132), though many values of λ other than those given here are possible.

In general	$Y' = Y^{\lambda}$.
Square	$Y' = Y^2$; $\lambda = 2$.
Square root	$Y' = Y^{1/2} = Y^{.5} = \sqrt{Y}$; $\lambda = .5$.
Logarithm	$Y' = \ln Y$; $\lambda = 0$ (a special case).[10]
Reciprocal	$Y' = \frac{1}{Y}$; $\lambda = -1$.

Transforming Individual Variables Using the Ladder of Re-expression and Changes in Skew

Transforming individual variables to be more symmetric is not our focus here (linearizing relationships through appropriate selection of a λ is), but it is useful to understand how the various powers on the ladder change the distribution of individual variables. These changes are the basis of straightening out nonlinear relationships. Values of $\lambda > 1$ compress the lower tail of a distribution and stretch out the upper tail; a negatively skewed (i.e., long, low tail) variable becomes less skewed when a transformation with $\lambda > 1$ is applied. Values of $\lambda < 1$ stretch the lower tail of a distribution and compress the upper tail; a positively skewed (i.e., long, high tail) variable becomes less skewed when a transformation with $\lambda < 1$ is applied. The farther from 1 on either side is the value of λ, the more extreme is the compression and stretching. This allows us to compare the familiar logarithmic and square root transformations: $\log X$ (associated with $\lambda = 0$) and $\sqrt{X}$ (where $\lambda = ½$). The logarithmic transformation is stronger; that is, it compresses the upper tail and stretches the lower tail of a distribution more than does the square root transformation.

The Bulging Rule

In addressing the problem of how to select a value of λ to apply to X or Y so as to linearize a relationship, Mosteller and Tukey (1977) proposed a simple graphical bulging rule. To use the bulging rule, one examines one's data in a scatterplot of Y against X, imposes a lowess curve to suggest the nature of the curvature in the data, and selects a transformation based on the shape of the curve. Suppose the curve in the data follows the curve in Figure 6.4.1(A), that is, Y rises rapidly for low values of X and then the curve flattens out for high values of X. There are two options for transforming that will straighten the relationship between X and Y. One option is to transform Y by moving up the ladder above $\lambda = 1$; this means applying a power transformation to Y with an exponent greater than 1, (e.g., $Y^{1.5}$, Y^2). This will stretch up the high end of Y (pulling the high values of Y even higher), straightening the relationship. Alternatively, one may transform X by moving down the ladder below $\lambda = 1$

[10]The logarithm bears special comment. In fact, the expression Y^0 transforms all values of Y to 1.0, since $Y^0 = 1$. However, as $\lambda \to 0$, the expression $(Y^{\lambda} - 1)/\lambda \to \ln Y$, leading to the use of the natural logarithm as the transformation when $\lambda = 0$.

(e.g., $X^{.5} = \sqrt{X}, \log X$). This will stretch the low end of X (to the left), again straightening out the relationship. Suppose one finds in one's data that Y increases as X increases, but with but with the shape of the curvature as in Fig. 6.4.1(B, left-hand panel), that is, a slow initial rise in Y as a function of X for low values of X and a rapid rise at high values of X. We may straighten the relationshhip by moving up the ladder for X (e.g., to X^2) or down the ladder for Y (e.g., $Y^{.5} = \sqrt{Y}, \log Y$). For a shape like that in Fig. 6.4.1 (C, left-hand panel) one could either move down the ladder for X or down the ladder for Y. One may try a range of values of λ applied to either X or Y, typically between -2 and $+2$. Mosteller and Tukey (1977) present a simple numerical method for deciding if straightening has been successful. Modern graphical computer packages[11] make this work easy by providing a "slider" representing values of λ that can be moved up and down with a mouse. As the slider is moved, the value of λ is changed; the data are graphically displayed in a scatterplot of Y against X with a lowess function superimposed, and one can visually select the value of λ that straightens the X–Y relationship. Sections 6.4.9 and 6.4.10 describe quantitative approaches to selecting value of λ to transfrom Y and X, respectively.

Should X or Y Be Transformed?

The bulging rule makes it clear that for linearizing a one-bend nonlinear relationship, we may transform either X or Y. The choice between X and Y is dictated by the nature of the residuals when untransformed Y is regressed on untransformed X. If the residuals are well behaved with the untransformed data, then transformation of Y will lead to heteroscedasticity; one should transform X. If, on the other hand, the residuals are problematic (heteroscedastic, non-normal) with the untransformed data, then transforming Y may improve the distribution of residuals, as well as linearize the relationship. Figure 6.4.2, discussed in Section 6.4.17, illustrates transformation of Y versus X.

What to Do with Zeros in the Raw Data: Started Logs and Powers

Use of the family of power functions assumes that the variables to be transformed have zero as their lowest value. Logarithms, a frequently used transformation from the power function family, are undefined for numbers less or equal to zero. Mosteller and Tukey (1977) proposed a remedy for distributions containing scores equal to zero—add a very small constant c to all the scores in the distribution and apply the logarithmic transformation to $\log(Y + c)$. For negative values of λ the same approach is used; one transforms $(Y + c)^{\lambda}$. These transformations are referred to as *started logs* and *started powers*.

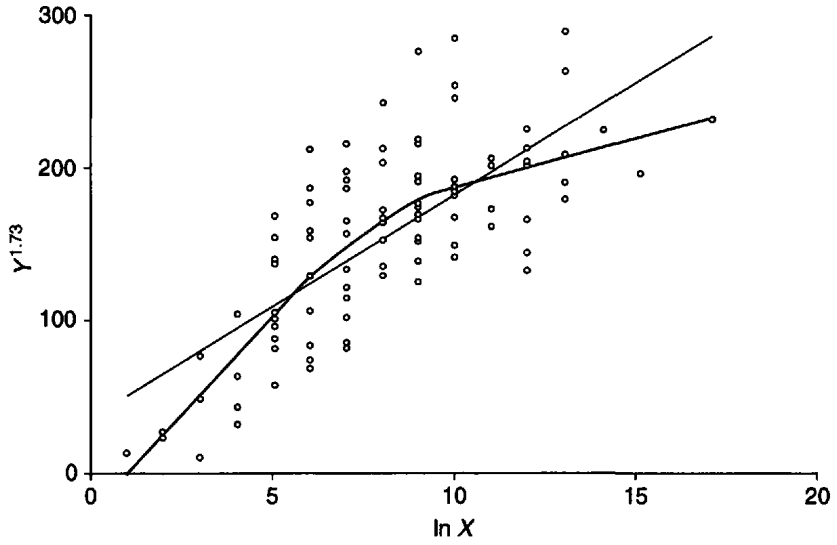
Variable Range and Power Transformations

Power transformations assume that all values of the variable being transformed are positive and without bound at the upper end. Power functions are most effective when the ratio of the highest to lowest value on a variable is large, at least 10 (e.g., Draper & Smith, 1998). If the ratio is small, then power transformations are likely to be ineffective, because for very small ratios, the power transformations are nearly linear with the original scores.

---

[11]The ARC software developed by R. D. Cook and Weisberg (1999) provides this technology, and much other technology useful for regression graphics and transformations. ARC is freeware accessible from the School of Statistics, University of Minnesota: www.stat.umn.edu/arc/.

(A) The regression of $Y^{1.73}$ on X.



(B) The regression of Y on $\ln(X)$.

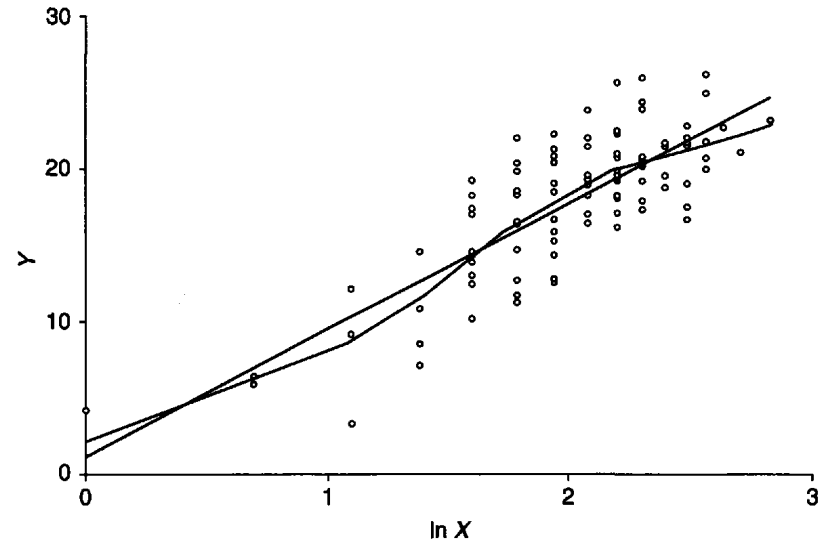


FIGURE 6.4.2 Power transformation of Y versus logarithmic transformation of X to linearize a relationship. The data are the same data as in Fig. 6.1.1 and Table 6.2.1.

6.4.9 Empirically Driven Transformation for Linearization in the Absence of Models: Box-Cox Family of Power Transformations on *Y*

Again suppose we find a one-bend monotonic relationship in our data and also observe problems with the residuals from an OLS regresion. In lieu of a trial and error approach to selecting λ to transform Y, Box and Cox (1964) provided a numerical procedure for selecting a value of λ to be applied the dependent variable Y (but not X). The goal of the Box-Cox procedure is to select λ to achieve a linear relationship with residuals that exhibit normality and homoscedasticity. Both linearization and improved behavior of residuals drive the choice of λ. Box-Cox is a

standard approach in statistics; although it has not been much used in some areas of behavioral science, Box-Cox transformation may be usefully applied to behavioral science data.

The mathematical details of the Box-Cox transformation are given in Box 6.4.1 for the interested reader. Maximum likelihood estimation (described in Section 13.2.9) is used to estimate λ in statistical software. Box 6.4.1 also provides a strategy for comparing the fit of regression models that use different transformations. Suppose we wished to try two transformations of Y, say $\sqrt{Y}$ and $\log Y$. One cannot simply fit two regression equations, one with $Y' = \sqrt{Y}$ as the DV and one with $Y' = \log Y$ as the DV and compare the fit of these models directly, because the dependent variables are on different scales (see also Section 6.4.16 for model comparison across transformations). Finally, Box 6.4.1 describes an approach to a diagnostic test whether a transformation is required; this approach also provides a preliminary estimate of λ. The approach can be implemented with OLS regression software; no specialized software is required. The value of λ produced by Box-Cox is often used to suggest the choice of a familiar transformation. For example, if $\hat{\lambda}$ from Box-Cox is .43, we might well choose a square root transformation, where $\lambda = .5$.

BOX 6.4.1
Achieving Linearization with the Box-Cox Transformation of *Y*

The Box-Cox transformation in its non-normalized form (Atkinson, 1985; Draper & Smith, 1998, p. 280) is given as

(6.4.22a) $$Y_i^{(\lambda)} = \frac{Y_i^{\lambda} - 1}{\lambda} \quad \text{for} \quad \lambda \neq 0$$

and

(6.4.22b) $$Y_i^{(\lambda)} = \ln Y_i \quad \text{for} \quad \lambda = 0.$$

The notation $Y^{(\lambda)}$, used by Ryan (1997), distinguishes the full Box-Cox transformation from simply raising Y to the power λ, written as Y^{λ}. Division by λ in Eq. (6.4.22a) preserves the direction of ordering from low to high after transformation of Y to $Y^{(\lambda)}$ when $\lambda < 0$ (Fox, 1997).

Expressions (6.4.22a) and (6.4.22b) are non-normalized, which means that we cannot try different values of λ, fit regression models, and compare the results directly to see which value of λ produces the best fit. Instead, we use the normalized transformation, which allows comparison across models of the same data using different values of λ (Draper & Smith, 1998, p. 280):

(6.4.23a) $$Z_i^{(\lambda)} = \frac{Y_i^{\lambda} - 1}{\lambda (Y_G)^{\lambda - 1}} \quad \text{for} \quad \lambda \neq 0$$

and

(6.4.23b) $$Z_i^{(\lambda)} = Y_G \ln Y_i \quad \text{for} \quad \lambda = 0$$

The term Y_G is the geometric mean of the Y scores in the untransformed metric and is computed as follows:

(6.4.24) $$Y_G = (Y_1 Y_2 Y_3 \cdots Y_n)^{1/n}$$

The geometric mean of a set of Y scores is easily calculated in two steps, following the transformation of each Y score into $\ln Y$. First, compute the arithmetic mean of the

ln Y scores:

$$\frac{\sum \ln(Y_i)}{n}$$

Then exponentiate this value to find the geometric mean

(6.4.25) $$Y_G = e^{\Sigma \ln(Y_i)/n}$$

The use of the geometric mean preserves scaling as Y is transformed to $Z^{(\lambda)}$. This, in turn, means that values of $SS_{residual}$ as a measure of lack of model fit may be compared across equations using different values of λ.

There are three ways to proceed in using Box-Cox to select a value of λ. One may try a series of values of λ to transform Y for a single data set (Draper & Smith, 1998). For each value of λ one would compute the values of $Z^{(\lambda)}$ according to normalized Eqs. (6.4.23a) and (6.4.23b), and predict $Z^{(\lambda)}$ in an OLS regression, retaining the value of residual sums of squares for that equation $SS_{residual:Z_i^{(\lambda)}}$. Then one would plot the values of $SS_{residual:Z_i^{(\lambda)}}$ against λ and select a value of λ that appeared to bring $SS_{residual:Z_i^{(\lambda)}}$ close to a minimum. A range of λ from about -2 to $+2$ might be tried, perhaps in increments of ½ : $-2, -1.5, -1.0, \ldots, 2$ (Draper & Smith, 1998).

Second, a form of statistical estimation (a method of selecting estimates of parameters) called *maximum likelihood estimation*[12] can be used mathematically to estimate the value of λ and simultaneously to estimate the values of the regression coefficients for $X_1, X_2, \ldots, X_K$ for predicting $Z^{(\lambda)}$ in the regression equation $\hat{Z}^{(\lambda)} = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k$. The value of λ selected by the method of maximum likelihood is referred to as the *maximum likelihood estimate* of λ. In addition, a confidence interval can be computed around the maximum likelihood estimate of λ. If the confidence interval includes the value $\lambda = 1$, this suggests that there is no need for transformation, since any score raised to the power 1 is simply the score itself.

Constructed variables and a diagnostic test of the need for transformation. A third method for estimating λ derives from a statistical test of whether a transformation of the dependent variable Y would improve prediction, suggested by Atkinson (1985) and described in detail in Fox (1997, p. 323). The null hypothesis is H_0: $\lambda = 1$, i.e., that no power transformation is needed. To operationalize the test, we create a *constructed variable* of the form:

(6.4.26) $$W_i = Y_i\left(\ln\frac{Y_i}{Y_G} - 1\right)$$

where Y_G is the geometric mean given in Eq. (6.4.25).

This constructed variable is included as an additional predictor in an OLS regression predicting Y in its original untransformed scale from the set of predictors $X_1, X_2, \ldots, X_k$:

(6.4.27) $$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + \theta W_i$$

If the θ coefficient is significant, this supports the need for transformation. The value $(1 - \hat{\theta}) = \hat{\lambda}$ provides a preliminary estimate of λ for use in transformation. The question then arises as to how we apply the estimated value of λ to generate a transformed Y score. We can use Eqs. (6.4.22a) and (6.4.22b) to generate transformed $Y^{(\lambda)}$. Alternatively, when $\lambda = 0$, we can simply compute Y^{λ}, using log Y.

[12]Maximum likelihood estimation for λ in Box-Cox is implemented in the ARC software of R. D. Cook and Weisberg (1999). The likelihood function to be maximized in selecting λ is monotonically related to $SS_{residual}$: $Z_i^{(\lambda)}$

6.4.10 Empirically Driven Transformation for Linearization in the Absence of Models: Box-Tidwell Family of Power Transformations on *X*

Suppose we observe a one-bend nonlinear relationship, but the residuals are well behaved (i.e., are homoscedastic and normal in form). To linearize the relationship we should transform X; transforming Y may introduce heteroscedasticity and/or non-normality of residuals. Again, we confront the question of how to choose the value of λ. Paralleling the Box-Cox procedure for transforming Y, Box and Tidwell (1962) provided a numerical strategy for the choice of transformations on predictors. The Box-Tidwell procedure may be simultaneously applied to several predictors, each with a different power transformation. Box 6.4.2 presents the procedure for a single predictor X. Atkinson (1985), Fox (1997), and Ryan (1997) present the multiple-predictor case in detail. As with Box-Cox, a test is provided for whether transformation of the predictor is required; the strategy also yields a preliminary estimate of λ and requires only OLS regression software. Again, this is a standard procedure in statistics that may well be useful for behavioral science data.

BOX 6.4.2
Achieving Linearization With the Box-Tidwell Transformation of *X*

The expressions for transformed X in Box-Tidwell closely resemble those in Eqs. (6.4.22a) and (6.4.22b) for Box-Cox:

(6.4.28a) $$X_i^{(\lambda)} = X_i^{\lambda} \quad \text{for} \quad \lambda \neq 0,$$

(6.4.28b) $$X_i^{(\lambda)} = \ln X \quad \text{for} \quad \lambda = 0.$$

Unlike Box-Cox, there is no need for normalization of the transformed scores in order to compare models using different values of λ, since the dependent variable Y is identical across equations being compared.

One may use a constructed variable strategy to provide a test of the need for transformation and a preliminary estimate of λ. For a single predictor X, the constructed variable is given as

(6.4.29) $$V_i = X_i \ln X_i.$$

In a series of steps provided by Box and Tidwell (1962) and described in detail in Fox (1997 p. 325), one can test whether a transformation of X will provide improved prediction of Y; again, H_0: $\lambda = 1$ (no transformation is required). One can also estimate the value of λ in Eqs. (6.4.28a) and (6.4.28b).

1. First, predict Y from untransformed X in the equation $\hat{Y}_i = B_0 + B_1 X_i$.
2. Then, predict Y from untransformed X plus constructed variable V from Eq. (6.4.29) in the following equation:

(6.4.30) $$\hat{Y} = B_0' + B_1' X + \phi V_i = B_0' + B_1' X + \phi X_i \ln X_i.$$

If the ϕ coefficient is significant, this supports the need for transformation.

3. An estimate of λ is given as follows:

$$\hat{\lambda} = \frac{\phi}{B_1} + 1, \tag{6.4.31}$$

where B_1 is taken from step 1, and ϕ is taken from step 2.

A second iteration of the same three steps, but using $X_i^{\hat{\lambda}}$ in place of X throughout, (i.e., in the regression equations in steps 1 and 2, and in the computation of $V_i = X_i \ln X_i$ for step 2 where $\hat{\lambda}$ is taken from the first pass of step 3), yields a better estimate of the maximum likelihood estimate of λ. These iterations continue until the estimates of λ change by only tiny amounts.

6.4.11 Linearization of Relationships With Correlations: Fisher z' Transform of r

Sometimes a variable is measured and expressed in terms of the Pearson product moment correlation r. Two examples arise from personality psychology. First, measures of consistency of judgments of personality by the same person over time are cast as correlations. Second, in the Q-sort technique for assessing personality, items are sorted into rating categories of prescribed size (usually defining a quasi-normal distribution) so as to describe a complex phenomenon such as personality. The similarity of two such Q-sort descriptions, for example, actual self and ideal self, is then indexed by the r between ratings over the set of items. The sampling distribution of r is skewed; the Fisher z' transformation described in Section 2.8.2 functions to normalize the distribution of r. The Fisher z' transformations of correlations are more likely to relate linearly to other variables than are the correlations themselves. The Fisher z' transformation has its greatest effect as the magnitude of r approaches 1.0. The z' transformation of r is given in Appendix Table B.

6.4.12 Transformations That Linearize Relationships for Counts and Proportions

In our presentation of transformations to linearize relationships, we have not mentioned dependent variables with special characteristics, such as counts of the number of events that occur in a given time period, or proportions. Counts have the special property that they are bounded at (cannot be lower than) zero and are positively skewed for rare events. Proportions are bounded at zero at the low end of the scale and at one at the high end of the scale. The fact that both counts and proportions are bounded means that they may not be linearly related to other continuous variables.

Arcsine, Logit, and Probit Transformations of Proportions

Since proportions are bounded at zero and one, the plot of a DV in the form of proportions against a continuous predictor may be S-shaped, as illustrated in Fig. 13.1.1(B), with values of Y compressed (flattened out) for low and high values of X. We note that there are two bends in the S-shaped curve. Therefore a simple power transformation will not straighten the function. Three transformations are commonly employed to transform dependent variables in the form of proportions; the arcsine, logit, and probit transformations. All three transformations linearize relationships by stretching out the tails of the distribution of proportions, eliminating the two bends of the S-shape. Hence, they are referred to as *two-bend transformations*. Of

the three, *only the arcsine transformation* stabilizes variances, as well a straightening out the relationship. Here we illustrate how these transformations can be calculated by hand to facilitate reader insight. In practice, standard statistical packages are used to calculate these transformations.

Arcsine transformation. The arcsine transformation is given as follows:

$$A = 2 \text{ arcsine } \sqrt{P}, \tag{6.4.32}$$

that is, twice the angle (measured in radians) whose trigonometric sine equals the square root of the proportion being transformed. Use of this transformation assumes that the number of scores on which the proportions in a data set are based is constant across cases (e.g., when the proportion of correct responses as a DV is taken as the proportion correct on a 40-item scale completed by all participants).

Table 6.4.2 gives the A values for proportions P up to .50. For A values for $P > .50$, let $P' = (1 - P)$, find A_P from the table, and then compute

$$A_{P'} = 3.14 - A_P. \tag{6.4.33}$$

For example, for the arcsine transformation of .64, find A for .36 (= 1 − .64), which equals 1.29, then find 3.14 − 1.29 = 1.85. Table 6.4.2 will be sufficient for almost all purposes. The transformation is easily calculated on a statistical calculator . First, set the calculator mode to Radians. Enter the value of the proportion, take the square root, hit $\mathbf{sin}^{-1}$, and multiply the result by 2. Statistical packages also provide this transformation.[13] See also Owen (1962, pp. 293–303) for extensive tables of the arcsine transformation.

The amount of tail stretching effected by a transformation may be indexed by the ratio of the length of the scale on the transformation of the P interval from .01 to .11 to that of the interval from .40 to .50, that is, two equal intervals, one at the end and one at the middle of the distribution, respectively. For A, this index is 2.4 (compared with 4.0 for the probit and 6.2 for the logit).

Probit transformation. This transformation is variously called *probit*, *normit*, or, most descriptively, *normalizing transformation of proportions*, a specific instance of the general normalizing transformation. We use the term *probit* in recognition of its wide use in bioassay, where it is so designated.

Its rationale is straightforward. Consider P to be the cumulative proportion of a unit normal curve (that is, a normal curve "percentile"), determine its baseline value, z_P, which is expressed in *sd* departures from a mean of zero, and add 5 to assure that the value is positive. The probit (PR) is

$$PR = z_P + 5. \tag{6.4.34}$$

Table 6.4.2 gives PR as a function of P for the lower half of the scale. When $P = 0$ and 1, PR is at minus and plus infinity, respectively, something of an embarrassment for numerical calculation. We recommend that for $P = 0$ and 1, they be revised to

$$P_0 = \frac{1}{2v} \tag{6.4.35}$$

and

$$P_1 = \frac{2v - 1}{2v}, \tag{6.4.36}$$

[13]Most statistical software provides the arcsine transformation: in SPSS, within the COMPUTE statement; in SAS, as a statement in PROC TRANSREG; in SYSTAT, in DATA(LET-ACS).

TABLE 6.4.2
Arcsine (*A*), Probit (*PR*), and Logit (*L*) Transformations for Proportions (*P*)

P	*A*	*PR*	*L*	*P*	*A*	*PR*	*L*
.000	.00	—[b]	—[b]	.16	.82	4.01	−.83
.002	.09	2.12	−3.11	.17	.85	4.05	−.79
.004	.13	2.35	−2.76	.18	.88	4.08	−.76
.006	.16	2.49	−2.56	.19	.90	4.12	−.72
.008	.18	2.59	−2.41	.20	.93	4.16	−.69
.010	.20	2.67	−2.30	.21	.95	4.19	−.66
.012	.22	2.74	−2.21	.22	.98	4.23	−.63
.014	.24	2.80	−2.13	.23	1.00	4.26	−.60
.016	.25	2.86	−2.06	.24	1.02	4.29	−.58
.018	.27	2.90	−2.00	.25	1.05	4.33	−.55
.020	.28	2.95	−1.96	.26	1.07	4.36	−.52
.022	.30	2.99	−1.90	.27	1.09	4.39	−.50
.024	.31	3.02	−1.85	.28	1.12	4.42	−.47
.026	.32	3.06	−1.81	.29	1.14	4.45	−.45
.028	.34	3.09	−1.77	.30	1.16	4.48	−.42
.030	.35	3.12	−1.74	.31	1.18	4.50	−.40
.035	.38	3.19	−1.66	.32	1.20	4.53	−.38
.040	.40	3.25	−1.59	.33	1.22	4.56	−.35
.045	.43	3.30	−1.53	.34	1.25	4.59	−.33
.050	.45	3.36	−1.47	.35	1.27	4.61	−.31
.055	.47	3.40	−1.42	.36	1.29	4.64	−.29
.060	.49	3.45	−1.38	.37	1.31	4.67	−.27
.065	.52	3.49	−1.33	.38	1.33	4.69	−.24
.070	.54	3.52	−1.29	.39	1.35	4.72	−.22
.075	.55	3.56	−1.26	.40	1.37	4.75	−.20
.080	.57	3.59	−1.22	.41	1.39	4.77	−.18
.085	.59	3.63	−1.19	.42	1.41	4.80	−.16
.090	.61	3.66	−1.16	.43	1.43	4.82	−.14
.095	.63	3.69	−1.13	.44	1.45	4.85	−.12
.100	.64	3.72	−1.00	.45	1.47	4.87	−.10
.11	.68	3.77	−1.05	.46	1.49	4.90	−.08
.12	.71	3.83	−1.00	.47	1.51	4.92	−.06
.13	.74	3.87	−.95	.48	1.53	4.95	−.04
.14	.77	3.92	−.91	.49	1.55	4.97	−.02
.15	.80	3.96	−.87	.50[a]	1.57	5.00	.00

[a] See text for values when $p > .50$.
[b] See text for transformation when $P = 0$ or 1.

where v is the denominator of the counted fraction. This is arbitrary, but usually reasonable. If in such circumstances this transformation makes a critical difference, prudence suggests that this transformation be avoided.

For *PR* values for $P > .50$, as before, let $P' = 1 - P$, find PR_P from Table 6.4.2, and then find

$$PR_{P'} = 10 - PR_P. \tag{6.4.37}$$

For example, the *PR* for $P = .83$ is found by looking up P for $.17 = (1 - .83)$ which equals 4.05, and then finding $10 - 4.05 = 5.95$. For a denser argument for probits, which maybe desirable in the tails, see Fisher and Yates (1963, pp. 68–71), but any good table of the inverse of the normal probability distribution will provide the necessary z_P values (Owen, 1962, p.12). Statistical computing packages also provide inverse distribution functions.[14]

Logit transformation. This transformation is related to the logistic curve, which is similar in shape to the normal curve but generally more mathematically tractable. The logistic distribution is discussed in more detail in Section 13.2.4 in the presentation of logistic regression. The logit transform is

$$L = \frac{1}{2} \ln \frac{P}{1 - P} \tag{6.4.38}$$

where ln is, as before, the natural logarithm (base e); the ½ is not a necessary part of the definition of the logit and is here included by convention. The relationship of values of L to P is illustrated in Fig. 13.2.1; the manner in which the logit stretches both tails of the distribution of proportions is clearly illustrated. As with probits, the logits for $P = 0$ and 1 are at minus and plus infinity, and the same device for coping with this problem (Eqs. 6.4.35 and 6.4.36) is recommended: replace $P = 0$ by $P = 1/(2v)$ and $P = 1$ by $(2v - 1)/(2v)$ and find the logits of the revised values. As before, Table 6.4.2 gives the L for P up to .50; for $P > .50$, let $P' = 1 - P$, find L_P and change its sign to positive for $L_{P'}$, that is,

$$L_{P'} = -L_P \tag{6.4.39}$$

For $P = .98$, for example, find L for $.02 (= 1 - .98)$, which equals -1.96, and change its sign, thus L for .98 is $+1.96$.

The logit stretches the tails of the P distribution the most of the three transformations. The tail-stretching index (described previously) for the logit is 6.2, compared with 4.0 for the probit and 2.4 for the arcsine.

The quantity $P/(1 - P)$ is the odds related to P (e.g., when $P = .75$, the odds are .75/.25 or simply 3). The logit, then, is simply half the natural logarithm of the odds. Therefore logits have the property that for equal intervals on the logit scale, the odds are changed by a constant multiple; for example, an increase of .35 on the logit scale represents a doubling of the odds, because .35 is ½ ln 2, where the odds are 2. The relationship of the logit to odds and their role in logistic regression is explained in detail in Section 13.2.4.

We also note the close relationship between the logit transformation of P and Fisher's z' transformation of the product-moment r (see Section 2.8.2 and Appendix Table B). If we let $r = 2P - 1$, then the z' transformation of r is the logit of P. Logit transformations are easily calculated with a statistical calculator: divide P by $(1 - P)$ and hit the ln key. Or, the computation is easily programmed within standard statistical software (see, for example, the SPSS code in Table 13.2.1).

Note that all three transformations are given in the form most frequently used or more conveniently tabled. They may be further transformed linearly if it is found convenient by the user to do so. For example, if the use of negative values is awkward, one can add a constant to L of 5, as is done for the same purpose in probits. Neither the 2 in the arcsine transformation in Eq. (6.4.32) nor the ½ in the logit transformation in Eq. (6.4.38) is necessary for purposes of correlation, but they do no harm and are tabled with these constants as part of them in accordance with their conventional definitions.

---

[14]The inverse normal function is also provided in SPSS, with the function IDF.NORMAL within the COMPUTE syntax.

The choice among the arcsine, probit, and logit transformations to achieve linearity may be guided by examining a scatterplot of each of the three transformations of the proportion against the variable with which it exhibited a nonlinear relationship in the untransformed state. A lowess line plus a linear regression line superimposed on the scatterplot will aid in discerning how well the linearization has been achieved by each transformation. Once again, the reader is warned that if the transformed proportion variable is the DV, then the fit of the regression equations with the three different transformed variables cannot be directly compared (see Section 6.4.16).

6.4.13 Variance Stabilizing Transformations and Alternatives for Treatment of Heteroscedasticity

Although we assume homoscedasticity in OLS regression, there are numerous data structures in which the predicted score $\hat{Y}_i$ is related to the variance of the residuals $sd^2_{Y|\hat{Y}}$ among individuals with that particular predicted score $\hat{Y}_i$. Often the variance increases as the predicted score increases; this is so for variables that have a lower bound of zero but no upper bound. Consider again "count" variables (e.g., the count of number of sneezes in an hour during cold season). If we take people with an average of 1 sneeze per hour, the variance in their number of sneezes over hours will be quite small. If we take people with an average of 20 sneezes per hour, the variance in number of sneezes over hours can be much larger. If we predict number of sneezes per hour in an OLS regression, we may well encounter heteroscedasticity of residuals. Now consider a measure of proportion (e.g., the proportion of days during winter flu season on which a person takes "flu" remedies). Here the variance does not simply increase as the mean proportion increases; rather, the variance increases as the mean proportion increases from 0 to .5, and then declines as the mean proportion increases further from .5 to 1.0.

Approaches to Variance Stabilization: Transformation, Weighted Least Squares, the Generalized Linear Model

Dependent variables that exhibit heteroscedasticity (nonconstant variance of the residuals) pose difficulties for OLS regression. Several approaches are taken to address the problem of heteroscedasticity. The first is transformation of the DV. Second is use of *weighted least squares regression*, presented in Section 4.5.4. Third and newest is the application of a class of regression methods subsumed under the name *generalized linear model*; this class of methods is composed of particular regression models that address specific forms of heterogeneity that commonly arise in certain data structures, such as dichotomous (binary) or count DVs. Most of Chapter 13 is devoted to two such methods: *logistic regression* for analysis of dichotomous and ordered categorical dependent variables, and *Poisson regression* for the analysis of count data. The availability of these three approaches reflects the evolution of statistical methodology. It is recommended that the reader carefully consider the developments in Chapter 13 before selecting among the solutions to the variance heterogeneity problem. Where the choice is available to transform data or to employ an appropriate form of the generalized linear model, current recommendations lean to the use of the generalized linear model.

Variance Stabilizing Transformations

The choice of variance stabilizing transformation depends on the relationship between the value of the predicted score $\hat{Y}_i$ in a regression analysis and the variance of the residuals $sd^2_{Y|\hat{Y}}$ among individuals with that particular predicted score. We discuss the use of four variance stabilizing transformations: square roots, logarithms, reciprocals, and the arcsine transformation.

The first three are one-bend transformations from the power family that we also employ to linearize relationships. The fourth is a *two-bend transformation*. That we encounter the same transformations for linearization and for variance stabilization illustrates that one transformation may, in fact, ameliorate more than one difficulty with data. We again warn, however, that a transformation that fixes one problem in the data (e.g., variance heterogeneity), may introduce another problem (e.g., nonlinearity).

We first present the three one-bend transformations from the power family Y^{λ}, the square root transformation ($\lambda = ½$), the logarithmic transformation ($\lambda = 0$) , and the reciprocal transformation ($\lambda = -1$). We then suggest approaches for selecting an approximate value of λ for variance stabilization.

Square Root Transformation ($\lambda = ½$) and Count Variables

The most likely use of a square root transformation occurs for count variables that follow a Poisson probability distribution, a positively skewed distribution of counts of rare events that occur in a specific time period, for example, counts of bizarre behaviors exhibited by individuals in a one-hour public gathering (see Section 13.4.2 for further description of the Poisson distribution). In a Poisson distribution of residuals, which may arise from a count DV, the variance of the residual scores $sd^2_{Y|\hat{Y}}$ around a particular predicted score $\hat{Y}_i$ is proportional to the predicted score $\hat{Y}_i$. Count data are handled by taking $\sqrt{Y}$. This will likely operate so as to equalize the variance, reduce the skew, and linearize relationships to other variables. A refinement of this transformation, $\sqrt{Y} + \sqrt{Y+1}$ suggested by Freeman and Tukey (1950) provides more homogeneous variances when the mean of the count variable across the data set is very low (i.e., the event being counted is very rare). Poisson regression, developed in Section 13.4, is a more appropriate approach to count dependent variables, when treatment of Y with a square root transformation fails to produce homoscedasticity.

Logarithmic Transformation ($\lambda = 0$)

The logarithmic transformation is most often employed to linearize relationships. If the variance of the residuals $sd^2_{Y|\hat{Y}}$ is proportional to the *square* of the predicted score $\hat{Y}_i^2$, the logarithmic transformation will also stabilize variances. Cook and Weisberg (1999) suggest the use of logarithmic transformations to stabilize variance when residuals are a percentage of the score on the criterion Y.

Reciprocal Transformation ($\lambda = -1$)

We encountered the reciprocal transformation in our consideration of linearizing relationships. If the residuals arise from a distribution in which the predicted score $\hat{Y}_i$ is proportional to the square of the variance of the residuals $(sd^2_{Y|\hat{Y}})^2$, the reciprocal transformation will stabilize variances.

An Estimate of λ for Variance Stabilization: The Family of Power Transformations Revisited

The choice among the square root ($\lambda = ½$); log($\lambda = 0$), or reciprocal ($\lambda = -1$) as a variance stabilizing transformation depends on the relationship between the predicted score $\hat{Y}_i$ and the variance of residuals $sd^2_{Y|\hat{Y}}$. An approach for selecting an appropriate λ for variance stabilization is described in Box 6.4.3. An alternative to this approach is to transform Y with each of the three transformations, carry out the regression analysis with each transformed DV, and examine the residuals from each analysis. The transformation that leads to the best behaved residuals is selected. Again, the reader is warned that measures of fit cannot be directly compared across

BOX 6.4.3
What Value of λ: Selecting a Variance Stabilizing Transformation From Among the Family of Power Transformations

To solve for a value of λ for variance stabilization, we find an estimate of δ that relates predicted score $\hat{Y}_i$ to the standard deviation of the residuals $sd_{Y|\hat{Y}}$, according to the expression $sd_{Y|\hat{Y}}$ is proportional to $\hat{Y}^{\delta}$ or, equivalently, $\ln sd_{Y|\hat{Y}} = \delta_0 + \delta \ln \hat{Y}$. Draper and Smith (1998) suggest that one regress untransformed Y on untransformed X, then select several vaues of the predicted score $\hat{Y}_i$, and for each of these values of $\hat{Y}_i$, find the band width (range) of the residuals (a procedure that requires a number of scores with essentially the same value of $\hat{Y}$). One then assumes that this range is approximately 4 sd_i, where sd_i is the standard deviation of the residuals for the value $\hat{Y}_i$, and plots ln sd_i as a function of ln $\hat{Y}_i$ to estimate the slope δ. To stabilize the variance of Y, we use $\lambda = (1 - \delta)$ to transform Y.

the regression equations because the dependent variables are on different scales. Strategies for model comparison across power transformations of the DV are discussed in Section 6.4.16 and in Box 6.4.1, with a complete numerical example provided in Section 6.4.17.

Box-Cox Transformation Revisited and Variance Stabilization

We introduced the Box-Cox approach to selection of λ in the context of linearization of relationships. The Box-Cox approach aims to simultaneously achieve linearization, homoscedasticity, and normality of residuals, and is applicable to the problem of variance stabilization.

Variance Stabilization of Proportions

Suppose our dependent variable were a proportion (e.g., the proportion of correct responses on a test comprised of a fixed number of items). The variance of a proportion is greatest when the proportion $P = .50$, and diminishes as P approaches either 0 or 1; specifically, $\sigma_P^2 = P(1 - P)$. The arcsine transformation introduced in Section 6.4.12 stabilizes variances.

Weighted Least Squares Regression for Variance Stabilization

Weighted least squares regression provides an alternative approach to the analysis of data that exhibit heteroscedasticity of residuals. This approach was described in detail in Section 4.5.4.

6.4.14 Transformations to Normalize Variables

We undertake transformations to normalize variables in several circumstances. One is that we have skewed Xs and/or Y. Another is that we are dealing with variables that are inherently not normally distributed, for example ranks.

Transformations to Eliminate Skew

Recall that inference in OLS regression assumes that *residuals* are normally distributed. If we analyze a data set with OLS regression and find that residuals are not normally distributed, for example, by examining a q-q plot of residuals against a normal variate (Section 4.3), then transformation of Y may be in order. Skew in the dependent variable may well be the source of the skewed residuals. Our approach, then, is to transform the DV in the hopes of achieving more normal residuals.

We can transform Y to be more normally distributed following the rules from the ladder of re-expression, that values of $\lambda > 1$ decrease negative skew, and values of $\lambda < 1$ decrease positive skew in the distribution of the transformed variable (see Section 6.4.8). Several values of λ can be tried, the transformed variable plotted as a histogram with a normal distribution overlayed or in a q-q plot against a normal variate (see Section 4.4.6). Modern statistical graphics packages provide a slider for values of λ and display the distribution of the variable as λ changes continuously. Alternatively, we may employ Box-Cox transformation of Y, which attempts to achieve more normally distributed residuals, as well as linearity and homoscedasticity.

Normalization of Ranks

A normalization strategy based on percentiles of the normal curve may be useful when data consist of *ranks*. When a third-grade teacher characterizes the aggressiveness of her 30 pupils by ranking them from 1 to 30, the resulting 30 values may occasion difficulties when they are treated numerically as measures. Ranks are necessarily rectangularly distributed; that is, there is one score of 1, one score of 2, . . . , one score of 30. If, as is likely, the difference in aggressiveness between the most and next-most (or the least and next-least) aggressive child is greater than between two adjacent children in the middle (e.g., those ranked 14 and 15), then the scale provided by the ranks is not likely to produce linear relationships with other variables. The need to stretch the tails is the same phenomenon encountered with proportions; it presupposes that the distribution of the construct to be represented has tails, that is, is bell shaped or normal. Because individual differences for many well-measured biological and behavioral phenomena seem to approximate this distribution, in the face of ranked data it is a reasonable transformation to apply in the absence of specific notions to the contrary. Even if the normalized scale is not optimal, it is likely to be superior to the original ranks.

The method for accomplishing this is simple. Following the procedure described in elementary statistics textbooks for finding centiles (percentiles), express the ranks as cumulative proportions, and refer these to a unit normal curve (Appendix Table C) to read off z_P, or use the *PR* column of Table 6.4.2, where 5 has been added to z_P to yield probits. Mosteller and Tukey (1977) suggest an alternative approach for transforming ranks, but with the same goal of normalization in mind. Other methods of addressing ordinal data are presented by Cliff (1996).

6.4.15 Diagnostics Following Transformation

We reiterate the admonition about transformation made in Section 6.4.2, that it is imperative to recheck the regression model that results from use of the transformed variable(s). Transformation may fix one difficulty and produce another. Examining whether relationships have been linearized, checking for outliers *produced by* transformation, and examining residuals are all as important after transformation as before. If difficulties are produced by transformation (e.g., heteroscedasticity of residuals), the decision may be made not to transform.[15]

---

[15]The constructed variable strategy described for Box-Cox transformation in Box 6.4.1, and Box-Tidwell in Box 6.4.2 provides an opportunity for the use of regression diagnostics to determine whether the apparent need for transformation signaled by the test of θ in Eq. (6.4.27), or of ϕ in Eq. (6.4.30) is being produced by a few outliers. An added variable plot (partial regression residual plot) is created in which the part of Y which is independent of untransformed X is plotted against the part of W in Eq. (6.4.26) for Box-Cox or V in Eq. (6.4.29) for Box-Tidwell, which is independent of untransformed X; the plot is inspected for outliers that may be producing the apparent need for transformation.

6.4.16 Measuring and Comparing Model Fit

We transform variables in part in the hope that our overall model will improve with transformation. In selecting transformations, we need to compare model fit among regression equations employing the same data but different transformations. We warn that when different nonlinear transformations of Y are employed, the R^2 values generated for the different models are not directly comparable. In other words, one cannot compare the R^2_Y resulting from predicting untransformed Y versus $R^2_{\sqrt{Y}}$ from predicting transformed $Y' = \sqrt{Y}$, versus $R^2_{\log Y}$ from predicting transformed $Y'' = \log Y$. Each dependent variable is on a different scale; the R^2s are not comparable (Kvålseth, 1985). Very misleading results with regard to the fit of models in the raw versus transformed metric may be reached by comparing the R^2 values that are reported in statistical software for these models (Alastair & Wild, 1991). To assess model fit after transformation, the predicted scores should be converted back to raw score units by reversing the transformation. For example, for $Y'' = \log Y$, the predicted scores $\hat{Y}_{\text{transformed}}$ are in logarithmic units. The antilog (Section 6.4.3) of each predicted score should be computed, yielding predicted scores in the original metric arising from the prediction of $Y' = \log Y$, that is, $e^{\hat{Y}_{\text{transformed}}} = \hat{Y}_{\text{original units}}$. (If the square root transformation were used, then we would square each predicted score to return to a predicted score in raw units.) Then two options are available for measuring fit. We may compute an index of fit as follows (Kvålseth, 1985):

$$R^2_1 = 1 - \frac{\sum (Y_i - \hat{Y}_{\text{original units}})^2}{\sum (Y_i - M_Y)^2}. \tag{6.4.40}$$

Alternatively, we may compute the correlation between the observed Y scores and $\hat{Y}_{\text{original units}}$ (Ryan, 1997), $R^2_{Y_i, \hat{Y}_{\text{original units}_i}}$, and compare these values across models. Ryan (1997) warns that both approaches may yield difficulties. First, if predicted scores are negative, then they cannot be transformed back to original units for many values of λ. Second, if $\hat{Y}_{\text{transformed}}$ is very close to zero, then the corresponding $\hat{Y}_{\text{original units}}$ may be a huge number, causing Eq. (6.4.40) to be negative. Ryan (1997) recommends use of $R^2_{Y_i, \hat{Y}_{\text{original units}_i}}$, with cases yielding negative predicted scores discarded.

6.4.17 Second-Order Polynomial Numerical Example Revisited

The data presented in Figs. 6.1.1 and 6.2.3 were actually simulated to follow a second-order polynomial with additive homoscedastic, normally distributed error. The second-order polynomial in Table 6.2.1 provides a well-fitting model with $R^2_{\text{second-order polynomial}} = .67$. In real life, we would not know the true form of the regression equation in the population that led to the observed data. We might try several transformations. What happens if we try a power transformation of Y to linearize the relationship? The bulge in the data follows Fig. 6.4.1(A), suggesting that we either transform X with $\lambda < 1.0$ or transform Y with $\lambda > 1.0$. Using Box-Cox transformation, the maximum likelihood estimate of λ is 1.73, derived from an iterative solution. We compute $Y_{\text{Box-Cox}} = Y^{1.73}$ and predict $Y_{\text{Box-Cox}}$ from untransformed X. The data, resulting linear regression line, $\hat{Y}_{\text{Box-Cox}} = 14.82X + 35.56$, plus a lowess line are shown in Fig. 6.4.2(A) (p. 236). From inspection of the lowess lines in Fig. 6.2.2(A) for untransformed Y versus Fig. 6.4.2(A) for transformed Y, the X-Y relationship appears more linear in Fig. 6.4.2(A), a result of transforming Y. However, the lowess curve in Fig. 6.4.2(A) tells us that we have not completely transformed away the curvilinear relationship in the data. Moreover, the transformation of Y has produced heteroscedasticity in Y: The spread of the Y

scores increases as X increases. As we have warned, transformation to fix one problem (here, nonlinearity) has produced another problem (nonconstant variance). We compare the fit of the polynomial model to that of the Box-Cox transformed Y model. Following Ryan (1997), we compute the predicted scores from $\hat{Y}_{\text{Box-Cox}} = 14.82X + 35.56$, which are in the transformed metric. We then convert the $\hat{Y}_{\text{Box-Cox}}$ predicted scores back to the original metric by computing $\hat{Y}_{\text{original units}} = (\hat{Y}_{\text{Box-Cox}})^{1/1.73}$. For example, for a single case $X = 7$, and observed $Y = 16.08$, $\hat{Y}_{\text{Box-Cox}} = Y^{1.73} = 16.08^{1.73} = 122.11$. The predicted score from the regression equation $\hat{Y}_{\text{Box-Cox}} = 14.82X + 35.56 = 139.29$. Finally $\hat{Y}_{\text{original units}} = (\hat{Y}_{\text{Box-Cox}})^{1/1.73} = 139.29^{1/1.73} = 17.35$. We then compute $R^2_{Y_i, \hat{Y}_{\text{original units}_i}} = .60$, the squared correlation between observed Y in its original units and the predicted score from the Box-Cox equation transformed back into original units. The Box-Cox transformation leads to a slightly less well fitting model than does the original polynomial equation. It also adds the woes of heteroscedasticity.

Suppose we focus on transforming X. The bulge rule suggests a value of $\lambda < 1$. With the left bulge, a logarithmic relationship is often helpful; we compute $\ln X$ and predict Y from $\ln X$. The resulting data, the regression equation $\hat{Y} = 8.34 \ln X + 1.34$, and a lowess curve are given in Fig. 6.4.2(B). The lowess line tells us that the logarithmic transformation succeeded in linearizing the relationship. The data look quite homoscedastic (though sparse at the low end). Because we have left Y in its original metric, the predicted scores are in the original metric as well. We do not have to transform the predicted scores before examining model fit; we may use the squared multiple correlation resulting from the regression equation $\hat{Y} = 8.34 \ln X + 1.34$, which is $R^2_{Y,\log X} = .67$, the same fit as from the second-order polynomial. With the data of Fig. 6.2.1, the second order polynomial and the logarithmic transformation are indistinguishable. The real difference between the logarithmic transformation and the quadratic polynomial is that the quadratic polynomial turns downward at the high end, as in Figure 6.1.1, but the logarithmic transformation, a one-bend transformation from the power family, does not. The data are too sparse at the high end to distinguish the polynomial equation from the logarithmic transformation. The lowess curve is not informative in this regard, due to the weakness of lowess at the ends of the X continuum. In contrast, the rectangularly distributed data in Fig. 6.2.4, with a number of cases with high values of X, would distinguish the polynomial versus logarithmic transformation; the downward turn in the data is obvious.

6.4.18 When to Transform and the Choice of Transformation

The choice between an untransformed versus a transformed analysis must take into consideration a number of factors: (a) whether strong theory, (as in psychophysics) dictates the use of transformation for estimation of critical model parameters, (b) whether the equation in the transformed metric provides a better explanation of the phenomenon under investigation than in the raw metric, for example, in the use of log dollars to reflect the utility of money, (c) whether overall fit is substantially improved by virtue of transformation, and (d) whether transformation introduces new difficulties into the model. In the behavioral sciences our focus is often on regression coefficients of particular predictors of strong theoretical interest, above and beyond an interest in overall level of prediction.

There are certainly examples of cases in which transformation yields new findings not detected in the original metric. For example, R. E. Millsap (personal communication, February 23, 2000) found evidence of salary discrimination in one of two demographic groups relative to another when salary as Y was transformed using a log metric, but not when salary was treated in the raw metric. When critical results like this differ across transformations, the researcher is pressed to develop an explanation of why the results in the transformed metric are more appropriate.

The opposite possibility exists, that is, that an important effect may be transformed away. There are instances in which we may predict a curvilinear relationship (e.g., a rise in performance as X increases to an asymptote) or an interaction between two variables (Chapter 7 is devoted to interactions). Transformation may remove the very effect we have proposed. In that case, we would obviously stay in the original metric, having once assured ourselves that the curvilinearity or interaction was not due to one or a few outliers. If the data in the original metric posed other problems (e.g., heteroscedasticity), we could retain the data in the original metric but use a more appropriate regression model, here weighted least squares regression instead of OLS regression.

In many instances, transformation may have little effect, particularly if scores contain substantial measurement error. In addition, if scores have a small range, the family of power transformations will have little effect. If data are in the form of proportions and most proportions fall between .3 and .7, or even .2 and .8, then the arcsine, logit, and probit transformation will have little effect; it is when events are very rare or very frequent (P close to 0 or 1) that transformations will make a difference. Reflection on these conditions leads us to expect that in a substantial number of cases in psychological research, (e.g., when our dependent variables are rating scales with small range), transformations will have little effect. In contrast, in areas where the DVs are physical measurements covering a large range, transformations will often be of considerable value.

An easy approach to examining whether a variable distribution (e.g., extreme skew in a predictor or the dependent variable) is producing an effect is to convert the variable to ranks[16] and repeat the analysis replacing the variable itself by its associated ranks. If the results remain the same, particularly whether theoretically important variables do or do not have an effect, then we have some confidence that the results in the raw metric are appropriate.

The choice among transformations, say the log versus square root for highly positively skewed data, will be guided by which transformation provides the better fit, given that there is no strong theoretical rationale for the choice of either. The choice will also be guided by the extent to which transformation leads to residuals that have constant variance and are normally distributed. However, the similarity of curves that are generated by different transformation equations (as illustrated in Fig. 6.4.1) coupled with random error in data mean that we may well not be able to distinguish among the transformations that may be applied to an individual data set. An interesting choice arises between polynomial regression, relatively often employed in psychology, and other transformations of the same data that lead to approximately the same fit (e.g., the use of a quadratic polynomial versus a logarithmic transformation of X). If one finds that with both transformations, the assumptions on residuals are similarly met, then interpretability in relationship to theory dictates choice. If the nonlinear relationship of X to Y is nonmonotonic, then polynomial regression must be employed; the family of power transformations handles only monotonic relationships. Finally, even when data properties point to a particular transformation, researchers should not act without simultaneously considering theoretical appropriateness.

Transformations should be tried when both violations of assumptions and evidence of nonlinearity exist and the researcher wishes to use OLS regression. The researcher should consider whether a form of the generalized linear model is more appropriate (Chapter 13). This may well be the case (e.g., the use of Poisson regression for counts of rare events).

Two alternatives exist to the use of either polynomial regression or the transformations described in Section 6.4: nonlinear least squares regression when an intrinsically nonlinear

[16]The Rank Cases procedure in SPSS ranks scores, as does rank transformation in SAS PROC TRANSREG and the rank option in the SYSTAT data module.

relationship is to be fitted, and nonparametric regression, in which no assumptions are made concerning the form of relationship of X to Y.

Sources on Transformation in Regression

The legacy of the ladder of re-expression and the bulge rule and much practical wisdom about transformation are found in Mosteller and Tukey (1977). Draper and Smith (1998) and Fox (1997) are useful starting points for further reading. Cook and Weisberg (1999) show the integration of the use of graphics and graphical software into transformation. Classic sources from mathematical statistics on transformation in regression include Atkinson (1985) and Carroll and Ruppert (1988).

6.5 NONLINEAR REGRESSION

Nonlinear regression (NR) is a form of regression analysis in which one estimates the coefficients of a nonlinear regression model that is *intrinsically nonlinear*, that is, cannot be linearized by suitable transformation (Section 6.4.4). Recall that whether an equation is intrinsically linear versus intrinsically nonlinear depends on whether the errors are assumed to be *multiplicative* versus *additive*, respectively. The nonlinear equations presented in Section 6.4.5 were all shown to be linearizable, but if and only if we assumed that the errors were multiplicative in the original metric, as was the assumption for all the models presented in Section 6.4.5. For example, when we assumed multiplicative error underlying the exponential growth model in Eq. (6.4.14), that is $Y = c(e^{dx})\varepsilon_i$, where ε represents error, the equation could be linearized to Eq. (6.4.15), $\log \hat{Y} = B_1X_1 + B_0$. If, on the other hand, we were to have assumed additive error, such that $Y = c(e^{dx}) + \varepsilon_i$, we would have needed to estimate the coefficients c and d using NR.

The use of NR begins with choice of a nonlinear model, either due to strong theory or some weaker evidence of the appropriateness of the model. The user of NR regression software must specify the particular nonlinear equation to be estimated. This is, of course, unlike the use of OLS regression or variants like WLS regression, which always employ a linear model. Ratkowsky (1990) provides graphical representations of relationships that can be useful in selecting a nonlinear model. The criterion for the choice of weights in NR is the same as in OLS regression, the least squares criterion (Section 4.3.2). However, there is not an analytic solution in the form of a set of equations (the normal equations) that we use to solve directly for the regression coefficients, as there are in OLS regression. The coefficients in NR must be found by trial and error, in an *iterative solution*. (Iterative solutions are explained in Section 13.2.9.) Iterative solutions require initial estimates of the coefficients, termed *start values* (e.g., initial estimates of the c and d coefficients in the equation $\hat{Y} = ce^{dx}$), in order that the iterative search for estimates of coefficients be successful. The values of coefficients obtained from using OLS regression to estimate the *corresponding* linearized equation (for example, the coefficients from fitting $\log \hat{Y} = B_1X_1 + B_0$) may serve as start values for NR on the same data. The regression coefficients from NR may be tested for significance under assumptions that the coefficients are asymptotically approximately normally distributed and that their variances are asymptotically approximately distributed as chi square; large sample sizes are required to approach these asymptotic conditions. An overall goodness of fit measure for the model follows the same approach as for transformed variables, given in Eq. (6.4.40).

Sources on Nonlinear Regression

In Chapter 13, we present logistic regression, a form of nonlinear regression, in some detail and also introduce another form of nonlinear regression, Poisson regression. Matters

of statistical inference, diagnostics, model fit are all explored for the logistic model and are applicable more generally to NR. Rawlings (1988) provides a highly readable introduction to nonlinear regression, and characterizes commonly used nonlinear models. Neter, Kutner, Nachtsheim and Wasserman (1996) provide an example of relevance to psychologists of fitting a common learning curve in two groups with an exponential growth model expanded to include an asymptote plus a variable representing group membership. Neter, Kutner, Nachtsheim and Wasserman (1996), Ryan (1997), and Draper and Smith (1998) provide useful practical advice and examples. Seber and Wild (1989) present a more advanced treatment.

6.6 NONPARAMETRIC REGRESSION

Nonparametric regression is an approach to discerning the pattern of the relationship of a predictor X (or set of predictors) to a dependent variable Y without first specifying a regression model, such as the familiar OLS regression model $\hat{Y} = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + B_0$. In nonparametric regression we discover the form of the relationship between X and Y by developing a smooth function relating X to Y *driven solely by the data themselves* absent any assumption about the form of the relationship. The nonparametric regression line (or curve) follows the trends in the data; the curve is smoothed by generating each point on the curve from a number of neighboring data points. The *lowess* (or *loess*) methodology explained in Chapter 4 and utilized in Fig. 6.2.2(A) is a central methodology in nonparametric regression. (See Section 4.2.1 for a discussion of smoothing and Section 4.2.2 for a discussion of lowess). Fox (2000a) provides a highly accessible introduction to nonparametric simple (one-predictor) regression; an accompanying volume (Fox, 2000b) extends to nonparametric multiple regression.

The lowess curve in Fig. 6.2.2(A) is a regression function. However, we note that it is not accompanied by a regression equation (i.e., there is no regression coefficient or regression constant). Yet we gain a great deal of information from the curve—that the relationship of X to Y is curvilinear, that there is one clearly discernable bend at low values of X, and that the relationship "bulges" to the upper left in the Mosteller and Tukey (1977) sense, illustrated in Fig. 6.4.2. We used the lowess curve in Fig. 6.2.2(A) to argue that quadratic polynomial regression was warranted to characterize the relationship. We could have gleaned further inferential information from the lowess analysis. The lowess curve in Fig. 6.2.2(A) provides a predicted score for each value of X on the lowess line: $\hat{Y}_{\text{lowess}}$. Thus it is possible to generate a measure of residual variation $SS_{\text{residual}} = \Sigma(Y_i - \hat{Y}_{\text{lowess } i})^2$, which leads to an F test of the null hypothesis that there is no relationship between X and Y. Further, since the linear regression line shown in Fig. 6.2.2(A) is nested in the more general lowess regression curve, we could have tested whether the lowess curve contributed significantly more predictability than the linear regression.

Nonparametric regression represents a new way of thinking about fitting functions to data, one that has been hardly exploited in the behavioral sciences at the time of this writing. How might we use nonparametric regression when considering the relation of X to Y? First, the lowess regression curve might be graphically presented, along with the statistical tests of relationship and nonlinearity, and the relationship described simply by the lowess curve. Second, the appearance of the lowess curve could guide the choice of transformation, either polynomial regression or one of the transformations reviewed in Section 6.4, or the selection of a function for nonlinear regression.

Nonparametric regression can be extended to multiple predictors. In the *additive nonparametric model*, a separate nonparametric regression function is fitted to each predictor (e.g., a lowess curve for each predictor). Overall fit can be tested, as can the partial contribution of

each predictor to prediction over and above the other predictors. Illustrating the shape of the regression function for two predictors as a two-dimensional irregular mountain rising from the regression plane is straightforward with modern graphical packages. Difficulty in visualizing the relationship arises with more than two predictors. Further, large sample sizes are required for multiple nonparametric regression in order to have sufficient cases at various combinations of values on all the predictors to generate the predicted scores for nonparametric regression (i.e., to develop the shape of the nonparametric regression surface). Nonetheless, nonparametric regression holds promise for highly informative exploration of relationships of predictors to a dependent variable. A classic reference in multiple nonparametric regression is Hastie and Tibshirani (1990).

6.7 SUMMARY

Multiple regression analysis may be employed to study the shape of the relationship between independent and dependent variables when these variables are measured on ordinal, interval, or ratio scales. Polynomial regression methods capture and represent the curvilinear relationship of one or more predictors to the dependent variable. Alternatively, transformations of variables in MR are undertaken to achieve linear relationships, and to eliminate heteroscedasticity and nonnormality of residuals as well so that data may be analyzed with linear MR. Nonlinear regression and nonparametric regression are also employed when data exhibit nonlinearity.

1. ***Power polynomials.*** The multiple representation of a research factor X by a series of predictors, X, X^2, X^3, etc., makes possible the fitting of regression functions of Y on X of any shape. Hierarchical MR makes possible the assessment of the size and significance of linear, quadratic, cubic (etc.), aspects of the regression function, and the multiple regression equation may be used for plotting nonlinear regression of Y on X (Section 6.2).

2. ***Orthogonal polynomials.*** For some purposes (for example, laboratory experiments where the number of observed values of X is not large), it is advantageous to code X so that the X_i not only carry information about the different curve components (linear, quadratic, etc.) but are orthogonal to each other as well. Some interpretive and computational advantages and alternate error models are discussed (Section 6.3).

3. ***Nonlinear transformations.*** Nonlinear transformations are one-to-one mathematical relationships that change the relative spacing of scores on a scale (e.g., the numbers 1, 10, 100 versus their base_{10} logs 0, 1, 2). Nonlinear transformations of predictors X and/or the dependent variable Y are carried out for three reasons. First, they are employed to simplify relationships between predictors and the DV; simplification most often means linearization of the relationship so that the relationship can be examined in linear MR. Second, they are employed to stabilize the variances of the residuals, that is, to eliminate heteroscedasticity of residuals. Third, they are used to normalize residuals. Homoscedasticity and normality of residuals are required for inference in linear MR. The circumstances in which logarithmic, square root, and reciprocal transformations are likely to be effective for linearization are described. Such transformations arise frequently in conjunction with formal mathematical models that are expressed in nonlinear equations, for example in exponential growth models. More generally, a full family of power transformations are employed for linearization. To select among transformations, graphical and statistical methods are employed; these include the ladder of re-expression and the bulging rule, plus the Box-Cox and Box-Tidwell methodologies. Tail-stretching transformations of proportions are also employed; they include the arcsine, probit, and logit transformations. Transformations also serve to render residuals homoscedastic and normal, so that data are amenable to treatment in linear MR (Section 6.4).

4. *Nonlinear regression* (NR). Nonlinear regression is a form of regression analysis in which one estimates the coefficients of a nonlinear regression model that is *intrinsically nonlinear*, that is, cannot be linearized by suitable transformation (Section 6.5).

5. *Nonparametric regression.* Nonparametric regression is an approach to discerning the pattern of the relationship of a predictor X (or set of predictors) to a dependent variable Y without first specifying a regression model. In nonparametric regression the form of the relationship between X and Y is discerned by developing a smooth function relating X to Y driven solely by the data themselves absent any assumption about the form of the relationship (Section 6.6).

7

Interactions Among Continuous Variables

7.1 INTRODUCTION

In this chapter we extend MR analysis to interactions among continuous predictors. By *interactions* we mean an interplay among predictors that produces an effect on the outcome Y that is different from the sum of the effects of the individual predictors. Many theories in the social sciences hypothesize that two or more continuous variables interact; it is safe to say that the testing of interactions is at the very heart of theory testing in the social sciences. Consider as an example how ability (X) and motivation (Z) impact achievement in graduate school (Y). One possibility is that their effects are additive. The combined impact of ability and motivation on achievement equals the sum of their separate effects; there is no interaction between X and Z. We might say that the whole equals the sum of the parts. A second alternative is that ability and motivation may interact synergistically, such that graduate students with both high ability and high motivation achieve much more in graduate school than would be expected from the simple sum of the separate effects of ability and motivation. Graduate students with both high ability and high motivation become "superstars"; we would say that the whole is greater than the sum of the parts. A third alternative is that ability and motivation compensate for one another. For those students who are extremely high in ability, motivation is less important to achievement, whereas for students highest in motivation, sheer native ability has less impact. Here we would say that the whole is less than the sum of the parts; there is some partial trade-off between ability and motivation in the prediction of achievement. The second and third alternatives exemplify interactions between predictors, that is, combined effects of predictors that differ from the sum of their separate effects.

When two predictors in regression analysis interact with one another, the regression of Y on one of those predictors *depends on* or is *conditional on* the value of the other predictor. In the second alternative, a *synergistic interaction* between ability X and motivation Z, the regression coefficient for the regression of achievement Y on ability X increases as motivation Z increases. Under the synergistic model, when motivation is very low, ability has little effect because the student is hardly engaged in the graduate school enterprise. When motivation is higher, then more able students exhibit greater achievement.

Continuous variable interactions such as those portrayed in alternatives two and three can be tested in MR analysis, treating both the original variables and their interaction as continuous

predictors. In this chapter we explore how to specify interactions between continuous variables in multiple regression equations, how to test for the statistical signficance of interactions, how to plot them, and how to interpret them through post hoc probing.

We suspect that some readers are familiar with the testing, plotting, post hoc probing, and interpretation of interactions between categorical variables in the analysis of variance (ANOVA) context. Historically, continuous variable interactions have often been analyzed by breaking the continuous variables into categories, so that interactions between them can be examined in ANOVA. For example, an analyst might perform median splits on ability and motivation to create four combinations (hi-hi, hi-lo, lo-hi, and lo-lo) of ability and motivation that could be examined in a 2×2 ANOVA. *This dichotomization strategy is ill-advised, and we strongly recommend against it.* The strategy evolved because methods were fully developed for probing interactions in ANOVA long before they were fully developed in MR. Dichotomization is problematic first because it decreases measured relationships between variables. For example, dichotomization at the median of a single continuous normally distributed predictor X reduces its squared correlation with a normally distributed dependent variable Y to .64 of the original correlation (Cohen, 1983). Dichotomization of a single predictor is equivalent to throwing out over a third of the cases in the data set. Dichotomization of two continuous variables X and Z so that their interaction can be examined in ANOVA lowers the power for detecting a true nonzero interaction between the two continuous predictors. As Maxwell and Delaney (1993) point out, if loss of power were the only impact of dichotomization and researchers found significance nonetheless after dichotomization, the practice might not seem so undesirable from a theoretical standpoint. But Maxwell and Delaney (1993) show much more deleterious effects from a validity standpoint. Carrying out median splits on two continuous predictors X and Z can produce spurious main effects, that is, effects of the individual predictors that are "significant" when the dichotomized data are analyzed, although the effects do not, in fact, exist in the population. Moreover, in one special circumstance in which there is no true interaction between two continuous predictors X and Z, a spurious interaction may be produced between the dichotomized predictors. This can happen if one of the predictors X or Z has a quadratic relationship to Y.

In this chapter we provide prescriptions for specifying, plotting, testing, post hoc probing, and interpretating interactions among continuous variables. In Chapter 8, we introduce the implementation of true categorical predictors (e.g., gender, ethnicity) in MR. In Chapter 9, we extend MR to interactions among categorical variables and between categorical and continuous variables.

7.1.1 Interactions Versus Additive Effects

Regression equations that contain as IVs only predictors taken separately signify that the effects of continuous variables such as X and Z are *additive* in their impact on the criterion, that is,

$$\hat{Y} = B_1X + B_2Z + B_0. \tag{7.1.1}$$

Note that Eq. (7.1.1) is the same equation as Eq. (3.2.1), except that the notation X and Z has been substituted for X_1 and X_2, respectively.

For a specific instance, consider the following numerical example:

$$\hat{Y} = .2X + .6Z + 2.$$

The estimated DV increases .2 points for each 1-point increase in X and another .6 points for each 1-point increase in Z. (Strictly speaking, this is correct only if X and Z are uncorrelated. If

they are correlated, these effects hold only when the two IVs are used together to estimate Y.) The effects of X and Z are additive. By *additivity* is meant that the regression of the criterion on one predictor, say predictor X, is constant over all values of the other predictor Z.

Interactions as Joint Effects

In Eq. (7.1.2) we add a predictor XZ to carry an interaction between X and Z:

$$\hat{Y} = B_1X + B_2Z + B_3XZ + B_0. \tag{7.1.2}$$

Literally, the predictor is the product of scores on predictors X and Z, calculated for each case. While the interaction is carried by the XZ product term, the interaction itself is actually that part of XZ that is independent of X and Z, from which X and Z have been partialed (more about this in Section 7.6).

Consider our numerical example, but with the product term added:

$$\hat{Y} = .2X + .6Z + .4XZ + 2.$$

If X and Z are uncorrelated, the criterion Y increases .2 points for each 1-point increase in X and an additional .6 points for each 1-point increase in Z. Moreover, the criterion Y increases an additional .4 points for a 1-point increment in the part of the cross-product XZ that is independent of X and Z. The partialed component of the cross-product represents a unique combined effect of the two variables working together, above and beyond their separate effects; here a *synergistic* effect, as in the example of ability X and motivation Z as predictors of graduate school achievement Y. Thus two variables X and Z are said to interact in their accounting for variance in Y when *over and above* any additive combination of their separate effects, they have a *joint effect.*

We can compare the joint or interactive effect of X and Z with the simple additive effects of X and Z in three-dimensional graphs. For data, we plot 36 cases for which we have scores on predictors X and Z (see Table 7.1.1A). Both X and Z take on the values 0, 2, 4, 6, 8, and 10; the 36 cases were created by forming every possible combination of one X value and one Z value. This method of creating cases makes X and Z uniformly distributed, that is, produces an equal number of scores at each value of X and Z. The method also assures that X and Z are uncorrelated. These special properties facilitate the example but are not at all necessary or typical for the inclusion of interactions in MR equations. The means and standard deviations of X and Z, as well as their correlations with Y, are given in Table 7.1.1B.

Figure 7.1.1(A) illustrates the additive effects (absent any interaction) of X and Z from the equation $\hat{Y} = .2X + .6Z + 2$. Predictors X and Z form the axes on the floor of the graph; all 36 cases (i.e., points representing combinations of values of the predictors) lie on the floor. Predicted $\hat{Y}$s for each case (unique combinations of X and Z) were generated from the regression equation. The *regression plane*, the tilted plane above the floor, represents the location of $\hat{Y}$ for every possible combination of values of X and Z. Note that the regression plane is a flat surface. Regardless of the particular combination of values of X and Z, the $\hat{Y}$ is incremented (geometrically raised off the floor) by a constant value relative to the values of X and Z, that is, by the value $(.2X + .6Z)$.

The regression plane in Fig. 7.1.1(B) illustrates the additive effects of X and Z plus the interaction between X and Z in the equation $\hat{Y} = .2X + .6Z + .4XZ + 2$. The same 36 combinations of X and Z were used again. However, $\hat{Y}$s were generated from the equation containing the interaction. Table 7.1.1B gives the mean and standard deviation of the product term that carries the interaction, and its correlation with the criterion Y. In Fig. 7.1.1(B) the regression plane is now a stretched surface, pulled up in the corner above the height of the

TABLE 7.1.1
Multiple Regression Equations Containing Interactions: Uncentered Versus Centered Predictors

A. Thirty-six cases generated from every possible combination of scores on predictors X and Z.

X (0, 2, 4, 6, 8, 10)
Z (0, 2, 4, 6, 8, 10)

Cases (X, Z combinations)
(0, 0), (0, 2), . . . , (4, 6), . . . , (6, 8), . . . , (10, 10)

B. Summary Statistics for X, Z, and XZ (uncentered, in raw score form).

Means and standard deviations			Correlation matrix				
	M	*sd*		*X*	*Z*	*XZ*	*Y*
X	5.000	3.464	*X*	1.00	0.00	.637	.600
Z	5.000	3.464	*Z*		1.00	.637	.709
XZ	25.000	27.203	*XZ*			1.00	.995

C. Unstandardized regression equations: prediction of Y from X and Z, and from X, Z, and XZ (uncentered, in raw score form).

1. Uncentered regression equation, no interaction:

$$\hat{Y} = .2X + .6Z + 2$$

2. Uncentered regression equation, with interaction:

$$\hat{Y} = .2X + .6Z + .4XZ + 2$$

D. Simple regression equations for Y on X at values of Z with uncentered predictors and criterion.

$$\text{At } Z_{high}: \quad \hat{Y} = 3.4X + 6.8$$
$$\text{At } Z_{mean}: \quad \hat{Y} = 2.2X + 5.0$$
$$\text{At } Z_{low}: \quad \hat{Y} = 1.0X + 3.2$$

E. Summary statistics for x, z and xz (centered, in deviation form).

Means and standard deviations			Correlation matrix				
	M	*sd*		*x*	*z*	*xz*	*Y*
x	0.000	3.464	*x*	1.00	.000	.000	.600
z	0.000	3.464	*z*		1.00	.000	.709
xz	0.000	11.832	*xz*			1.00	.372

F. Unstandardized regression equations: prediction of Y from x and z, and from x, z, and xz (centered, in deviation form).

1. Centered regression equation, no interaction:

$$\hat{Y} = .2x + .6z + 6$$

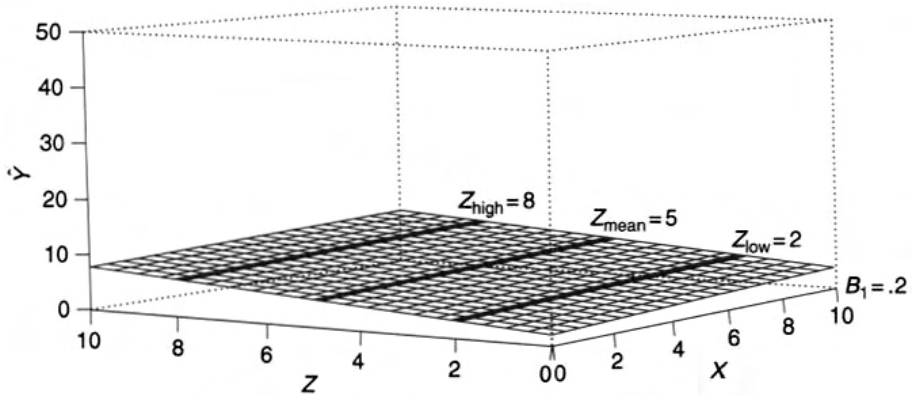
2. Centered regression equation, with interaction:

$$\hat{Y} = 2.2x + 2.6z + .4xz + 16$$

G. Simple regression equations for Y on x at values of z with centered predictors and criterion.

$$\text{At } z_{high}: \quad \hat{Y} = 3.4x + 23.8$$
$$\text{At } z_{mean}: \quad \hat{Y} = 2.2x + 16.0$$
$$\text{At } z_{low}: \quad \hat{Y} = 1.0x + 8.2$$

(A) Regression surface: $\hat{Y} = .2X + .6Z + 2$



(B) Regression surface: $\hat{Y} = .2X + .6Z + .4XZ + 2$

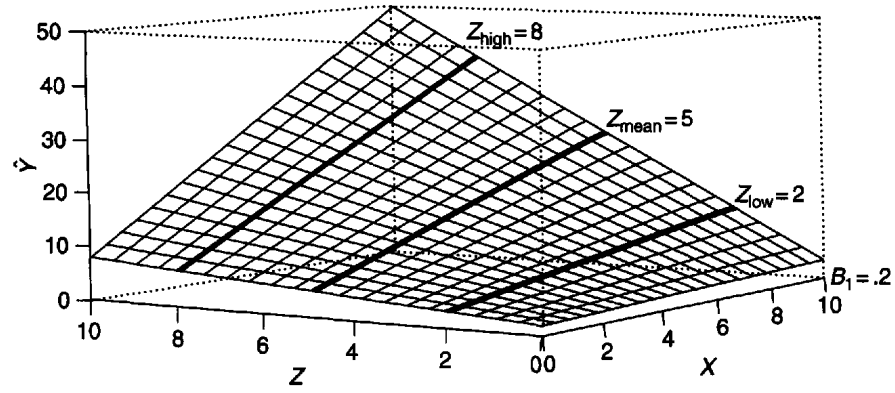


FIGURE 7.1.1 Regression surface predicated in (A) an additive regression equation containing no interaction and (B) a regression equation containing an interaction. Predictors and criterion are in raw score (uncentered) form.

flat regression plane in Fig. 7.1.1(A). The amount by which the stretched surface is lifted above the flat regression plane represents unique variance due to the interaction of X and Z, over and above the individual additive effects of X and Z. What is the source of the upward stretching? The stretching occurs because the increment in Y depends not only on additive values of X and Z but also on their product XZ, and the product XZ increases in a curvilinear fashion as X and Z increase linearly. Note the dramatic rise in the product XZ relative to the sum $X + Z$:

X	0	2	4	6	8	10
Z	0	2	4	6	8	10
$X+Z$	0	4	8	12	16	20
XZ	0	4	16	36	64	100

7.1.2 Conditional First-Order Effects in Equations Containing Interactions

As in polynomial regression explained in Chapter 6, we make the distinction between *first-order effects* and *higher order effects* in regression equations containing interactions. First-order effects refer to the effects of the individual predictors on the criterion. Higher order effects refer

to the partialed effects of multiplicative functions of the individual predictors, for example the XZ term with X and Z partialed out in Eq. (7.1.2).

When the effects of individual predictors are purely additive, as in Eq. (7.1.1), the first-order regression coefficient for each predictor is constant over all values of the other predictor (again, this is the definition of *additivity*). The constancy is illustrated in Fig. 7.1.1(A). In Fig. 7.1.1(A), three lines on the regression plane defined by $\hat{Y} = .2X + .6Z + 2$ are darkened: at $Z = 2, Z = 5$ (i.e., the mean of Z, M_Z) and $Z = 8$. These lines show the regression of Y on X at each of these three values of Z: $Z_{\text{low}}, Z_{\text{mean}}$, and Z_{high}, respectively. These three regression lines are parallel, signifying that the regression of Y on X is constant over values of Z. Thus the regression coefficient for the X predictor applies equally across the range of Z. The only characteristic that varies across the three regression lines is the overall height of the regression line (distance from the floor of the graph). The displacement upward of the lines as Z increases signifies that as Z increases, the criterion Y increases as well (a first-order effect). On average, values of Y are higher for higher values of Z.

Figure 7.1.1(B) represents the regression equation $\hat{Y} = .2X + .6Z + .4XZ + 2$. Regression lines for the regression of Y on X are drawn at the same three values of Z as in Fig. 7.1.1(A): $Z = 2, 5, 8$. We see immediately that the regression line for Y on X becomes steeper as Z increases. The regression of Y on X is not constant over all values of Z but depends specifically on the particular value of Z at which the regression of Y on X is taken. Predictors X and Z are no longer additive in their effects on Y; they are interactive. The regression of Y on X is *conditional upon* (i.e., depends upon) the value of Z. In regression equations containing interactions, the *first-order effects* of variables are conditional on (depend upon, or are *moderated by*) the values of the other predictors with which they interact.

We have cast this discussion of *conditional effects* in terms of the regression of Y on X at values of Z. However, the interaction between X and Z is symmetric. We could examine the regression of Y on Z at values of X. The result would be the same: the regression of Y on Z would differ as a function of X; that is, the regression of Y on Z is again conditional upon the value of X.

Now we focus on the angle formed between the regression plane and the floor of Fig. 7.1.1(A). This angle is best seen at the right edge of Fig. 7.1.1(A) (predictor X), where $Z = 0$. In Fig. 7.1.1(A), with no interaction, the slope of the regression of Y on X equals .2 at $Z = 0$. Recall that .2 is the regression coefficient for Y on X in Eq. (7.1.1). This same angle is maintained across the range of Z, which is another way of saying that the regression of Y on X is constant across all values of Z, meeting the definition of additivity.

Examine the right edge of Fig. 7.1.1(B) (predictor X), where $Z = 0$. The regression of Y on X also equals .2 at $Z = 0$ in Fig. 7.1.1(B), and the regression coefficient B_1 for Y on X in our numerical example containing an interaction is .2. However, in Fig. 7.1.1(B), the slope of the regression of Y on X is only .2 at $Z = 0$. As Z increases, the slope of Y on X also increases. Thus the numerical value of the regression coefficient $B_1 = .2$ is only an accurate representation of the regression of Y on X at one point on the regression plane. In general, in a regression equation containing an interaction, the first-order regression coefficient for each predictor involved in the interaction represents the regression of Y on that predictor, *only at the value of zero on all other individual predictors with which the predictor interacts*. The first-order coefficients have different meanings depending on whether the regression equation does or does not include interactions. To reiterate, without an interaction term the B_1 coefficient for X represents the overall effect of X on Y across the full range of Z. However, in Eq. (7.1.2), the B_1 coefficient for X represents the effect of X on the criterion only at $Z = 0$.

7.2 CENTERING PREDICTORS AND THE INTERPRETATION OF REGRESSION COEFFICIENTS IN EQUATIONS CONTAINING INTERACTIONS

The interpretation of the first-order coefficients B_1 and B_2 in the presence of interactions is usually problematic in typical social science data. The B_1 coefficient represents the regression of Y on X at $Z = 0$, and the B_2 coefficient represents the regression of Y on Z at $X = 0$. Only rarely in the social sciences is zero a meaningful point on a scale. For example, suppose, in a developmental psychology study, we predict a level of language development (Y) of children aged 2 to 6 years from mother's language development (D), child's age (A), and the interaction of mother's language development and child's age, carried by the DA term. In the regression equation $\hat{Y} = B_1D + B_2A + B_3DA + B_0$, the regression coefficient B_1 of child's language development on mother's language development D is at child's age $A = 0$, not a useful value in that all children in the study fall between ages 2 and 6. To interpret this B_1 coefficient, we would have to extrapolate from our sample to newborns in whom the process of language development has not yet begun. (Our comments about the dangers of extrapolation in Section 6.2.5 apply here as well.)

7.2.1 Regression With Centered Predictors

We can make a simple linear transformation of the age predictor that renders zero on the age scale meaningful. Simply, we *center* age, that is, put age in deviation form by subtracting M_A from each observed age (i.e., $a = A - M_A$). If age were symmetrically distributed over the values 2, 3, 4, 5, and 6 years, $M_A = 4$ years, and the centered age variable a would take on the values $-2, -1, 0, 1, 2$. The mean of the centered age variable a necessarily would be zero. When a is used in the regression equation $\hat{Y} = B_1D + B_2a + B_3Da + B_0$, the B_1 coefficient represents the regression of child's language development on mother's language development at the mean age of the children in the sample. This strategy of centering to make the regression coefficients of first-order terms meaningful is identical to the use of centering in polynomial regression (Section 6.2.3.).

The symmetry in interactions applies to centering predictors. If we center mother's language development into variable $d = D - M_D$ and estimate the regression equation $\hat{Y} = B_1d + B_2A + B_3dA + B_0$, then the B_2 coefficient represents the regression of child's language development on child's age at the mean of mother's language development in the sample.

Finally, suppose we wish to assess the interaction between age and mother's language development. We center both predictors and form the product of the centered variables da to carry the interaction and estimate the regression equation $\hat{Y} = B_1d + B_2a + B_3da + B_0$. Both the B_1 and B_2 coefficients represent the first-order relationships at the *centroid* (mean on both predictors) of the sample. The regression equation characterizes the typical case. In sum, if all the predictors in a regression equation containing interactions are centered, then each first-order coefficient has an interpretation that is meaningful in terms of the variables under investigation: the regression of the criterion on the predictor at the sample means of all other variables in the equation.

With centered predictors, each first-order regression coefficient has yet a second meaningful interpretation, as the *average regression* of the criterion on the predictor across the range of the other predictors. In the developmental study, if the d by a interaction were nonzero, then the regression of child's language development on mother's language development would differ at each age. Assume that there were an equal number of children at each age. Imagine computing the regression coefficient B_1 of child's language development on mother's language

development separately at each age and then averaging all these B_1 coefficients. The B_1 coefficient for the impact of mother's language development in the overall centered regression equation containing all ages would equal the average of the individual B_1 coefficients at each child's age. If there were an unequal number of children at each age, then the overall B_1 coefficient would equal the weighted average of the individual B_1 coefficients, where the weights were the number of children at each age. In sum, when predictors are centered, then each first-order coefficient in a regression equation containing interactions is the *average regression of the criterion on a predictor* across the range of the other predictors in the equation.

7.2.2 Relationship Between Regression Coefficients in the Uncentered and Centered Equations

As noted in Chapter 2, correlational properties of variables do not change under linear transformation of variables. Linear transformations include adding or subtracting constants, and multiplying and dividing by constants. If we correlate height in inches with weight in pounds, we obtain the same value as if we correlate height in inches with weight in ounces or kilograms. *Centering*, or putting predictors in deviation score form by subtracting the mean of the predictor from each score on the predictor, is a linear transformation. *Thus our first intuition might be that if predictors were centered before they were entered into a regression equation, the resulting regression coefficients would equal those from the uncentered equation. This intuition is correct only for regression equations that contain no interactions.*

As we have seen, centering predictors provides tremendous interpretational advantages in regression equations containing interactions, but centering produces a very puzzling effect. When predictors are centered and entered into regression equations containing interactions, the regression coefficients for the first-order effects B_1 and B_2 are different numerically from those we obtain performing a regression analysis on the same data in raw score or *uncentered* form. We encountered an analogous phenomenon in Chapter 6 in polynomial regression; when we centered the predictor X, the regression coefficient for all but the highest order polynomial term changed (see Section 6.2.3). The explanation of this phenomenon is straightforward and is easily grasped from three-dimensional representations of interactions such as Fig. 7.1.1(B). An understanding of the phenomenon provides insight into the meaning of regression coefficients in regression equations containing interactions.

7.2.3 Centered Equations With No Interaction

We return to the numerical example in Table 7.1.1 and Fig. 7.1.1. The means of both predictors X and Z equal 5.00. Uncentered and centered X and Z would be as follows:

$X_{\text{uncentered}}$	0	2	4	6	8	10
x_{centered}	−5	−3	−1	1	3	5

and

$Z_{\text{uncentered}}$	0	2	4	6	8	10
z_{centered}	−5	−3	−1	1	3	5

Now, assume that we keep the criterion Y in its original uncentered metric, but we use x and z, and re-estimate the regression equation without an interaction. The resulting regression equation is

$$\hat{Y} = .2x + .6z + 6.$$

The regression coefficients for x and z equal those for uncentered X and Z. Only the regression intercept has changed. From Chapter 3, Eq. (3.2.6), the intercept is given as

$B_0 = M_Y - B_1M_X - B_2M_Z$. Centering X and Z changed their means from 5.00 to 0.00, leading to the change in B_0. In fact, there is a simple algebraic relationship between B_0 in the centered versus uncentered equations. For the uncentered regression equation $\hat{Y} = B_1X + B_2Z + B_0$ versus the centered regression equation $\hat{Y} = B_1x + B_2z + B_0$,

(7.2.1) $$B_{0,\text{centered}} = B_{0,\text{uncentered}} + B_{1,\text{uncentered}}M_{X,\text{uncentered}} + B_{2,\text{uncentered}}M_{Z,\text{uncentered}}$$

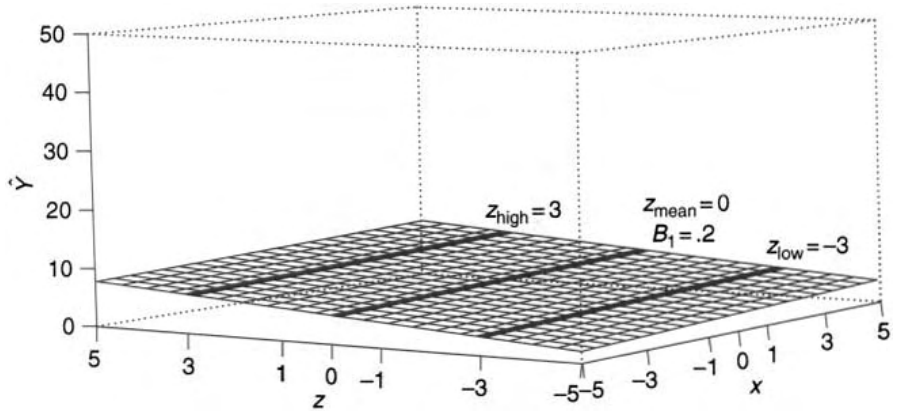
For our example, this is

$$B_{0,\text{centered}} = 2 + .2(5.00) + .6(5.00) = 6.$$

The centered regression equation is plotted in Fig. 7.2.1(A). The only difference between Fig. 7.1.1(A) and Fig. 7.2.1(A) is that the scales of the X and Z axes in Fig. 7.2.1(A) have been changed from those in Fig. 7.1.1(A) to reflect centering. Note that $x = 0$ and $z = 0$ in Fig. 7.2.1(A) are now in the *middle of the axes*, rather than at one end of the axes, as in Fig. 7.1.1(A). Note also that the criterion Y is left uncentered.

Figure 7.2.1(A) confirms the numerical result that the regression coefficients B_1 and B_2 do not change when we center predictors in regression equations containing no interactions.

(A) Regression surface from centered regression equation: $\hat{Y} = .2x + .6z + 6$



(B) Regression surface from centered regression equation: $\hat{Y} = 2.2x + 2.6z + .4xz + 16$

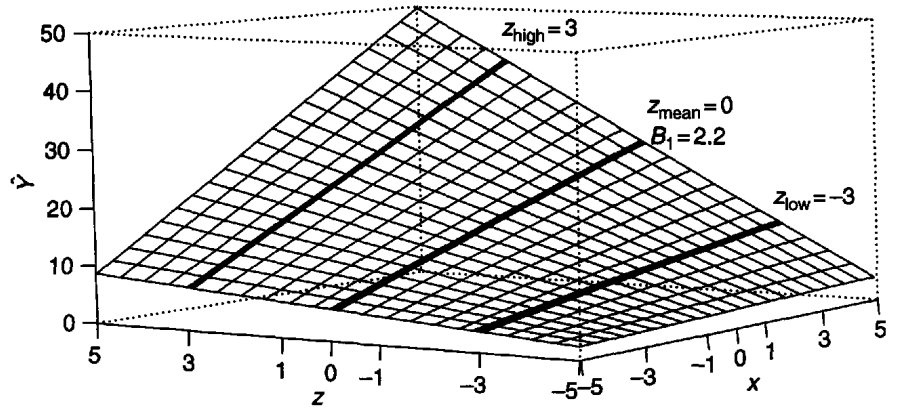


FIGURE 7.2.1 Regression surface predicted in (A) an additive regression equation containing no interaction and (B) a regression equation containing an interaction. Predictors are in centered (deviation) form.

Consider B_1; the slope of Y on X at $Z = 0$ in Fig. 7.1.1(A) is the same as that for Y on x at $z = 0$ in Figure 7.2.1(A), though the location of $Z = 0$ differs across the figures.

A comparison of Figures 7.1.1(A) and 7.2.1(A) also confirms the change in intercept. The intercept is the height of the regression plane from the floor at the point $X = 0, Z = 0$. In Fig. 7.1.1(A) for the uncentered equation, this point is in the lower right-hand corner of the plane; here the plane is only two units from the floor, so $B_0 = 2$. As pointed out earlier, in Fig. 7.2.1(A) for the centered equation, the point $x = 0, z = 0$ is now in the center of the regression plane. (When all predictors have been centered, the value 0, 0 is the *centroid* of the predictor space.) The overall elevation of the plane is farther from the floor at this point, specifically six units from the floor, so $B_0 = 6$. The change in the location of the point $X = 0, Z = 0$ produced by centering also produces the change in the intercept.

7.2.4 Essential Versus Nonessential Multicollinearity

The correlation matrix among the centered predictors, including xz, and of the centered predictors with the criterion is given in Table 7.1.1E. This correlation matrix should be compared with that in Table 7.1.1B for the uncentered predictors. There is one dramatic change when predictors are uncentered versus centered. The correlations of the X and Z terms with XZ ($r = .637$ in each case) are substantial in the uncentered case but fall to zero in the centered case. This drop is another example of *essential* versus *nonessential multicollinearity* (Marquardt, 1980), also encountered in our work with polynomial regression equations (Section 6.2.3).

Algebraically, the covariance (numerator of the correlation coefficient) between X and XZ is in part a function of the arithmetic means of X and Z. If X and Z are each completely symmetrical, as in our numerical example, then the covariance (cov) between X and XZ is as follows (Aiken & West, 1991, p. 180, eq. A.15):

$$\text{cov}(XZ, X) = sd_X^2 M_Z + \text{cov}(X, Z) M_X$$

If X and Z are centered, then M_X and M_Z are both zero, and the covariance between x and xz is zero as well. Thus the correlation between x and xz is also zero. The same holds for the correlation between z and xz. The amount of correlation that is produced between X and XZ or Z and XZ by the nonzero means of X and Z, respectively, is referred to as *nonessential multicollinearity* (Marquardt, 1980). This nonessential multicollinearity is due purely to scaling—when variables are centered, it disappears. The amount of correlation between X and XZ that is due to skew in X cannot be removed by centering. This source of correlation between X and XZ is termed *essential multicollinearity*. The same is true for the correlation between Z and XZ.

7.2.5 Centered Equations With Interactions

Now consider the use of centered predictors in an equation containing an interaction. How do the coefficients of the uncentered equation $\hat{Y} = B_1X + B_2Z + B_3XZ + B_0$ relate to those in the centered equation $\hat{Y} = B_1x + B_2z + B_3xz + B_0$? For the same reason as in the equation without an interaction, the intercept changes here. However, B_1 and B_2 also change, often dramatically, in association with the changes in the correlation matrix of predictors just described.

In the numerical example of Table 7.1.1, the centered equation is

$$\hat{Y} = 2.2x + 2.6z + .4xz + 16.$$

This equation was found by retaining the criterion Y in raw score form, centering X and Z into x and z, respectively, forming the cross-product of centered X and Z (i.e., xz), and predicting

Y from x, z, and xz. Note that the intercept B_0 has changed from $B_0 = 2$ in the uncentered regression equation to 16 in the centered equation. Coefficients B_1 and B_2 have changed from .2 and .6, respectively, to 2.2 and 2.6, respectively. As we will see, these changes do not mean that the relationships of X and Z to the criterion Y have somehow changed with centering.

The centered regression equation containing an interaction is plotted in Fig. 7.2.1(B). As noted earlier, the value $x = 0, z = 0$ has moved from the lower right-hand corner of the regression plane in Fig. 7.1.1(B) to the middle of the regression plane in Fig. 7.2.1(B) due to centering.

A comparison of Fig. 7.1.1(B) with Fig. 7.2.1(B) gives insight into the source of the change in regression coefficients. In the uncentered equation, the B_1 coefficient represented the regression of Y on X at $Z = 0$, at the far right edge of Fig. 7.1.1(B). For higher values of Z (moving left along Fig. 7.1.1(B), the regression of Y on X became increasingly steep. With centered z, in Fig. 7.2.1(B), the value $z = 0$ is no longer at the right edge of the figure; it is halfway up the regression plane. At $z_{\text{mean}} = 0$ the regression of Y on x has risen to 2.2, the value of B_1 in the centered regression equation.

In general, centering predictors moves the value of zero on the predictors along the regression surface. If the regression surface is a flat plane (i.e., the regression equation contains no interaction), then the regression of Y on X is constant at all locations on the plane. Moving the value of zero by linear transformation has no effect on the regression coefficient for the predictor. If the regression surface is not flat (i.e., the regression equation contains an interaction), then the regression of Y on X varies across locations on the plane. The value of the B_1 regression coefficient will always be the slope of Y on X at $Z = 0$ on the plane, but the location of $Z = 0$ on the plane will change with centering.

What about the interpretation of B_1 as the *average* regression slope of Y on x across all values of z in the centered regression equation, $\hat{Y} = B_1x + B_2z + B_3xz + B_0$? A closer examination of Fig. 7.2.1(B) confirms this interpretation. In Fig. 7.2.1, the far right-hand edge now is at $z = -5$; at this point the regression of Y on X is .2. At the far left edge, $z = 5$ and the slope of the regression of Y on X is 4.2. The distribution of Z is uniform, so the average slope across all cases represented in the observed regression plane is $(.2 + 4.2)/2 = 2.2$; this is the value of the B_1 coefficient. Thus B_1 is the average slope of the regression of Y on X across all values of centered predictor Z.

There are straightforward algebraic relationships between the B_0, B_1, and B_2 coefficients in the uncentered versus centered regression equation containing the interactions:

$$\begin{aligned} B_{1,\text{centered}} &= B_{1,\text{uncentered}} + B_{3,\text{uncentered}}M_{Z,\text{uncentered}}; \\ B_{2,\text{centered}} &= B_{2,\text{uncentered}} + B_{3,\text{uncentered}}M_{X,\text{uncentered}}. \end{aligned} \tag{7.2.2}$$

For our numerical example,

$$B_{1,\text{centered}} = .2 + .4(5.00) = 2.20, \quad \text{and} \quad B_{2,\text{centered}} = .6 + .4(5.00) = 2.60.$$

Note that if there is no interaction (i.e., $B_3 = 0$), then the B_1 and B_2 coefficients would remain the same if X and Z were centered versus uncentered. This confirms what we know—*only if there is an interaction does rescaling a variable by a linear transformation change the first order regression coefficients.*

For the relationship of the intercept $B_{0,\text{centered}}$ to $B_{0,\text{uncentered}}$, we have

$$\begin{aligned} B_{0,\text{centered}} = B_{0,\text{uncentered}} &+ B_{1,\text{uncentered}}M_{X,\text{uncentered}} + B_{2,\text{uncentered}}M_{Z,\text{uncentered}} \\ &+ B_3M_{X,\text{uncentered}}M_{Z,\text{uncentered}}. \end{aligned} \tag{7.2.3}$$

For our numerical example

$$B_{0,\text{centered}} = 2 + .2(5.00) + .6(5.00) + .4(5.00)(5.00) = 16.$$

Equations (7.2.1), (7.2.2), and (7.2.3) pertain only to Eq. (7.1.1). These relationships differ for every form of regression equation containing at least one interaction term; they would be different for more complex equations, for example, Eqs. (7.6.1) and (7.9.2) given below. Aiken and West (1991, Appendix B) provide an extensive mapping of uncentered to centered regression equations.

7.2.6 The Highest Order Interaction in the Centered Versus Uncentered Equation

By inspection the shapes of the regression surfaces in Fig. 7.1.1(B) for uncentered data and Fig. 7.2.1(B) for centered data are identical. Consistent with this, there is no effect of centering predictors on the value of regression coefficient B_3 in Eq. (7.1.2). The B_3 coefficient is for the highest order effect in the equation; that is, there are no three-way or higher order interactions. The interaction, carried by the XZ term, reflects the shape of the regression surface, specifically how this shape differs from the flat regression plane associated with regression equations having only first-order terms. This shape does not change when variables are centered. In general, *centering predictors has no effect on the value of the regression coefficient for the highest order term* in the regression equation. For Eq. (7.1.2) we have

(7.2.4) $$B_{3,\text{centered}} = B_{3,\text{uncentered}}.$$

7.2.7 Do Not Center *Y*

In computing the centered regression equations and in displaying the regression surfaces in Figs. 7.1.1 and 7.2.1, Y has been left in uncentered form. There is no need to center Y because when it is in its original scale, predicted scores will also be in the units of the original scale and will have the same arithmetic mean as the observed criterion scores.

7.2.8 A Recommendation for Centering

We recommend that continuous predictors be centered before being entered into regression analyses containing interactions. Doing so has no effect on the estimate of the highest order interaction in the regression equation. Doing so yields two straightforward, meaningful interpretations of each first-order regression coefficient of predictors entered into the regression equation: (1) effects of the individual predictors at the mean of the sample, and (2) average effects of each individual predictors across the range of the other variables. Doing so also eliminates nonessential multicollinearity between first-order predictors and predictors that carry their interaction with other predictors.[1]

There is one exception to this recommendation: If a predictor has a meaningful zero point, then one may wish to keep the predictor in uncentered form. Let us return to the example of language development. Suppose we keep the predictor of child's age (A). Our second predictor

[1]The issue of centering is not confined to continuous variables; it also comes into play in the coding of categorical variables that interact with other categorical variables or with continuous variables in MR analysis, a topic developed in Chapter 9.

is number of siblings (S). Following our previous argument, we center age. However, zero siblings is a meaningful number of siblings; we decide to retain number of siblings S in its uncentered form. We expect age and number of siblings to interact; we form the cross-product aS of centered a with uncentered S and estimate the following regression equation:

$$\hat{Y} = B_1 a + B_2 S + B_3 aS + B_0.$$

The interpretation of the two first-order effects differs. The effect of number of siblings is at $a = 0$; since a is centered, B_2 is the regression of language development on number of siblings at the mean age of children in the sample. The effect of child's age is at $S = 0$, where $S = 0$ stands for zero siblings. Hence B_1 is the regression of language development on age *for children with no siblings.* If this is a meaningful coefficient from the perspective of data summarization or theory testing, then centering is not advised. But even if the variable has a meaningful zero point, it may be centered for interpretational reasons. If number of siblings had been centered, then B_1 would be interpreted as the regression of language development on age at mean number of siblings. Finally, B_3 is not affected by predictor scaling and provides an estimate of the interaction between the predictors regardless of predictor scaling.

Our discussion of centering predictors has been confined to those predictors that are included in the interaction. But it is entirely possible that we include a predictor that is not part of any interaction in a regression equation that contains interactions among other variables. Suppose in the example of language development, we wish to control for mother's education level (E) while studying the interaction between child's age and number of siblings in predicting child's language development. Assume we wish to center number of siblings for interpretational reasons. We estimate the following regression equation:

$$\hat{Y} = B_1 a + B_2 s + B_3 as + B_4 E + B_0.$$

It is not necessary to center E. The B_1, B_2, and B_3 coefficients will not be affected by the scaling of E because E does not interact with any other predictors in the equation. In addition, since E does not interact with the other predictors, the B_4 coefficient will be completely unaffected by changes in scaling of age and number of siblings. In fact, the only effect of centering E is on the intercept B_0. However, we recommend that for simplicity, if one is centering the variables entering the interaction, one should also center the remaining variables in the equation.

To reiterate our position on centering, *we strongly recommend the centering of all predictors* that enter into higher order interactions in MR prior to analysis. The cross-product terms that carry the interactions should be formed from the centered predictors (i.e., center each predictor first and then form the cross-products). Centering all predictors has interpretational advantages and eliminates confusing nonessential multicollinearity.

There is only one exception to this recommendation to center. If a predictor has a meaningful zero point, then one may wish to have regression coefficients in the overall regression equation refer to the regression of the criterion on predictors at this zero point. For the remainder of this chapter, we will assume that all predictors in regression equations containing an interaction have been centered, unless otherwise specified.

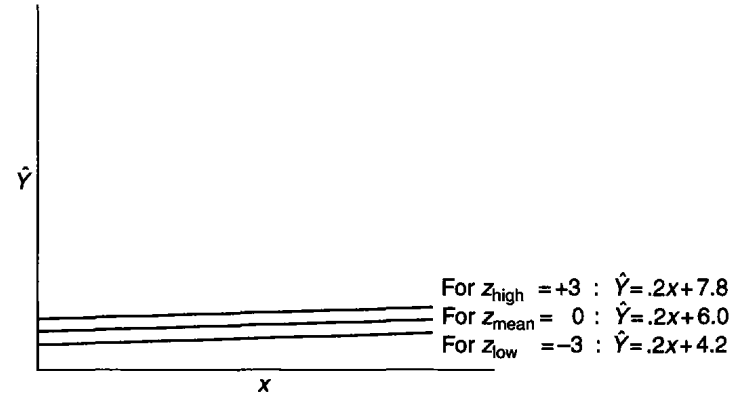
7.3 SIMPLE REGRESSION EQUATIONS AND SIMPLE SLOPES

If an interaction is found to exist in a regression equation, the issue becomes one of interpretation of the interaction. The approach we take harkens back to the idea of conditional effects

in MR with interactions: When X and Z interact, the regression of each predictor depends on the value of the other predictor. To characterize interactions, we examine the regression of the criterion Y on one predictor X at each of several values of the other predictor Z, as when we examine the regression of Y on x at z_{low}, z_{mean}, and z_{high} in Fig. 7.2.1(B). Following Aiken and West (1991), we call the regression line of Y on X at one value of Z a *simple regression line.* Hence, Figs. 7.2.1(A) and 7.2.1(B) each contain three simple regression lines.

In Fig. 7.3.1, we plot the centered simple regression lines of Fig. 7.2.1 in more familiar two-dimension representations. In Fig. 7.3.1(A), the regression lines of Y on x at z_{low}, z_{mean}, and z_{high} are reproduced from Fig. 7.2.1(A). Similarly, the three regression lines of Y on x in Fig. 7.3.1(B) are those from Fig. 7.2.1(B). Each line in Figs. 7.3.1(A) and 7.3.1(B) is the regression of Y on x at one value of the other predictor z, *a simple regression line.* The rule for discerning the presence of an interaction is straightforward. If the lines are parallel, there is no interaction, since the regression of Y on X is constant across all values of Z. If the lines are not parallel, there is an interaction, since the regression of Y on X is changing as a function of Z.

(A) Simple regression lines and equations based on Eq. (7.1.1), no interaction. Simple regression lines correspond to those in Fig. 7.2.1(A).



(B) Simple regression lines and equations based on Eq. (7.1.2), with interaction. Simple regression lines correspond to those in Fig. 7.2.1(B).

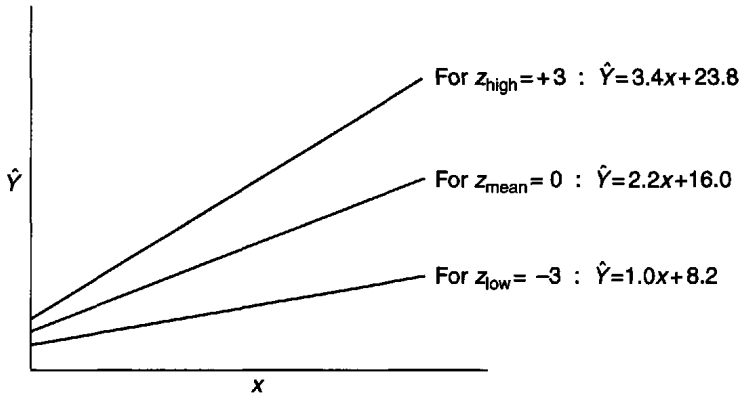


FIGURE 7.3.1 Simple regression lines and equations for Y on centered x at three values of centered z. The simple regression lines correspond directly to the simple regression lines in Fig. 7.2.1.

7.3.1 Plotting Interactions

Plotting interactions is the first step to their interpretation. We recommend plotting the regression of Y on X at three values of Z: the mean of Z plus a low and a high value of Z. Often a convenient set of values to choose are the mean of Z (Z_{mean}), one standard deviation below the mean of Z (Z_{low}), and one standard deviation above the mean of Z (Z_{high}). However, there may be specific meaningful values of Z—for example, clinical cutoffs for diagnostic levels of illness, or the income in dollars defined as the poverty level for a family of four. The symmetry of interactions means that the choice of plotting Y on X at values of Z as compared to Y on Z at values of X will depend on the theoretically more meaningful characterization of the data.

7.3.2 Moderator Variables

Psychological theories often hypothesize that a relationship between two variables will depend on a third variable. The third variable is referred to as a *moderator* (Baron & Kenny, 1986). These third variables may be organismic (e.g., gender, ethnicity, personality traits, abilities) or situational (e.g., controllable versus uncontrollable stressful events). They may be merely observed or manipulated. Of course, they are characterized statistically in terms of interactions. If a theory predicts that a variable M will moderate the relationship of another variable X to the criterion, then it is appropriate to plot regression of Y on X at meaningful values of the *moderator* M.

7.3.3 Simple Regression Equations

We can write a *simple regression equation* for each of the simple regression lines of Figs. 7.3.1(A) and 7.3.1(B). The use of simple regression equations is the key to the interpretation of interactions in MR analysis. *A simple regression equation* is the equation for the regression of the criterion on one predictor at a specific value of the other predictor(s), here Y on x at specific values of z.

For Figs. 7.2.1(A) and 7.3.1(A) with centered x and z, we place brackets in the regression equation with no interaction, $\hat{Y} = .2x + .6z + 6$, to show the regression of Y on x at values of z, in the form of a *simple regression equation*:

$$\hat{Y} = .2x + [.6z + 6].$$

Here the intercept of the simple regression equation $[.6z + 6]$ depends on the value of z; the slope of .2 does not. For each of the three values of z, we generate a simple regression equation. Recall that centered z takes on the values $(-5, -3, -1, 1, 3, 5)$, with $z_{mean} = 0$. We choose $z_{low} = -3$, and $z_{high} = 3$.

$$\begin{aligned}
&\text{For } z_{low} = -3: && \hat{Y} = .2x + [.6(-3) + 6] = .2x + 4.2;\\
&\text{For } z_{mean} = 0: && \hat{Y} = .2x + [.6(0) + 6] = .2x + 6.0;\\
&\text{For } z_{high} = 3: && \hat{Y} = .2x + [.6(3) + 6] = .2x + 7.8.
\end{aligned}$$

We note that in all three equations the regression coefficient for x has the constant value .2. The intercept increases from 4.2 to 6.0 to 7.8, as z increases from -3 to 0 to 3.

To plot a simple regression line, we follow standard practice for plotting lines: we substitute into the equation two values of x, and find $\hat{Y}$ corresponding to those two values, giving us two points for plotting. For example, for z_{high}, where $\hat{Y} = .2x + 7.8$, if $x = -3$, then $\hat{Y} = 7.2$; if $x = 3, \hat{Y} = 8.4$. To plot the simple regression line for z_{high} in Fig. 7.3.1(A), we used the points $(-3, 7.2)$ and $(3, 8.4)$.

The numerical result corresponds completely with the graphical results in Figs. 7.2.1(A) and 7.3.1(A). The *simple slopes* of the simple regression lines (i.e., the regression coefficients for Y on x in the simple regression equations) are constant at .2. The *simple intercepts*, that is, the regression constants in the simple regression equations (values of Y at $x = 0$ for specific values of z), increase with increasing values of z.

For Figs. 7.2.1(B) and 7.3.1(B), we first rearrange the regression equation containing the interaction, $\hat{Y} = 2.2x + 2.6z + .4xz + 16$, placing the terms involving x at the beginning of the equation:

$$\hat{Y} = 2.2x + .4xz + 2.6z + 16$$

We then factor out x and include some brackets to show the regression of Y on x at z in the form of a simple regression equation:

$$\hat{Y} = [2.2 + .4z]x + [2.6z + 16]$$

The expression $[2.2 + .4z]$ is the simple slope of the regression of Y on x at a particular value of z; $[2.6z + 16]$ is the simple intercept. In an equation with an xz interaction, both the simple slope and simple intercept for the regression of Y on x depend on the value of z.

For each of the three values of z, we generate a simple regression equation:

$$\begin{aligned}
&\text{For } z_{\text{low}} = -3: && \hat{Y} = [2.2 + .4(-3)]x + [2.6(-3) + 16] = 1.0x + 8.2;\\
&\text{For } z_{\text{mean}} = 0: && \hat{Y} = [2.2 + .4(0)]x + [2.6(0) + 16] = 2.2x + 16.0;\\
&\text{For } z_{\text{high}} = -3: && \hat{Y} = [2.2 + .4(3)]x + [2.6(3) + 16] = 3.4x + 23.8.
\end{aligned}$$

The numerical result is the same as the graphical results in Fig. 7.2.1(B) and 7.3.1(B): The simple slopes of the simple regression lines increase from 1.0 to 2.2 to 3.4 as z increases; the simple intercepts (values of Y at $x = 0$), increase from 8.2 to 16.0 to 23.8 as z increases. To plot a simple regression we follow the same approach as described earlier, that is, to substitute two values of x and solve for $\hat{Y}$. For z_{low}, where $\hat{Y} = 1.0x + 8.2$, if $x = -3$, then $\hat{Y} = 5.2$; if $x = 3$, then $\hat{Y} = 11.2$. To plot the regression line for z_{low} in Fig 7.3.1(B), we used the points $(-3, 5.2)$ and $(3, 11.2)$.

7.3.4 Overall Regression Coefficient and Simple Slope at the Mean

The overall regression coefficient B_1 for the regression of Y on x in the centered regression equation containing the interaction is 2.2 and represents the regression of Y on x at $z = 0$. The simple regression coefficient for Y on centered x at $z_{\text{mean}} = 0$ is also 2.2. This equality of coefficients is expected, since both coefficients represent the regression of Y on x at $z = 0$. In general, the simple regression coefficient for the regression of Y on x at the mean of z will equal the overall regression coefficient of Y on x in the centered regression equation.

We may cast simple regression equations in a general form. First, we have the overall regression equation containing predictors X and Z and their interaction:

$$\hat{Y} = B_1X + B_2Z + B_3XZ + B_0, \tag{7.1.2}$$

where B_3 is the regression coefficient for the interaction. We rearrange Eq. (7.1.2) to show the regression of Y on X at values of Z:

$$\begin{aligned}
\hat{Y} &= [B_1X + B_3XZ] + [B_2Z + B_0]\\
\hat{Y} &= [B_1 + B_3Z]X + [B_2Z + B_0],
\end{aligned} \tag{7.3.1}$$