6

Quantitative Scales, Curvilinear Relationships, and Transformations

6.1 INTRODUCTION

In Chapter 6 we continue our treatment of quantitative scales; that is, scales that can take on a continuous range of values. We include ordinal, interval, and ratio scales (Section 1.3.3), and note that they share the minimum property that the scale numbers assigned to the objects order the objects with regard to the measured attribute. Ordinal (rank-order) scales have only this property, interval scales add the property of equal units (intervals), and ratio scales have both equal intervals and equal ratios, hence a true scale zero. Despite these differences in the amount of information they yield, we find it convenient to treat them together. To this point we have treated quantitative variables as if they have only a *linear* relationship to the criterion. We would argue that in some research problems there may be other *nonlinear* aspects that should be used in addition to, or in place of, this linear aspect. This chapter thus addresses itself to the question of how nonlinear relationships can be detected, represented, and studied within the confines of *linear* MRC, by the use of multiple and/or nonlinear aspects of research variables.

6.1.1 What Do We Mean by Linear Regression?

Linear regression refers to regression models that take the form we have used throughout the book:

$$\hat{Y} = B_1X_1 + B_2X_2 + \cdots + B_kX_k + B_0. \tag{6.1.1}$$

More formally, regression equations of the form of Eq. (6.1.1) are said to be *linear in the parameters* (or linear in the coefficients), where the parameters refer to the coefficients $B_0, B_1, \ldots, B_k$. If a regression equation is linear in the parameters, then the predicted score is a *linear combination* of the predictors: Each predictor is simply multiplied by the regression coefficient and the products are added to produce the predicted score. Any relationship between a set of predictors and a criterion that can be expressed in the form of Eq. (6.1.1) can be handled in linear MR. The actual relationship between a predictor and the criterion need not be a linear (straight line) relationship in order to use linear MR to analyze the relationship.

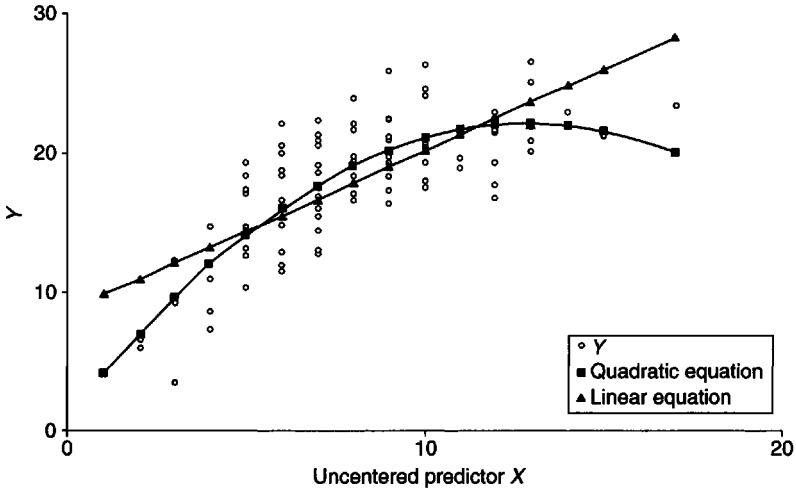


FIGURE 6.1.1 Quadratic relationship between X (number of credits taken in minor subject) and Y (interest in further course work in the minor). The triangles characterize the best fit linear regression of Y on X; the squares, the best fit quadratic regression of Y on X.

Theories in the social sciences sometimes predict a curvilinear relationship between two variables, such as that illustrated in Fig. 6.1.1. In Fig. 6.1.1, the outcome Y increases as X increases up to some maximum value; thereafter, as X increases, Y remains the same or even declines. For example, Janis (1967) hypothesized that one could motivate people to engage in health protective behaviors (e.g., to quit smoking, exercise) through fear communication (i.e., through strong threats about disease) but only up to a point. If the fear induced became too great, then people would deny the existence of the threat and health protective behavior would decline, rather than continue to rise with increasing fear. We will be able to capture this curvilinear relationship between X and Y with *polynomial regression*, a special case of linear MR, which we develop in Sections 6.2 and 6.3. The point here is that any regression model of the form of Eq. (6.1.1) can be analyzed in linear MR, even if the relationship of the predictor to the criterion is not linear.

6.1.2 Linearity in the Variables and Linear Multiple Regression

It is worth reviewing what we mean by a *linear relationship,* that is, a relationship of X to Y best summarized by a straight line (as apart from linear regression). The conditional means of Y, that is, the means of Y at each value of X, $\mu_{Y|X}$, lie on a straight line (see Section 4.3.1). More formally, *linearity in the variables* of a relationship means that the regression of Y on X is constant across all values of X; that is, a one-unit increase in X is associated with a constant magnitude of increase in Y (namely, B_{YX}), regardless of where in the X scale it occurs. If we use linear MR to characterize a predictor-criterion relationship, we are forcing this constant regression of Y on X across the range of X.

Of course, linearity is a special case of relationship; there are a variety of nonlinear relationships that can occur between predictors and a dependent variable. For example, we say that the cost of college tuition has "risen exponentially over the years"; what we mean is that the cost of college tuition has been rising at an ever increasing rate over the years. Obviously, the relationship of time (X) to tuition (Y) fails to meet the definition of linearity in the variables, which would require a constant increase in tuition with each year.

In Fig. 6.1.1 linearity in the variables does not hold, since the slope of the regression of Y on X changes over the range of X. If we predict Y from X in Fig. 6.1.1 with linear regression equation $\hat{Y} = B_1X + B_0$, the best fit straight line, shown in Fig. 6.1.1 with the triangles, does not capture the curvilinearity. The linear regression suggests that Y continues to increase as X increases; the leveling off and then declining of Y at high levels of X is not captured. To anticipate our discussion of polynomial regression, the curvilinear relationship of variable X to Y in Fig. 6.1.1 can be handled within linear MR through a regression equation in which a single research variable X is entered both as a linear predictor X, and as a curvilinear (second-order) predictor such as X^2. The second-order predictor is simply X^2, and together the two predictors X and X^2 represent the relationship of the variable to the criterion:

$$\hat{Y} = B_{1.2}X + B_{2.1}X^2 + B_0. \tag{6.1.2}$$

This equation is *linear in the parameters*; by this we mean that predictors are simply multiplied by the regression coefficients and the products summed to form the predicted score, rather than being in some more complex form, for example, $X^{1/B}$. Note that Eq. (6.1.2) is linear in the parameters, even though it is not linear in the variables, since X^2 is a predictor. So long as an equation is linear in the parameters, it can be analyzed with MR. The characterization of the relationship of the variables to the criterion in Eq. (6.1.2) is novel in that it requires two distinct predictors to capture the relationship: X representing a linear aspect and X^2 or some other transform, representing a curvilinear aspect.

6.1.3 Four Approaches to Examining Nonlinear Relationships in Multiple Regression

There are four broad classes of approaches to examining nonlinear relationships in MR. Of traditional and common use in the behavioral sciences is *polynomial regression*, explored here in depth. Power polynomials are a convenient method of fitting curves of almost any shape, although other functions such as a log function will often work as well. Second is the use of *monotonic nonlinear transformations* (i.e., transformations that shrink or stretch portions of the scale differentially). That is, these transformations change the relative spacing between adjacent points on the scale (i.e., the nonlinearity) but maintain the rank order of the scores (i.e., the monotonicity). We choose the particular transformation in order to create rescaled variable(s) after transformation that bear a close to linear relationship to one another so that they may be treated in linear MR. The choice of transformation may be either theory driven or empirically driven by the data. Transformations are treated in Section 6.4. Third is *nonlinear regression*, a distinctly different class of analysis in which the central point of the analysis is estimation of complex (nonlinear) relationships among variables that may be implied by theory. We introduce nonlinear regression here in Section 6.5 and devote much of Chapter 13 to two common forms of nonlinear regression: logistic regression and Poisson regression. Fourth are *nonparametric regression approaches* (Hastie & Tibshirani, 1990), introduced briefly in Section 6.6.

Our presentation is a mix of approaches already familiar in the behavioral sciences and approaches that have not heretofore been much used in the behavioral sciences, though they are standardly used in related fields. We hope that the presentation of the unfamiliar approaches may lead to innovation in the analysis of behavioral science data.

For our treatment of curvilinear relationships we must distinguish between a variable $\boldsymbol{X}$, and the predictors that carry the various aspects of the relationship of that variable to the criterion (e.g., X and X^2 of Eq. 6.1.2). The variable will be represented in bold italics (here $\boldsymbol{X}$), and its aspects will be characterized by the same letter, in regular type, with the particular function, $X, X^2, \log X$ specifically indicated. Capital letters will represent raw scores X, squares

of raw scores X^2, etc. Lowercase letters will represent deviation (centered) scores of the form $x = (X - M_X)$; $x^2 = (X - M_X)^2$, as in earlier chapters.

Boxed Material in the Text

In this chapter we adopt a strategy initiated in Chapter 4 of putting some material into boxes, typically material of interest to the more mathematically inclined reader. The boxes provide supplementation to the text; the text can be read without the boxes. Boxes appear within the section in which the boxed material is relevant. Readers not interested in boxed material should simply skip to the beginning of the next numbered section.

6.2 POWER POLYNOMIALS

6.2.1 Method

It is a most useful mathematical fact that in a graph of n data points relating Y to X (where the values of X are all different), an equation of the following form will define a function that fits these points *exactly*:

$$\hat{Y} = BX + CX^2 + DX^3 + \cdots + QX^{n-1} + A. \tag{6.2.1}$$

This *polynomial equation* relates the one variable X to Y by using $(n - 1)$ aspects of X to the criterion Y. Each term X, X^2, etc., is said to have *order* equal to its exponent,[1] (e.g., X^2 is of order 2). The order of the polynomial equation is the order of the highest term, here $(n - 1)$. The term with the highest exponent is referred to as the *highest order term* (here X^{n-1}) and all other terms are referred to as *lower order terms*. The relationship of variable X to Y is nonlinear, and several powers of the linear X term serve as predictors in addition to the linear term. Put another way, the regression equation includes stand-in variables (X^2, X^3, etc.) that possess a known nonlinear relationship to the original variables. Yet the regression equation is linear in the parameters and can be analyzed with MR. By structuring nonlinear relationships in this way, we make it possible to determine *whether* and specifically *how* a relationship is nonlinear and to write an equation that *describes* this relationship. The higher order terms X^2, X^3, etc. in the polynomial equation are nonlinear transformations of the original X; thus polynomial regression falls within the general strategy of creating nonlinear transformations of variables that can be handled in linear MR.

The linear, quadratic, and cubic polynomials follow. The highest order term in a polynomial equation determines the *overall shape* of the regression function within the range between $-\infty$ and $+\infty$, that is, the number of bends in the function. The $B_{2.1}X^2$ term in the quadratic Eq. (6.2.3) causes the regression line to be a parabola (one bend), as in Fig. 6.1.1. The $B_{3.12}X^3$ term in the cubic Eq. (6.2.4) produces an S-shaped function (two bends), as in Fig. 6.2.1. There are $(q - 1)$ bends in a polynomial of order q.

$$\text{Linear:} \quad \hat{Y} = B_1X + B_0 \qquad \text{with no bends;} \tag{6.2.2}$$

$$\text{Quadratic:} \quad \hat{Y} = B_{1.2}X + B_{2.1}X^2 + B_0 \qquad \text{with one bend;} \tag{6.2.3}$$

$$\text{Cubic:} \quad \hat{Y} = B_{1.23}X + B_{2.13}X^2 + B_{3.12}X^3 + B_0 \qquad \text{with two bends.} \tag{6.2.4}$$

In reality, a bend may occur outside the range of the observed values of the predictor.

---

[1] In more complex terms considered in Chapter 7, the order is equal to the sum of the exponents (e.g., X^2Z is of order 3).

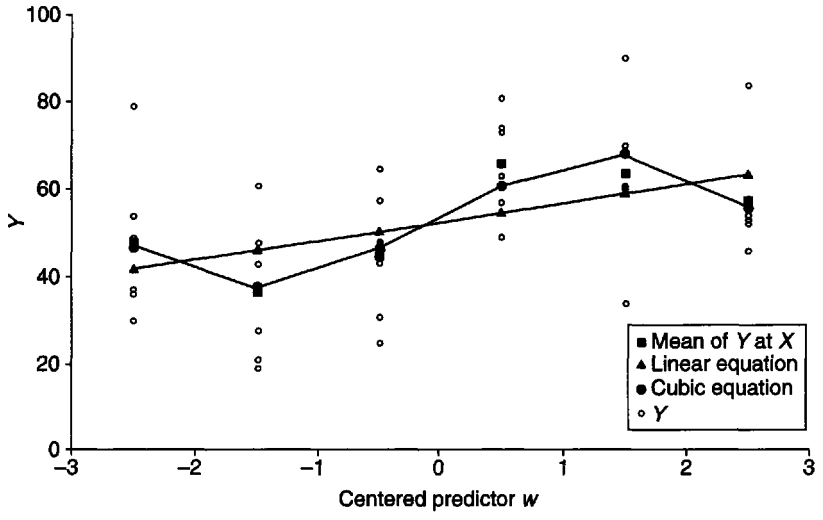


FIGURE 6.2.1 Cubic relationship between w and Y. Predictor W is in centered (deviation) form, $w = (W - M_w)$. The triangles characterize the best fit linear regression of Y on w; the solid circles, the best fit cubic regression of Y on w. The conditional dependent variable means at each value of w are represented by squares. Note that the cubic equation does not track the conditional means with complete accuracy.

The signs (positive, negative) of the highest order terms in polynomial regressions determine the direction of the curvature. In the quadratic equation, positive B_2 indicates a curve that is U-shaped (concave upward, as a smile); negative B_2 indicates a curve that is inverted U-shaped (concave downward, as a frown), as in Fig. 6.1.1. In the cubic equation, negative B_3 indicates a curve that is first concave upward and then concave downward as X increases, as in Fig. 6.2.1; positive B_3 yields the opposite pattern, concave downward followed by concave upward.

What Order Polynomial to Estimate?

From a mathematical perspective, we may fit up to a polynomial of order $(q - 1)$ for a variable whose scale contains q distinct values (e.g., a 19th-order polynomial for a 20-point scale). From a research perspective, we have neither theoretical rationale nor practical benefit from fitting a polynomial of high order that captures every random jiggle in the data. Our nonlinear curve fitting with power polynomials should make substantive sense. We argue that theory should guide the choice and that for the most part theory in the social sciences predicts quadratic, and at most cubic, relationships. For example, opponent process theories predict phenomena that are properly cubic (i.e., a reaction in one direction followed by a compensatory over-response in the opposite direction, followed by a return to baseline, as in physiological responses to stressors). The quality of data must also be considered: social science data typically do not support explorations above the cubic level. Finally, coefficients of higher order polynomials (above cubic) are difficult to interpret (Neter, Kutner, Nachtsheim, & Wasserman, 1996). In this presentation we focus on quadratic and cubic equations.

We realize that theorizing may be weak and that there may be an exploratory character to the analysis. In the behavioral sciences, the prediction is often that the relationship is "nonlinear" absent the specific form of the nonlinearity. One may suspect that behavior will approach a peak (asymptote) and then remain essentially level, or will decline and then level off at some minimal

level. A quadratic equation can be used to represent either of these relationships. We caution that *polynomial equations may be only approximations* to nonlinear relationships. Finally, we must distinguish nonlinear relationships from relationships that exhibit cyclic variation over time (e.g., activity levels of people over 24-hour cycles, or mood levels over 7 days of the week); here time series analysis is appropriate (see Section 15.8).

Detecting Curvilinear Relationships Through Graphical Displays

In the absence of specific predictions concerning nonlinearity, the detection of nonlinearities begins with *scatterplots* of predictors against the criterion (see Section 4.2.2). These plots should be augmented with superimposed curves for better visualization of the data, for example, a curve connecting the means of the criterion Y at specific values of X. A *lowess* (or, equivalently, *loess*) curve, a nonparametric curve that follows the data and traces the X–Y relationship (Section 4.2.2), will help to visualize nonlinearity; in Fig. 6.2.2(A), the data of Fig. 6.1.1 are plotted with a lowess curve. *Residual scatterplots* (Section 4.4.1) are even more useful. The residual scatterplot in Fig. 6.2.2(B) is generated from the data in Fig. 6.1.1 by predicting the criterion from only the linear aspect of X: $\hat{Y} = B_1X + B_0$, and plotting the residuals from this analysis against the predictor. Once the strong linear increasing trend in the data has been removed, the curvilinearity is clearly apparent. The residuals are systematically related to the value of X: below zero for low and high values of X, and above zero for moderate values of X. The detection of curvilinearity in a residual scatterplot is also enhanced with a lowess line. Should nonlinearity be suspected, a polynomial regression equation is specified. Higher order predictor(s) are created simply by squaring (and perhaps cubing) the original linear variable X. Multiple regression is applied to the polynomial equation.

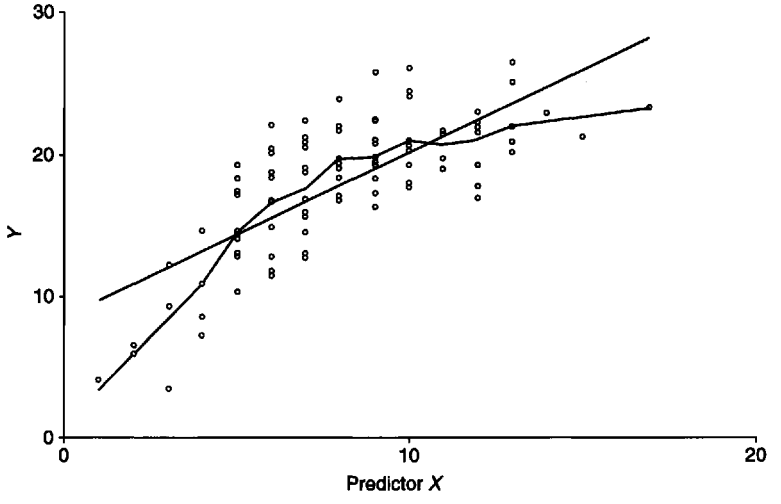
6.2.2 An Example: Quadratic Fit

Consider again the gestalt of the 100 points in Fig. 6.1.1, which suggests that Y reaches a maximum and then declines slightly as variable X increases. As an example, we might think of variable X as the number of elective credits undergraduate students have taken in a particular minor subject and Y as their expressed interest in taking further electives in this minor. We hypothesize that interest in course work in a minor will increase as students take more courses in the minor, but only up to a point. After several courses, students then will shift their interest to electives in other topics and interest in the particular minor will begin to decline.

The linear correlation between variable X and Y is substantial, $r = .75$, and reflects the strong positive relationship between variable X and Y. However, this linear correlation does not capture the onset of decline in interest above about 12 credits (four 3-credit courses). In Table 6.2.1 we summarize a series of polynomial regressions fitted to these data, the linear, quadratic, and cubic polynomials of Eqs. (6.2.2), (6.2.3), and (6.2.4), respectively. We present the regression equations, tests of significance of each of the individual regression coefficients, and 95% confidence intervals on the regression coefficients. The triangles in Fig. 6.1.1 are predicted scores from the linear regression equation $\hat{Y} = 1.15X_1 + 8.72$. The linear regression coefficient B_1 in the linear equation is significant, due to the strong increasing trend in the data, $B_1 = 1.15$, confidence interval (*CI*): [.95, 1.35], $t(98) = 11.19, p < .01$. This coefficient only signifies the generally increasing trend in the data and does not test our curvilinear prediction.

In Fig. 6.1.1 the squares are predicted scores from the quadratic regression equation $\hat{Y}_2 = 3.27X - .13X^2 + 1.00$. The $B_{2.1}$ coefficient is negative, $B_{2.1} = -.13$, the 95% confidence interval does not include zero, *CI*: [−.17, −.08], and $t(97) = -5.75, p < .01$. Negative $B_{2.1}$ reflects the hypothesized initial rise followed by decline in interest in taking further courses as number of credits in the minor accumulates.

(A) Scatterplot of data from Fig. 6.1.1 with best fit linear regression of Y on X and lowess curve superimposed. The lowess curve is useful in visual detection of the curvilinear relationship of X to Y.



(B) Residual scatterplot of data from Fig. 6.1.1 following prediction from linear equation $\hat{Y} = B_1X + B_0$. The increasing linear trend shown in Fig. 6.2.2(A) has been removed and the residuals exhibit the remaining curvilinearity. A lowess curve is superimposed to highlight the remaining curvilinearity.

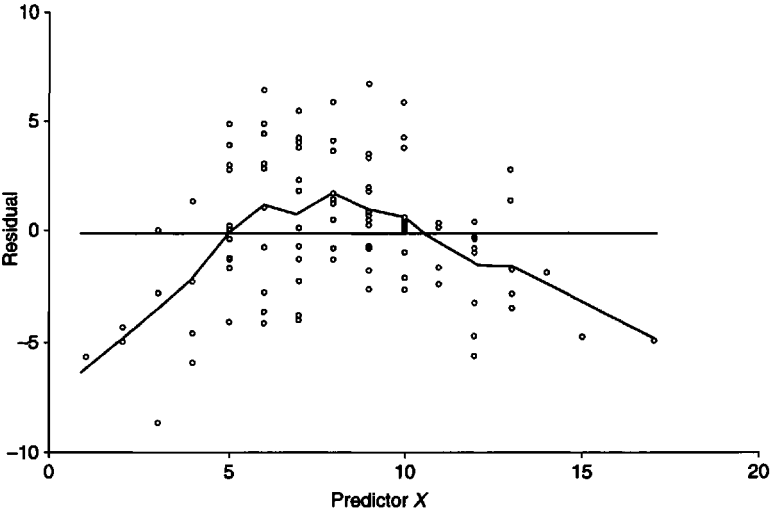


FIGURE 6.2.2 Exploration of data to detect the presence of curvilinearity. The data are the same as in Fig. 6.1.1.

We might wonder whether the students' interest once again rises among those students who take many credits in the minor (say after 15 credits or five 3-credit courses); hence we examine the cubic equation. In the cubic equation, the confidence interval for the $B_{3.12}$ coefficient includes zero, *CI*: [−.0007, .02]. There is no evidence of a cubic relationship in the data, that is, after decline in student interest above about 12 credits, interest does not increase again with higher credits.

TABLE 6.2.1
Polynomial Multiple Regression Analysis of Regression of Y on X with Uncentered Data

	Mean	*sd*	Uncentered correlation matrix			
			Y	X	X^2	X^3
Y	18.11	4.75	Y 1.000	.749	.650	.554
X	8.17	3.10	X .749	1.000	.972	.911
X^2	76.23	53.45	X^2 .650	.972	1.000	.981
X^3	781.99	806.44	X^3 .554	.911	.981	1.000

Regression Equations

Linear:	$\hat{Y}_{linear} =$	$1.15X$	$+8.72$		
95% *CI*:		[.95, 1.35]			
t_{B_i} (98):		11.19			
Quadratic:	$\hat{Y}_{quadratic} =$	$3.27X$	$-.13X^2$	$+1.00$	
95% *CI*:		[2.52, 4.03]	[−.17, −.08]		
t_{B_i} (97):		8.62	−5.75		
Cubic:	$\hat{Y}_{cubic} =$	$4.94X$	$-.34X^2$	$+.01X^3$	-2.58
95% *CI*:		[2.99, 6.89]	[−.58, −.10]	[−.0007, .02]	
t_{B_i} (96):		5.03	−2.86	1.83	

Hierarchical Model

Equation	IVs	R^2	F	*df*	I	F_I	*df*
Linear	X_1	.56	125.14**	1, 98	.56	125.14**	1, 98
Quadratic	X_1 X_2	.67	99.62**	2, 97	.11	33.11**	1, 97
Cubic	X_1 X_2 X_3	.68	69.15**	3, 96	.01	3.35	1, 96

**$p < .01$
Note: I is the increment in R^2.

Structuring Polynomial Equations: Include all Lower Order Terms

Each polynomial equation in Table 6.2.1 contains all lower order terms, (i.e., the quadratic equation contains the linear term; the cubic equation, the linear and quadratic terms). In order that higher order terms have meaning, all lower order terms must be included, since higher order terms are reflective of the specific level of curvature they represent only if all lower order terms are partialed out. For example, in Table 6.2.1, had we not included the linear term in the quadratic equation, the regression coefficient for X^2 would have confounded linear and quadratic variance.

Conditional Effects: Interpretation of Lower Order Coefficients in Higher Order Equations Based on Uncentered Variables

In equations containing linear predictors and powers of these predictors, the actual slope of the regression of Y on X differs for each value of X. In Fig. 6.1.1, one can imagine drawing a tangent line to each of the darkened squares representing the quadratic relationship. The tangent line summarizes the linear regression of Y on X *at that particular value of* X as characterized

in the quadratic equation. The actual slope of each tangent line[2] at a particular value of X for the quadratic equation (Eq. 6.2.3) is given by the expression $(B_{1.2} + 2B_{2.1}X)$. The value of this expression depends on X, i.e., is different for every value of X. This expression equals $B_{1.2}$ only at $X = 0$. Hence the $B_{1.2}$ coefficient in the quadratic equation represents the linear regression of Y on X at only the point $X = 0$, as characterized by the quadratic equation. The conditional nature of the $B_{1.2}$ coefficient makes sense when we consider that the slope of the linear regression of Y on X is different for every value of X if the relationship is curvilinear. In our example, the $B_{1.2}$ coefficient in the quadratic equation ($B_{1.2} = 3.27$) represents the linear regression of Y on X at $X = 0$. A glance at Fig. 6.1.1 tells us that this is not a meaningful value, since there are no data points in which $X = 0$; all students under consideration have taken at least one credit in the subject area; that is, observed scores on the predictor range from 1 through 17. To understand what the $B_{1.2}$ coefficient represents, imagine projecting the quadratic curve (the squares) downward to $X = 0$. The slope of the regression of Y on X at the point $X = 0$ would be steeply positive; the large positive $B_{1.2} = 3.27$ is this slope. That is, if we extrapolate our quadratic relationship to students who have never taken even a one-credit course in the subject in question, we would predict a rapid rise in their interest in the subject. We warn the reader not to report this $B_{1.2}$ coefficient or test of significance of this coefficient unless (a) zero is a meaningful value that occurs in the data set, (b) one wishes to consider the linear trend in the data only at the value zero, and (c) the meaning of the coefficient is carefully explained, since these coefficients are little understood and will be grossly misinterpreted.

6.2.3 Centering Predictors in Polynomial Equations

Lower order coefficients in higher order regression equations (regression equations containing terms of higher than order unity) only have meaningful interpretation if the variable with which we are working has a meaningful zero. There is a simple solution to making the value zero meaningful on any quantitative scale. We *center* the linear predictor, that is we convert X to deviation form:

$$\text{centered linear predictor } x: \qquad x = (X - M_X).$$

With centered variables, the mean M_X is, of course, zero. Thus the regression of Y on x at $x = 0$ becomes meaningful: it is the linear regression of Y on Z at the mean of the variable X. Once we have centered the linear predictor, we then form the higher order predictors from centered x:

$$\text{centered quadratic predictor } x^2: \qquad x^2 = (X - M_X)^2,$$

and

$$\text{centered cubic predictor } x^3: \qquad x^3 = (X - M_X)^3.$$

We use these predictors in our polynomial regression equations. For example, the cubic equation becomes

$$\begin{aligned}\hat{Y} &= B_{1.23}(X - M_X) + B_{2.13}(X - M_X)^2 + B_{3.12}(X - M_X)^3 + B_0 \\ &= B_{1.23}x \qquad\quad + B_{2.13}x^2 \qquad\quad + B_{3.12}x^3 \qquad\quad + B_0.\end{aligned}$$

To gain the benefits of interpretation of lower order terms, we do not need to center the criterion Y; we leave it in raw score form so that predicted scores will be in the metric of the observed criterion.

---

[2]The expression $(B_{1.2} + 2B_{2.1}X)$ is actually the first derivative of Eq. (6.2.3).

TABLE 6.2.2
Polynomial Multiple Regression Analysis of Regression of Y on X with Centered Data

	Mean	*sd*	Centered correlation matrix			
				Y	x	x^2	x^3
Y	18.11	4.75	Y	1.000	.749	−.250	.586
x	0.00	3.10	x	.749	1.000	.110	.787
x^2	9.48	12.62	x^2	−.250	.110	1.000	.279
x^3	4.27	102.70	x^3	.586	.787	.279	1.000

Regression Equations

Linear:	$\hat{Y}_{\text{linear}} =$	$1.15x$	$+18.10$		
95% *CI*:		[.95, 1.35]			
$t_{B_i}(98)$:		11.19			
Quadratic:	$\hat{Y}_{\text{quadratic}} =$	$1.21x$	$-13x^2$	$+19.30$	
95% *CI*:		[1.03, 1.38]	[−.17, −.08]		
$t_{B_i}(97)$:		13.45	−5.75		
Cubic:	$\hat{Y}_{\text{cubic}} =$	$1.00x$	$-.14x^2$	$+.01x^3$	$+19.39$
95% *CI*:		[.70, 1.28]	[−.19, −.09]	[−.0007, .02]	
$t_{B_i}(96)$:		6.86	−6.10	1.83	

Hierarchical Model

Equation	IVs	R^2	F	*df*	I	F_I	*df*
Linear	x_1	.56	125.14**	1, 98	.56	125.14**	1, 98
Quardratic	x_1, x_2	.67	99.62**	2, 97	.11	33.11**	1, 97
Cubic	x_1, x_2, x_3	.68	69.15**	3, 96	.01	3.35	1, 96

**$p < .01$
Note: I is the increment in R^2.


CH06EX01

Table 6.2.2 provides a reanalysis of the data of Fig. 6.1.1, following centering of X. To prepare the table, the original X variable with $M_X = 8.17$ was centered by subtracting the mean from each value of X, yielding predictor x with $M_X = 0$. Figure 6.2.3 shows the data set and the linear and quadratic regression functions based on centered x; $x = 0$ is approximately at the middle of the x axis. Note that Figs. 6.1.1 and 6.2.3 are identical in form. The shapes of the regression functions do not change on centering X; only the scaling of the x axis changes.

Essential Versus Nonessential Multicollinearity

There are substantial differences, however, between the results presented in Table 6.2.1 for uncentered X versus Table 6.2.2 for centered x. The means of the three predictors have changed, as have the standard deviations of all but the linear term. Dramatic changes have occurred in the correlation matrix. Consider first the correlation between X and X^2. In the uncentered case (Table 6.2.1), this correlation is .972. However, in the centered case (Table 6.2.2), the corresponding correlation is only .110. Whenever we take an uncentered predictor X (a predictor with a nonzero mean) and compute powers of the predictor (X^2, X^3, etc.), these powers will be highly correlated with the original linear predictor X. There are two sources of correlation between a predictor and an even power of the predictor, say between X and X^2 (Marquardt,

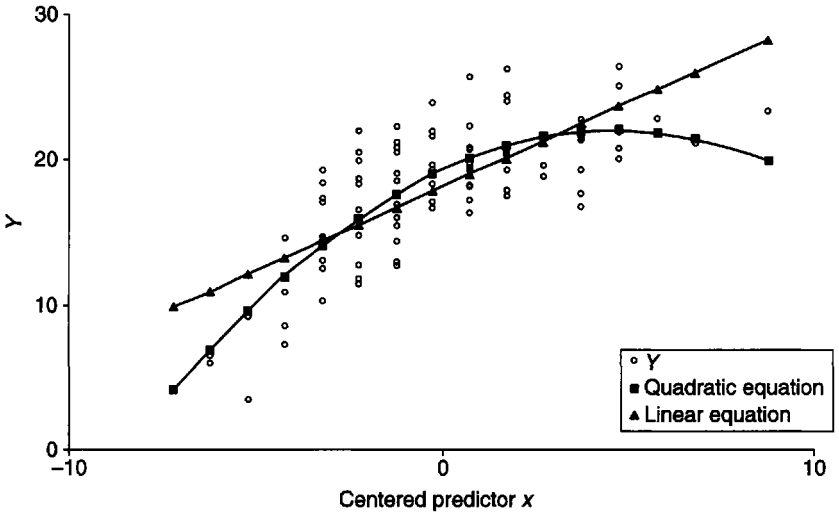


FIGURE 6.2.3 Data of Fig. 6.1.1 with predictor X in centered (deviation) form $x = (X - M_X)$. The triangles characterize the best fit linear regression of Y on x; the squares, the best fit quadratic regression of Y on X.

1980). First is *nonessential multicollinearity* that exists merely due to the scaling (nonzero mean) of X. If we set the mean of X to zero, then all nonessential multicollinearity is removed. This accounts for the drop in correlation from .972 to .110. Second is *essential multicollinearity*, correlation that exists because of any nonsymmetry in the distribution of the original X variable. For a perfectly symmetric predictor X, if we center it, producing x, and then square the centered x, producing x^2, x and x^2 will be completely uncorrelated. In fact, this will also be so for the correlation of x with all even powers of x (i.e., x^2, x^4, x^6, etc.). The correlation between x and x^2 in our example of .110 is attributable to the very slight asymmetry of X. Centering leads to computational advantages because very high correlations between predictors may cause computational difficulties. The correlation between X and X^3 is .911 in the uncentered case and .787 in the centered case. Correlations between a predictor and odd powers (i.e., x^3, x^5, x^7, etc.) of the predictors will not drop dramatically when we center.

Now consider the correlations of each predictor with the criterion Y. The linear transformation of X to x leaves the correlation with Y unchanged, $r = .749$. However, the correlation of the X^2 term with Y changes dramatically, from .650 in the uncentered case to $-.250$ in the centered case. We know from Figs. 6.1.1 and 6.2.3 that the shape of the relationships in the data has changed *not at all* on centering. Consistent with our consideration of MR with correlated predictors, the examination of hypotheses about the effects of individual predictors is considered in the context of all other predictors. We interpret partial effects of individual predictors with all other predictors held constant. The polynomial regression case is no exception. The zero-order correlations of the linear, quadratic, and cubic predictors with the criterion should not be interpreted in examining hypotheses about fit. Interpretation of the overall shape of the regression function should be based on the *partial regression coefficient for the highest order predictor in the context of the full polynomial equation*, that is, the regression equation containing all lower order terms. Again, if we omit lower order terms, then the variance attributed to higher order terms will be confounded with variance attributable to the omitted lower order terms. Finally, we would not look at effects in reverse order, for example, asking whether the linear term contributes over and above the quadratic term, even if the quadratic relationship were expected.

Regression Coefficients in Centered Versus Uncentered Polynomial Regressions

Although the regression functions in Figs. 6.1.1 and 6.2.3 are identical, there is a striking difference between the coefficients of the regression equations. First, the intercepts differ because they reflect the scaling of the means of the variables. For the uncentered predictor, zero represents no previous credit hours, whereas for the centered predictor, 0 represents the mean hours of credit ($M_X = 8.17$). In addition, the lower order terms differ. Consider the centered quadratic equation: $\hat{Y}_2 = 1.21X - .13X^2 + 19.30$. Examine Fig. 6.2.3. The $B_{1.2} = 1.21$ coefficient gives the linear regression of Y on x at the point $x = 0$, now the mean of x. That this coefficient is positive tells us that at the mean of x, the criterion Y is still increasing. In terms of our numerical example, at the mean number of credits taken in the minor (raw $M_X = 8.17$), interest in the minor is still increasing. Note that the centered $B_{1.2}$ of 1.21 is much smaller than the uncentered $B_{1.2}$ of 3.27. Recall that uncentered $B_{1.2}$ of 3.27 represented the slope of Y on X if the regression curve for uncentered X in Fig. 6.1.1 were extended downward to $X = 0$, a very steep portion of the curve.

The $B_{1.2}$ coefficient in the centered quadratic regression equation also has another useful interpretation; it is the average linear slope of the regression of Y on X *in the quadratic equation.* That $B_{1.2} = 1.21$ tells us that the overall linear trend in the data is positive. It is important to note that we can use both coefficients of the *centered* quadratic equation for interpretation. Both the linear and quadratic terms in the quadratic equation are meaningful once we have centered predictor X.

The $B_{1.2} = 1.21$ coefficient from the centered quadratic equation is not identical to that from the centered linear equation ($B_1 = 1.15$). In the quadratic equation, the $B_{1.2}$ coefficient is that for x with x^2 partialed out; in the linear equation, the B_1 coefficient has no terms partialed from it.

The $B_{2.1}$ coefficient ($B_{2.1} = -.13$) in the quadratic equation is constant over additive transformations (adding, subtracting a constant) and thus does not change from the uncentered to centered equation. In a polynomial regression equation of any order the regression coefficient for the highest order term is identical in the uncentered and centered solutions, but all lower order coefficients are different.

Predictors in Polynomial Regression Equations Should Be Centered

Centering renders all the regression coefficients in a polynomial regression equation meaningful as reflecting the regression function at the mean of the predictor. Centering also eliminates the extreme multicollinearity associated with using powers of predictors in a single equation. We therefore strongly recommend the use and reporting of centered polynomial equations. Although lower order terms have interpretations in uncentered equations, these interpretations often are not meaningful in terms of the range of the data and are guaranteed to be misleading to the naive consumer of polynomial regression equations. Only in cases in which X is measured on a ratio scale with a true 0 point is the use of polynomial equations with the raw data likely to lead to easily interpretable results.

6.2.4 Relationship of Test of Significance of Highest Order Coefficient and Gain in Prediction

The tests of the highest order coefficients in polynomial equations are actually tests of whether these aspects of the relationship of variable X to Y contribute to overall prediction above and beyond all lower order terms. The hierarchical tests reported in Tables 6.2.1 and 6.2.2 are the now familiar F tests of the contribution of a set B of predictors to an equation containing

set A (see Section 5.5). Consider the centered analysis in Table 6.2.2. The squared multiple correlation R^2_{linear} for the linear equation is .56, for the quadratic equation, $R^2_{\text{quadratic}} = .67$. This represents an increment of .11 (an 11% gain in total prediction) from the addition of the quadratic term to the linear term. This increment I is the squared semipartial correlation sr^2_2 of the quadratic term with the criterion, over and above the linear term (for x^2 with x partialed). In general the F test for this increment in any order polynomial is computed by treating the highest order term as set B and all other lower terms as Set A, as in Section 5.5, Eq. (5.5.2A):

$$\text{(5.5.2A)} \qquad F = \frac{R^2_{Y.AB} - R^2_{Y.A}}{1 - R^2_{Y.AB}} \times \frac{n - k_A - k_B - 1}{k_B} \qquad (df = k_B, n - k_A - k_B - 1).$$

For the quadratic term,

$$F = \frac{.67257 - .56082}{1 - .67257} \times \frac{97}{1} = 33.11 \qquad (df = 1, 97).$$

Actually, the F test for the gain in prediction by the addition of the quadratic term is the square of the t test for the significance of the quadratic term[3] in the quadratic equation, $t(97) = -5.75$.

Our emphasis on the increment in prediction by the quadratic term over and above the linear term should highlight once again that it is not the X^2 term per se that carries the quadratic component of the relationship of variable X to Y. Hence inspection of the correlation of the X^2 term with the criterion does not reflect the curvilinear nature of the data, as we warned in Section 6.2.3. Rather it is the *partialed* $X^2_{2.1}$, that is, X^2 with linear X partialed out, that represents the pure quadratic variable (Cohen, 1978).

6.2.5 Interpreting Polynomial Regression Results

Plots of Curvilinear Relationships

Plotting curvilinear relationships has now been made easy with the graphics contained in widely available statistical software, but plotting by hand can be accomplished by substituting values of X into the regression equation and generating the predicted scores. This must be done for a number of points to capture the shape of the curve. Be certain that if you are using the centered regression equation, you substitute centered values of X. Further, be certain that you use the highest order regression equation you have selected, *not* the linear term from the linear equation, the quadratic term from the quadratic equation, etc. The graph of the quadratic function in Fig. 6.2.3 is generated from centered equation $\hat{Y} = 1.21x - .13x^2 + 19.30$. For example, for centered $x = 0, \hat{Y} = 1.21(0) - .13(0^2) + 19.30 = 19.30$. For centered $x = 5, \hat{Y} = 1.20(5) - .13(5^2) = 19.30 = 22.05$. It is straightforward to translate the values of centered x into the original scale. Centered $x = 0$ corresponds to $M_X = 8.17$; centered $x = 5$ corresponds to $X = 5 + 8.17 = 13.17$.

Maxima and Minima

The polynomial of Eq. (6.2.1) defines a function that contains powers of X up to (let us say) the kth, with $k - 1$ bends. We may be interested in identifying the value of X at which each of the bends occurs,[4] for example, after what number of credits taken does interest in taking further courses begin to decline. Box 6.2.1 shows the computation of this value for the one bend in the quadratic polynomial equation.

---

[3]Note that this equation is equivalent to that of the test for sr_i (Eq. 3.6.8).

[4]In the quadratic equation, Eq. (6.2.3), the first derivative $B_{1.2} + 2B_{2.1}X$ is set to zero to solve for the value of X at which the function reaches a maximum or minimum.

BOX 6.2.1
Maximum and Minimum for a Quadratic Equation

For the quadratic, Eq. (6.2.3), there is one bend at the value X_M,

$$X_M = \frac{-B_{1.2}}{2B_{2.1}}. \tag{6.2.5}$$

In the centered quadratic equation of Table 6.2.2, $X_M = -1.21/2(-.13) = 4.65$. Recall that the mean of X was 8.17. The value 4.65 is the number of points above the mean after centering, which is equivalent to $8.17 + 4.65 = 12.82$ in raw score units. Students' interest in further course work in a minor subject is estimated to peak at just under 13 credits of course work in the minor and to decline thereafter. We know from inspection of Fig. 6.2.3 that the value 4.65 is the value of centered x at which $\hat{Y}$ is maximum. In a quadratic equation with a negative $B_{2.1}$ coefficient, the value of $\hat{Y}$ is a maximum because the overall shape of the curve is concave downward; with $B_{2.1}$ positive, the value of $\hat{Y}$ is a minimum, since the overall shape of the curve is concave upward. Note that although X_M falls within the range of the observed data in this example, it need not. It may be that we are fitting a quadratic equation to data that rise to some point and then level off (reach asymptote). The maximum of a quadratic equation may fall outside the meaningful range of the data (e.g., be higher than the highest value of predictor X) and will be expected to do so if the data being fitted are asymptotic in the observed range.

The value of $\hat{Y}$ at its maximum or minimum value (here at centered $x = 4.65$) can be found by substituting Eq. (6.2.5) into the quadratic regression equation, which yields

$$\hat{Y}_M = \frac{4\,(B_{2.1})(B_0) - B_{1.2}^2}{4\,(B_{2.1})} \tag{6.2.6}$$

For the quadratic numerical example in Table 6.2.2, this value is 22.12 on the 30-point scale of interest. Note that because we did not center Y, all predicted values for Y, including the predicted maximum, are in the original scale of the criterion.

Maxima and minima of polynomial regression equations are of interest in some applications, but they need not be routinely determined. They also require some caution in interpretation, since they are subject to sampling error, like other statistics, and hence are only approximations of the corresponding population values. The values of X_M identified by Eq. (6.2.5) are themselves sample estimates and exhibit sampling variability. For large samples, approximate standard errors of these sample X_M values can be estimated (see Neter, Wasserman, & Kutner, 1989, p. 337, Eq. 9.22).

Simple Slopes: Regression of Y on X at Specific Values of X

Once again refer to Fig. 6.2.3, and focus on the quadratic equation, identified with the squares. Imagine that we place a tangent line to the curve at each square, which represents the linear regression of Y on X at the particular value of X represented by the square. Recall from Section 6.2.2 that we can actually calculate the value of the linear regression of Y on X for each value of X using the expression

$$B_{1.2} + 2B_{2.1}X. \tag{6.2.7}$$

These values are referred to as *simple slopes* (Aiken & West, 1991). In polynomial equations, the simple slopes represent the linear regression of Y on X at a particular value of X; each is the

slope of a tangent line to the polynomial curve at a particular value of X. These values are useful in describing the polynomial regression function. For the centered quadratic regression equation in Table 6.2.2, the simple slope is $B_{1.2} + 2B_{2.1}x = 1.21 + 2(-.13)x$. For a case with centered $x = -3.10$ (one standard deviation below the mean of centered x), for example, the simple slope $= 1.21 + 2(-.13)(-3.10) = 2.02$. For centered $x = 0$, the simple slope is $1.21 + 2(-.13)(0.00) = 1.21$, the value of B_1. For centered $x = 3.10$ (one standard deviation above the mean of centered x), the simple slope $= 1.21 + 2(-.13)(3.10) = .40$. Finally, for centered $x = 6.20$ (two standard deviations above the mean of centered x), the simple slope becomes negative, $1.21 + 2(-.13)(6.20) = -.40$. Considered together, these four values of the simple slope confirm that interest rises strongly as students accumulate a few credits in a subject, then begins to level off, and finally diminishes after they have taken a relatively high number of credits in the subject.

As has been illustrated here, a series of these simple slopes is useful in describing the nature of the relationship of X to Y across the range of X (or another predictor). In fact, the simple slopes may be tested for significance of difference from zero in the population (see Aiken & West, 1991, p. 78). For example, we might examine whether interest is still rising significantly after three 3-credit courses (or 9 credit hours). The reader is warned that the algebraic expression for the simple slope changes as the order of the polynomial changes.

A Warning About Polynomial Regression Results

The quadratic polynomial regression equation plotted in Fig. 6.2.3 shows a relatively strong downward trend in the data at high values of predictor X. On the other hand, the lowess curve in Fig. 6.2.2(A) does not. (Recall from Section 4.2.2 that the lowess curve is a nonparametric curve that is completely driven by the data—no model is imposed.) The question is how to interpret the outcome of the polynomial regression. The data at hand are very, very sparse at high values of X. There is insufficient information to make a judgment about whether Y drops at very high values of X; hence judgment must be suspended. Both lowess curves and polynomial regression are uninformative at extreme values if data at these extreme values are sparse. In Section 6.2.7 we suggest a strategy of sampling heavily from extreme values on X to ameliorate this problem. It is always important to examine the actual data against both the polynomial regression and some nonparametric curve such as lowess with graphs that show both the fitted curves and the data points, as in Figs. 6.1.1, 6.2.1, 6.2.2, and 6.2.3. One must ask whether the data support the interpretation at the extremes of the predictors.

A Warning About Extrapolation

Extrapolation of a polynomial regression function beyond the extremes of observed X is particularly dangerous. Polynomial regression functions may be very steep and/or change directions at the extremes of the data. If we fit a second order polynomial to an asymptotic relationship that levels off but does not decline at high values of observed X and we project that function to even higher values of X, the function will eventually reverse direction. If we were to extrapolate beyond the highest observed X to predict the criterion, we would make incorrect predictions of scores that differed from the observed asymptote.

6.2.6 Another Example: A Cubic Fit

CH06EX02

We offer another example of polynomial regression to demonstrate its operation for a more complex relationship and to further exemplify the general method. The variable W is of a different nature than X; it has only 6 integer values (1, 2, 3, 4, 5, 6 in uncentered form), they are equally spaced, and for each of these values there is the same number of points, 6 per value

TABLE 6.2.3
Polynomial Multiple Regression Analysis of Regression of Y on W with Centered Data

	Mean	*sd*	Centered correlation matrix			
			Y	w	w^2	w^3
Y	52.75	18.09	Y 1.000	.415	−.052	.251
w	0.00	1.73	w .415	1.000	.000	.934
w^2	2.92	2.53	w^2 −.052	.000	1.000	.000
w^3	0.00	9.36	w^3 .251	.934	.000	1.000

Regression Equations

Linear:	$\hat{Y}_{\text{linear}} =$	$4.34w$	$+52.75$		
95% *CI*:		[1.03, 7.65]			
$t_{B_i}(34)$:		2.66			
Quadratic:	$\hat{Y}_{\text{quadratic}} =$	$4.34w$	$-.38w^2$	$+53.84$	
95% *CI*:		[.98, 7.70]	[−2.67, 1.92]		
$t_{B_i}(33)$:		2.63	−.33		
Cubic:	$\hat{Y}_{\text{cubic}} =$	$14.89w$	$-.38w^2$	$-2.09w^3$	$+53.84$
95% *CI*:		[6.20, 23.58]	[−2.49, 1.74]	[−3.70, −.48]	
$t_{B_i}(32)$:		3.49	−.36	−2.65	

Hierarchical Model

Equation	IVs	R^2	F	*df*	I	F_I	*df*
Linear	w_1	.17	7.09**	1, 34	.173	7.09**	1, 34
Quadratic	w_1, w_2	.18	3.51*	2, 33	.003	.11	1, 33
Cubic	w_1, w_2, w_3	.32	5.10**	3, 32	.148	7.01**	1, 32

**$p < .01$
Note: I is the increment in R^2.

of W. These features suggest that W is a variable produced by experimental manipulation (e.g., number of exposures to a stimulus, or drug dosage level) and that the data structure is the product of a laboratory experiment rather than a field study. These features are important to substantive interpretation of the results but are not relevant to the present analysis; we would proceed as we do here with data at unequal intervals and/or unequal *n*s at each level of W (see Section 6.3.3). The mean of W is 3.5, and we center W, yielding w (−2.5, −1.5, −.5, .5, 1.5, 2.5) in centered form. Consider the plot of points in Fig. 6.2.1, relating Y to centered w and the analysis in Table 6.2.3 for the centered data.

The correlations among the centered predictors are instructive. The centered linear predictor w is perfectly symmetric. Its correlation with the centered quadratic predictor w^2 is 0.00. (Recall that in the previous example the correlation between centered x and x^2 was only .110, with this very small correlation due to minor asymmetry in x.) However, the correlation between w and w^3, the centered cubic predictor, is .934. Even with centering, all odd powers of the linear predictor (w, w^3, w^5, etc.) will be highly intercorrelated; similarly, all even powers will be highly intercorrelated (w^2, w^4, w^6, etc.). Even with these high interpredictor correlations, the analysis can proceed.

The correlation of Y with w (the centered linear aspect of W) is found to be .415, so that 17.2% of the Y variance can be accounted for by w, corresponding to a moderate to large effect

size (Cohen, 1988). Again we caution the reader that this only means that in the population (as well as in the sample) a straight line accounts for some variance and not necessarily that it provides an optimal fit. The equation for this line is given in the middle of Table 6.2.3; the confidence interval for B_1, *CI*: [1.03, 7.65], does not include zero and reflects the strong positive trend in the data, with $t(34) = 2.66, p < .05$, or, equivalently, $F(1, 34)$ for prediction from the linear prediction equation $= 7.07, p < .05$, where $t = \sqrt{F}$. The linear equation is plotted in Fig. 6.2.1, noted by triangles. Since this data set is structured to have replications at each value of w (i.e., 6 cases at each value of w), we can examine the arithmetic mean observed Y score at each value of w, represented by the squares in Fig. 6.2.1. Although the straight line accounts for substantial variance in Y, we note the S-shaped function of these means; they decrease from $w = -2.5$ to $w = -1.5$, then rise to $w = 1.5$, and then drop again, a two-bend pattern.

When the quadratic term w^2 is added into the equation, R^2 increases by only .003, and the confidence interval *CI*: [−2.67, 1.92] for the $B_{2.1}$ coefficient ($B_{2.1} = -.38$) in the quadratic equation includes zero. The data do not support the relevance of the quadratic aspect of W to Y. Note that it is not curvilinearity that is being rejected, but quadratic curvilinearity, that is, a tendency for a parabolic arc to be at least partially descriptive of the regression; a higher order, necessarily curvilinear, aspect of W may characterize the overall form of the relationship of W to Y.

The addition of the cubic term w^3, in contrast, does make a substantial difference, with an increment in R^2 over the quadratic R^2 of .15, a moderate effect size (J. Cohen, 1988). The confidence interval for the cubic coefficient, *CI*: [−3.70, −.48] does not include zero, $t(32) = -2.65, p < .05$. Table 6.2.3 gives the cubic equation, which is also plotted in Fig. 6.2.1 (noted by filled circles). The cubic always gives a two-bend function (although the bends need not appear within the part of the range of the independent variable under study—see Section 6.2.5), and the fit of the cubic equation to the data is visibly improved; the cubic equation better tracks the means of Y at each value of w, though the match is not perfect. We conclude that this regression equation is curvilinear and, more particularly, that it is cubic. By this we mean that the cubic aspect of W relates to Y, and also that the best fitting equation utilizing w, w^2, and w^3 will account for more Y variance in the population than one that has only w and w^2, or (necessarily) only w.

Since we have centered W, the coefficients for the w and w^2 terms in the cubic equations have meaning, though the utility of these terms seems less clear than for the quadratic equation. The $B_{1.23}$ coefficient is the slope of a tangent line to the cubic function at the value $w = 0$. Note in Fig. 6.2.3 that the slope of the cubic function is much steeper at this point than the slope of the linear function, and the $B_{1.23}$ coefficient from the cubic equation is substantially larger (14.89) than the $B_{1.2}$ coefficient in the quadratic equation (Eq. 4.34). The $B_{2.13}$ coefficient indicates the curvature (concave upward, concave downward) at the value $w = 0$. Once again we caution that if one reports these coefficients, their meaning should be explained, because they are little understood.

The cubic function has both a maximum and a minimum; values of the independent variable W at these points may be useful in interpretation. Simple slopes can also be calculated for a cubic function and can be used to describe the function. Computations are given in Box 6.2.2.

6.2.7 Strategy and Limitations

What Order Polynomial, Revisited

In Section 6.2.1 we argued that theory should drive the choice of the order polynomial to be estimated. Yet there may be some exploratory aspect to the analysis. In the quadratic example of student interest in academic minor subjects, we hypothesized a quadratic relationship of

BOX 6.2.2
Maximum, Minimum, and Simple Slopes for a Cubic Equation

In the case of the quadratic equation there is either a minimum or maximum point; in the full cubic equation (Eq. 6.2.4), there are both a minimum and a maximum value. The values of W at which the minimum and maximum occur are given as the two solutions to Eq. (6.2.8).

(6.2.8) $$W_M = \frac{-B_{2.13} \pm \sqrt{B_{2.13}^2 - 3B_{1.23}B_{3.21}}}{3B_{3.21}}$$

For the cubic numerical example in Table 6.2.3, the two solutions to W_M are -1.60 and 1.48, on the centered X scale. These values correspond to uncentered .90 and 3.98 respectively on the original 1 to 6 point X scale. The corresponding values of $\hat{Y} = 37.60$ (the minimum) and $\hat{Y} = 68.28$ (the maximum), respectively.

The *simple slope* of the regression of Y on W in the cubic equation[5] is given as

(6.2.9) $$B_{1.2} + 2B_{2.1}W + 3B_{3.12}W^2.$$

Once again, the simple slope indicates the slope of Y on W at a particular value of W in the polynomial equation. For centered $w = -1.5$, just above the value of w at which $\hat{Y}$ attains a minimum, the simple slope is $14.89 + 2(-.38)(-1.5) + 3(-2.09)(-1.5)^2 = 1.92$ (i.e., the function is rising). For centered $w = 1.5$, just above the value of w at which $\hat{Y}$ attains a maximum, the simple slope is $14.89 + 2(-.38)(1.5) + 3(-2.09)(1.5)^2 = -.36$ (i.e., the function has begun to fall below the maximum).

rising followed by falling interest. We also had some curiosity about whether interest returned among students who had taken a large number of credits, as would be detected with a cubic polynomial. We thus examined the quadratic and cubic relationships. We recommend that the application of polynomial regression begin with setting out the theoretical rationale that might be provided for each polynomial examined. It makes little sense to fit a series of increasingly higher order polynomials and to identify a high-order polynomial term as significant when there is no rationale for the meaning of this term. We also caution that vagaries of the data, particularly outliers in Y with X values at the ends of the range of observed values, may dramatically alter the order of the polynomial that is detected.

Two approaches to exploring polynomial equations exist. First is a *build-up* procedure in which one examines first the linear equation, then the quadratic equation, and so forth; each equation examined includes all lower order terms. A *tear-down* procedure reverses the process, beginning with the highest order equation of interest and simplifying by working down to lower order equations. First, the highest order term in the complete highest order equation also containing all lower order terms is examined by some criterion (traditional significance level, effect size, confidence interval). If the term accounts for a proportion of variance deemed material by the researcher, then the polynomial at this level is retained. If not, then the highest order term is eliminated from the equation, the next lower order equation is estimated, and the highest order term in this equation is examined.

[5] In the cubic equation, Eq. (6.2.4), the first derivative is $B_{1.23} + 2B_{2.13}W + 3B_{3.12}W^2$. This derivative is set to zero to determine the minimum and maximum of the cubic function, given in Eq. (6.2.5).

Each of these approaches has its advantages and its limitations. The difficulty in the build-up procedure is in deciding when to stop in adding terms. Consider the cubic numerical example of Table 6.2.3, in which the linear term is significant in the linear equation and accounts for 17% of the criterion variance: The quadratic term is not significant in the quadratic equation, accounting for less than 1% of variance over and above the linear term, but the cubic term is significant in the cubic equation, accounting for an additional 15% of the variance in the criterion. A rule that indicates we should stop adding terms at the point at which there is no longer a gain in predictability (measured in terms of significance or effect size) would lead us, in the cubic example, to stop with the linear equation. The quadratic term does not contribute material accounted for variance over and above the linear effect. But the relationship is cubic; by build-up cutoff rules we would not reach the cubic equation.

The tear-down procedure has the advantage of insuring a test of the highest order term of interest. In the cubic numerical example, the tear-down procedure would begin with a test of the cubic term in the cubic equation (assuming we had some hypothesis of a cubic relationship). Having found the cubic term to be significant (or to have a large effect size), we would stop the exploration; that is, we would retain the cubic equation and not test terms from lower order equations. Of course, we would still miss any higher order terms that might have been statistically significant although unanticipated.

Must one pick one of the two strategies and rigidly adhere to it in deciding upon the order of the polynomial? Absolutely not. One can first target the theoretically predicted equation (e.g., the quadratic equation in our earlier example of taking further credits in one's minor, depicted in Fig. 6.1.1.) If there is some reason to suspect a higher order trend in the data (e.g., the cubic term), we may then examine the cubic equation. There are no rigid rules for identifying the appropriate equation. Consider a cubic example. Suppose we identified the cubic equation as appropriate in a large sample based on the highest order term but in plotting the equation we saw that the curvature was slight compared to the strong linear trend and that the cubic equation very closely approximated the linear equation plotted on the same graph. We could estimate the linear equation to determine the percentage of variance accounted for by only the linear trend. We might even decide, based on effect size (additional variance accounted for) that we would not retain the cubic equation but would use the linear equation, particularly if there were not a strong theoretical rationale for the higher order term.

There are no hard and fast rules for the inclusion of higher order terms in polynomial equations. An aspect of selecting the final equation is the behavior of residuals, as illustrated in Fig. 6.2.2(B). Model checking (described in Section 4.4) will be informative and is advised. For example, an added variable plot for a higher order term may help clarify the role of this higher order term. To decide between two equations of adjacent order, one of several criteria may be employed:

1. ***Statistical significance.*** We would examine the loss (or gain) in prediction attributable to the highest order term in a tear-down (or build-up) procedure, employing some conventional level of significance.

2. ***Change in R^2.*** The difference in R^2 between two adjacent polynomial equations is closely related to the measure of effect size for change in prediction specified by J. Cohen (1988, pp. 412–413). The change from $R^2_{Y.A}$ to $R^2_{Y.AB}$ is the squared semipartial correlation of B with the criterion, over and above A. The squared partial correlation is given as Eq. (5.4.11):

$$R^2_{YB\cdot A} = \frac{R^2_{Y.AB} - R^2_{Y.A}}{1 - R^2_{Y.A}}. \tag{5.4.11}$$

Cohen (1988) suggested that squared partial correlations of .02, .13, and .26 were reflective of small, moderate, and large effect sizes, respectively.

3. *Change in $\tilde{R}^2$.* Thus far in this chapter, we have been concerned only with the observed R^2 and changes in R^2. However, we know that the addition of any independent variable to a set, even a random, nonpredictive variable, will result in an increase in R^2—a decrease is mathematically impossible and an increment of exactly zero is exceedingly rare. It can be instructive to track the change in the shrunken $\tilde{R}^2$, the estimated proportion of variance in Y accounted for in the population by a polynomial of that order (see Section 3.5.3). Unlike the increment in the R^2 observed, this increment may be positive or negative. A reasonable criterion for deciding between two equations is that the change in $\tilde{R}^2$ between the equations is some arbitrary minimum constant, say between .02 and .05.

The overriding preference in using polynomial equations is that the order of the polynomial be small. Relatively weak data quality in the social sciences (i.e., the presence of substantial measurement error), mitigates against fitting high order equations. Our theories do not predict higher order terms. We gain no benefit in interpretation from fitting an equation of high order. We do not advocate a curve fitting approach in which higher order terms are added merely because they increase the R^2 by small amounts. The likelihood of replication is low indeed. Finally, the polynomial terms may be included in more complex equations. The variable represented by a polynomial may be hypothesized to interact with other variables. As we will see in Chapter 7, each interaction requires predictors involving all the terms in the polynomial equation.

Impact of Outliers and Sampling to Increase Stability

Polynomial equations may be highly unstable and can be grossly affected by individual outliers. Examine Fig. 6.1.1 for the quadratic polynomial. Note that at the highest value of $X, X = 17$, there is only one data point. Its Y value ($Y = 23.38$) lies between the linear and quadratic equations. Changing the value of Y for this point to the highest scale value ($Y = 30.00$, or an increment of 6.62 points) produces substantial changes in the regression equation. In the original analysis, summarized in Table 6.2.2, the increments in prediction were .56, .11, and .01 for the linear, quadratic, and cubic terms, respectively; the cubic term was not significant. In the analysis with the one modified data point, the corresponding increments are .59, .07, and .03. The cubic term is significant, $B_{3.12} = .01$, *CI*: [.01, .02], $t(96) = 3.22, p < .01$. One data point out of 100 cases has produced the cubic effect.

Sparseness of the data points at the ends of the X distribution contributes to the instability. We may sample cases systematically to increase both the stability of regression equations and the power to detect effects (McClelland & Judd, 1993; Pitts & West, 2001). The X variable in the quadratic example is normally distributed, with relatively few cases in the two tails. If X were rectangularly (uniformly) distributed, with an approximately equal number of data points at each value of X, then there would be many more points in the tails, on average just under 6 data points for each of the 17 observed values of X (see Fig. 6.2.4). In a simulation in which 100 values of X were rectangularly distributed, and the same population regression equation was employed, there were 6 points with $X = 17$. Of these 6 points, the point with the highest Y value, initially 30.93, was modified by increasing the Y value the same amount as was done in the normally distributed case, 6.62 points, to 37.55. Before modifying the point the increments in prediction were .63, .24, and .00, for linear, quadratic, and cubic, respectively. After modifying the point, these increments were .64, .22, and .00, respectively. The cubic term accounted for essentially no variance either before or after the modification of the single point.

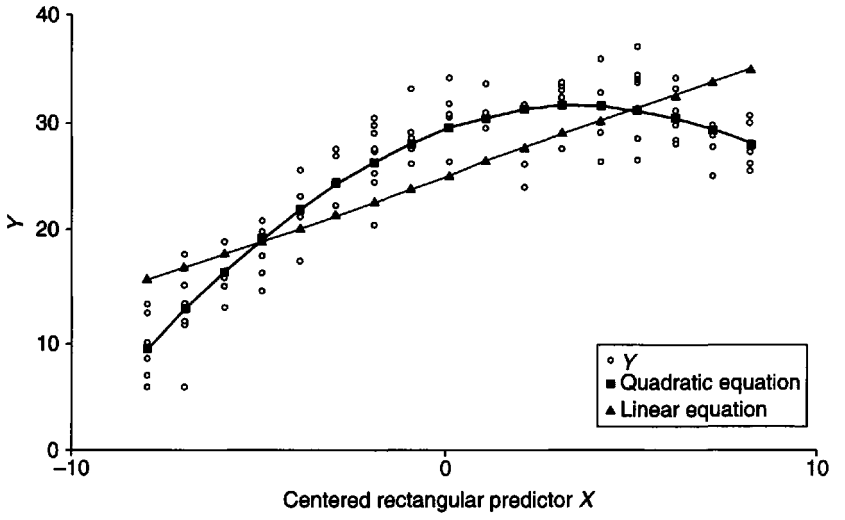


FIGURE 6.2.4 Quadratic relationship with predictor X sampled to have a rectangular distribution. The population regression equation for the quadratic regression equation is identical to that generating the data of Figure 6.1.1 (or centered data of Fig. 6.2.3). The triangles characterize the best fit linear regression of Y on x; the squares, the best fit quadratic regression of Y on x. Note that the density of points in the extremes of the distribution is much greater here than with normally distributed predictor X in Fig. 6.2.3.

There are other more complex rules for sampling X than simple rectangular sampling. These rules arise from an area of statistics termed *optimal design*; McClelland and Judd (1993) and Pitts and West (2001) provide easily accessible introductions to the topic. See Atkinson and Donev (1992) for a complete development.

6.2.8 More Complex Equations

Polynomial equations may also include other predictors. Recall that Janis (1967) predicted that compliance with medical recommendations (Y) would first increase and then decrease as fear of a health threat (X) increases. Medical compliance also increases with the belief that health can be controlled by medical practitioners (W) (Wallston, Wallston, & DeVellis, 1978). We would specify these predictions in the equation $\hat{Y} = B_{1.23}X + B_{2.13}X^2 + B_{3.12}W + B_0$. A simulated data set that follows these predictions is illustrated in Fig. 6.2.5. Instead of the usual flat regression plane generated by two predictors that bear linear relationships to the criterion, we have a curved regression surface. The curve is produced by predictor X that follows the Janis (1967) prediction. A strategy for examining this equation is first to estimate the full equation containing X, X^2 and W as predictors, as well as an equation containing only the linear X and linear W terms. The difference in prediction between these two equations is used to gauge whether the curvilinear term is required. Once this is decided, the effect of the W predictor should be considered in the presence of the X terms retained. Note that in this strategy, the test of curvilinearity of the relationship of X to Y is carried out in the presence of W. The test of W is carried out in the presence of X in the form in which it apparently relates to the criterion. Since predictors X and/or X^2 may be correlated with W, it is appropriate to treat each variable in the presence of the others, as is usual in multiple regression with correlated predictors. The relationship between predictors X and W may not be linear. Darlington (1991)

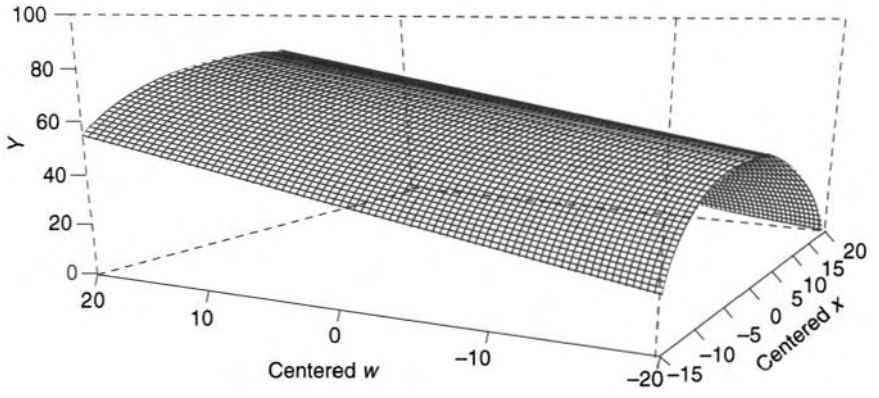


FIGURE 6.2.5 Regression surface for equation $\hat{Y} = .06x - .12x^2 + .86w + 66.08$. Predictor variables X and W produce a regression surface. The quadratic relationship of variable X to Y produces the concave downward shape of the surface. The positive linear relationship of W to Y produces the tilt of the plane. Note that the front right hand corner of the figure represents the lowest values of centered predictors x and w.

warns that difficulties are introduced when there are nonlinear relationships between predictors. In this case, nonlinear relationships of the predictor to the criterion may be obscured in residual scatterplots.

6.3 ORTHOGONAL POLYNOMIALS

In instances in which the values of a variable W form ordered categories, as in the cubic numerical example, it is possible to examine nonlinearity with a special class of variables, *orthogonal polynomials*. Orthogonal polynomials are unique variables that are structured to capture specific curve components—linear, quadratic, cubic, etc. Orthogonal polynomials exist in sets (one linear, one quadratic, one cubic, etc.). For a variable W with u categories, there are $(u - 1)$ orthogonal polynomials in the complete set; for the cubic example with $u = 6$, there are $(u - 1) = 5$ orthogonal polynomials (linear, quadratic, cubic, quartic, quintic). The orthogonal polynomials are used as predictors in a regression equation to represent the variable W. The members of a set of orthogonal polynomials have the special property of being mutually orthogonal, that is, they are uncorrelated with one another. Thus they account for independent portions of variance in a criterion when W represents a complete set of categories ($X = 1, 2, 3, \ldots, u$, with no skipped categories) and there are equal ns at each value of W.

Each orthogonal polynomial is a series of integer coefficients or weights; the weights for each orthogonal polynomial sum to zero. Orthogonal polynomials have two defining properties: mutually orthogonal codes that sum to zero.[6] The specific integer weights of an orthogonal polynomial depend upon the number of values (ordered categories) of the predictors. Table 6.3.1 provides sets of orthogonal polynomials for the linear, quadratic and cubic terms of predictors

[6]Orthogonal polynomials are a special case of code variables called *contrast codes*, which have the properties of being mutually orthogonal and summing to zero. In general, code variables are structured variables specifically designed to be applied to categorical predictors. Chapter 8 is devoted to code variables; contrast codes are presented in Section 8.5.

TABLE 6.3.1
Orthogonal Polynomial Coding for u-Point Scales; First-, Second-, and Third-Order Polynomials for $u = 3$ to 12[a]

$u = 3$		$u = 4$			$u = 5$			$u = 6$			$u = 7$		
X_1	X_2	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
1	1	−3	1	−1	−2	2	−1	−5	5	−5	−3	5	−1
0	−2	−1	−1	3	−1	−1	2	−3	−1	7	−2	0	1
−1	1	1	−1	−3	0	−2	0	−1	−4	4	−1	−3	1
		3	1	1	1	−1	−2	1	−4	−4	0	−4	0
					2	2	1	3	−1	−7	1	−3	−1
								5	5	5	2	0	−1
											3	5	1

$u = 8$			$u = 9$			$u = 10$			$u = 11$			$u = 12$		
X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
−7	7	−7	−4	28	−14	−9	6	−42	−5	15	−30	−11	55	−33
−5	1	5	−3	7	7	−7	2	14	−4	6	6	−9	25	3
−3	−3	7	−2	−8	13	−5	−1	35	−3	−1	22	−7	1	21
−1	−5	3	−1	−17	9	−3	−3	31	−2	−6	23	−5	−17	25
1	−5	−3	0	−20	0	−1	−4	12	−1	−9	14	−3	−29	19
3	−3	−7	1	−17	−9	1	−4	−12	0	−10	0	−1	−35	7
5	1	−5	2	−8	−13	3	−3	−31	1	−9	−14	1	−35	−7
7	7	7	3	7	−7	5	−1	−35	2	−6	−23	3	−29	−19
			4	28	14	7	2	−14	3	−1	−22	5	−17	−25
						9	6	42	4	6	−6	7	1	−21
									5	15	30	9	25	−3
												11	55	33

[a] This table is abridged from Table 20.1 in Owen (1962). (Courtesy of the U.S. AEC.)

with from $u = 3$ to 12 categories.[7] Although for u ordered categories there exist a set of $u - 1$ codes, those above cubic are not likely to be useful for reasons we have discussed (Section 6.2.7). Note that within any set for $u = k$ categories, the sum of the products of corresponding weights for any pair of polynomials is zero, for example for the quadratic and cubic terms for $u = 5$: $(2)(-1) + (-1)(2) + (-2)(0) + (-1)(-2) + (2)(1) = 0$; this is the orthogonality property. Finally, each column of coefficients in Table 6.3.1, when plotted against an equal interval scale from low to high has the distinctive shape of its order (e.g., the quadratic coefficients for X_2 form a concave upward parabola).

Orthogonal polynomials are applied to *ordered categories* which are assumed to be *equally spaced* along some underlying continuum. For example, if the continuum were a drug dosage continuum, the categories should result from equal increments in dosages across the continuum. There are two further requirements for the use of orthogonal polynomials to produce orthogonal portions of variance accounted for by individual curve components. First, the ordered values of the original predictor must be a *full set of numerically equally spaced categories*; that is, if

[7] High-order polynomials and larger numbers of points are available elsewhere. The most extensive are those of Anderson and Houseman (1942), which go up to the fifth order and to $u = 104$. Pearson and Hartley (1954) give up to the sixth order and to $u = 52$. Kleinbaum, Kupper, and Muller (1988) provide all $(u - 1)$ polynomial coefficients for $u = 3, \ldots, 10$ categories.

there are six categories, all six categories must be represented in the data set. Second, there must be *equal numbers of cases* in each ordered category.[8] These conditions for the use of orthogonal polynomials are most likely to be met in laboratory experiments, as in the example of dose responses to drugs.

The polynomials we have used as predictors to this point are *natural polynomials*, generated from the linear predictor by centering and then powering the linear predictor. Natural polynomials are typically correlated so that effects (linear, quadratic, etc.) are necessarily partialed effects. Orthogonal polynomials eliminate this complexity by eliminating the correlation among polynomials. With correlated natural polynomials, we examined increments of variance accounted for by individual polynomial terms. Since, given equal ns, the orthogonal polynomials are uncorrelated with each other, the square of the correlation of each orthogonal polynomial with the criterion r_{Yi}^2 indicates the proportion of variation accounted for by that curve component in the regression equation containing the set of orthogonal polynomials. This squared correlation is exactly equal to the increment in prediction of the criterion by the inclusion of the particular polynomial. In other words, since the orthogonal polynomials are mutually uncorrelated (with equal ns) and there is thus nothing to partial, the correlations r_{Yi} will equal the semipartial correlations sr_{Yi}. These correlations r_{Yi} will also equal the standardized regression coefficients for the orthogonal polynomials β_i.

6.3.1 The Cubic Example Revisited

To accomplish polynomial regression, the natural polynomials are replaced as predictors in the regression equation by the orthogonal polynomials. The linear orthogonal polynomial $X_{\text{linear}} = -5, -3, -1, 1, 3, 5$ replaces w; the quadratic polynomial $X_{\text{quadratic}} = 5, -1, -4, -4, -1, 5$ replaces w^2; similarly, X_{cubic} replaces w^3.

Results of a simultaneous MR analysis are given in Table 6.3.2. Note first that the correlations among the predictors $X_{\text{linear}}, X_{\text{quadratic}}$, and X_{cubic} are all zero. This occurs because the data set contains an equal number of cases at each value of W. Were this not the case, then despite the orthogonality of the polynomial coefficients, the predictors created from substituting the coefficients of the orthogonal polynomials would be correlated. Since the predictors are uncorrelated, the squared correlation of each predictor with the criterion is the proportion of variation accounted for by that curve component. These are *identical* to the respective increments I for the individual curve components in Table 6.2.3. The partialing out of lower order terms for correlated natural polynomials in Table 6.2.3 is replaced in Table 6.3.2 by the use of orthogonal predictors, thus rendering partialing unnecessary. Because the orthogonal polynomial predictors are uncorrelated, $R_{Y.123}^2$ is simply the sum of the three r_{Yi}^2: $R_{Y.123}^2 = .172 + .003 + .148 = .323$, as in the original cubic example. The unstandardized regression coefficients for $X_{\text{linear}}, X_{\text{quadratic}}$, and X_{cubic} in Table 6.3.2 do not equal the corresponding coefficients for w, w^2 and w^3 in Table 6.2.3. Nonetheless, the two equations generate identical predicted scores; that is, the same cubic function is generated by both equations.

Tests of Significance and Confidence Intervals for Curve Components

We describe three approaches to testing the significance of coefficients in orthogonal polynomials. These approaches are all special cases of the general strategy for testing sets of

[8]The derivation of orthogonal polynomials for unequal ns is given in Gaito (1965) and in Appendix C of Kirk (1995).

TABLE 6.3.2
Simultaneous Orthogonal Polynomial Multiple Regression Analysis of Regression of Y on W (same data as Table 6.2.3).

	Mean	*sd*	Uncentered correlation matrix					$r^2_{r_i}$	t_{r_i}
				Y	linear	quadratic	cubic		
Y	52.75	18.09	Y	1.000	.415	−.052	−.385		
linear	.00	3.46	linear	.415	1.000	.000	.000	.172	2.663*
quadratic	.00	3.80	quadratic	−.052	.000	1.000	.000	.003	−.306
cubic	.00	5.56	cubic	−.385	.000	.000	1.000	.148	2.433*

Regression Equation

Cubic:	$\hat{Y}_{\text{cubic}} =$	2.17 linear	−.25 quadratic	−1.25 cubic	+52.75
95% *CI*:		[.62, 3.72]	[−1.66, 1.16]	[−2.22, 2.89]	
t_{B_i} (32):		2.86*	−.36	−2.65*	

R^2 (linear, quadratic, cubic) = .32356 with 3, 32 *df*.

Analysis of Variance of Trend Components

Source	SS	*df*	MS	*F*
Treatment	4020.9167	5		
Linear	1976.0024	1	1976.0024	7.97*
Quadratic	31.5000	1	31.5000	.12
Cubic	1697.5148	1	1697.5148	6.85*
Higher order	315.8995	2	157.9497	.64
Error (within)	7429.8333	30	247.6611	
Total	11450.7500			

R^2 (linear, quadratic, cubic, quartic, quintic) = .35115 with 5, 30 *df*.

* $p < .05$.

predictors in Section 5.5, in which there is a set $\boldsymbol{B}$ of predictor(s) that is tested while another set $\boldsymbol{A}$ is partialed out, the Model 1 approach. What varies from approach to approach is the specific terms partialed out, (i.e., set $\boldsymbol{A}$).

1. The first strategy involves specifying a single polynomial equation at the highest order trend level of interest and testing each regression coefficient in the equation against the $MS_{residual}$ from that highest order regression equation. Tests of terms in the cubic equation in Table 6.2.3 exemplify this strategy. The $MS_{residual}$ is a Model 1 error in that the particular term being tested constitutes set $\boldsymbol{B}$, and all other terms in the equation, all of which are partialed, constitute set $\boldsymbol{A}$. In the cubic equation of Table 6.2.3, the test of the quadratic coeffcient (set $\boldsymbol{B}$) is carried out with both the linear and cubic terms (set $\boldsymbol{A}$) partialed out.

2. The second strategy is a hierarchical strategy, as described in Section 6.2.4, in which one tests the linear term in the linear equation, the quadratic term in the quadratic equation, and so forth. The $MS_{residual}$ for each test comes from a different regression equation. Again this is a form of the Model 1 strategy; the particular term being tested constitutes set $\boldsymbol{B}$; all *lower order* terms constitute set $\boldsymbol{A}$. In this strategy, the quadratic term (set $\boldsymbol{B}$) is tested in the quadratic equation, with the linear term (set $\boldsymbol{A}$) partialed out.

3. The third strategy is one that we have not heretofore encountered. In this strategy, all $(u-1)$ possible trend components are included in a single regression equation. The coefficient for each trend component is tested against $MS_{residual}$ from the complete equation. This is again a form of Model 1 testing, in which the term being tested constitutes set ***B*** and all other terms, both lower and higher order, constitute set ***A***. For the cubic example, with $(u-1)=5$ orthogonal polynomials, the test of the quadratic term (set ***B***) would be carried out with the linear, cubic, quartic, and quintic terms (set ***B***) partialed out.

Coefficients in different equations. With orthogonal polynomials and equal *n*s, since the predictors are uncorrelated, each coefficient in the equation containing one set of curve components equals the corresponding coefficient in an equation with fewer or more curve components. For the cubic example, the linear, quadratic, and cubic equations are as follows: linear, $\hat{Y} = 2.17$ linear $+ 52.75$; quadratic: $\hat{Y} = 2.17$ linear $-.25$ quadratic $+ 52.75$; cubic: $\hat{Y} = 2.17$ linear $- .25$ quadratic $- 1.25$ cubic $+ 52.75$.

Residual variance in different equations and resulting statistical power. Although the three approaches will yield the same regression coefficients for corresponding polynomial components, these approaches differ in the specific sources of variance included in $MS_{residual}$, as we have seen for the test of the quadratic term in each strategy described earlier. Thus they yield different values of $MS_{residual}$ with different associated degrees of freedom, and thus different tests of significance of individual coefficients and different confidence intervals on the coefficients. The relative statistical power of the three approaches depends on the magnitude of trend components other than the specific individual trend being tested and whether these other trends are included as predictors in the model or are pooled into $MS_{residual}$.

Table 6.3.3 shows the *t* tests for the linear coefficient in the data from Table 6.3.2 according to the three testing approaches. There is a general principle that emerges: If one includes in the model extra terms that do not account for material amounts of variance (e.g., the quartic and quintic terms), one reduces power. In contrast, including a term that accounts for a material amount of variance (here the cubic term), increases power for the tests of other effects.

TABLE 6.3.3
t Tests of Alternative Models for Testing Orthogonal Polynomials Under Three Approaches to Testing

		Approach					
		(1) Simultaneous subset		(2) Hierarchical buildup		(3) Simultaneous full set	
Aspect of *W*	Increment	*t*	*df*	*t*	*df*	*t*	*df*
Linear	.173	2.857	32	2.663	34	2.825	30
Quadratic	.003	−.361	32	−.332	33	−.357	30
Cubic	.148	−2.648	32	−2.648	32	−2.618	30
Quartic						.755	30
Quintic						.840	30

Note: Approach (1): each term tested in a model contained the same subset of all possible trend components (linear, quadratic, and cubic); approach (2): each term tested in a model in which all lower order terms are included; approach (3); each term tested in a model containing all possible trend components (linear, quadratic, cubic, quartic, and quintic).

Choosing Among the Three Approaches

The choice among the three approaches requires resolution of the competing demands of maximizing statistical power, minimizing the risk of having negatively biased statistical tests (i.e., tests that underestimate statistical signficance), and obtaining confidence intervals that are as narrow as possible.

Approach 1 requires that a particular level of polynomial be specified in advance, hopefully driven by theory. It is recommended when k is small (e.g., 3) and n is large. We expect that k chosen based on theorizing in the behavioral sciences will, in fact, be small—at most cubic. It is generally sensible to exclude from error the k terms in which one is seriously interested, and when n is sizable, the df for $MS_{residual}$, $n - k - 1$ df, are sufficient to maintain power.

Approach 2 in the build-up mode is very much exploratory. If there are substantial higher order components, then tests of lower order components are negatively biased, as in the cubic numerical example. When there are clear hypotheses about the order of the appropriate equation, approach 1 is preferred.

Approach 3 is safe in its avoidance of underestimating tests of significance because all trend components are included. It requires that $n - u$ be large enough to achieve adequate power for the effect size expected for the critical term. It should be used if the test of gain in prediction from the higher order terms (e.g., the quartic and quintic test in our cubic example) is of substantial effect size (i.e., accounting for more than a few percent of variance). In such a case, the approach 1 tests would be negatively biased.

Trend Analysis in Analysis of Variance and Orthogonal Polynomial Regression

The third approach to testing orthogonal polynomials in MR is isomorphic with trend analysis in a one factor non-repeated measures ANOVA applied to a factor consisting of ordered categories. In trend analysis, $SS_{between\ cell}$ from the overall design is partitioned into $(u - 1)$ trend components (linear, quadratic, etc.) for the u levels of the factor. $MS_{residual}$ from the regression analysis with all trend components is identical to $MS_{within\ cell}$ from the one factor ANOVA. The bottom section of Table 6.3.2 presents a trend analysis of the cubic data set from Table 6.3.2, showing the partition of $SS_{treatment}$ into trend components. The ratio of the SS for each component (linear, quadratic, etc.) to SS_{total} yields the squared correlation of each trend component with the criterion. The F tests for the individual components, each with (1, 30) df are the squares of the t tests reported in Table 6.3.3, approach 3. The overall $R^2 = SS_{treatment}/SS_{total} = .35$ with the five components included represents the maximum predictability possible from a full set of trend components. Here trend analysis is applied to non-repeated measures data, with different subjects in each ordered category of predictor W. Trend analysis may also be applied to repeated measures data, in which the same individuals appear at each level of the repeated measured factor (see Section 15.3.2).

6.3.2 Unequal *n* and Unequal Intervals

The previous example had equal ns at each of the u points of variable W. This property is required in order that the curve components of variable W account for orthogonal portions of variance in the criterion Y. When the ns are not equal, the correlations r_{ij} among the aspects of W are generally not zero, because the orthogonal polynomial coefficients are unequally weighted. With the curve components not independent, equality among r^2_{Yi}s and sr^2_{Yi}s in the simultaneous model is lost and r^2_{Yi} no longer equals the amount of variance accounted for purely by the ith trend component, I_i. This, however, constitutes no problem in analyzing the

data—we simply revert to strategies involving the exploration of partialed effects, as discussed in Section 6.2.4.

Another circumstance that defeats the simplicity of the use of orthogonal polynomials but can be handled in MR is inequality of the given intervals in the variable *W*. A scale with unequal given intervals can be conceived as one with equal intervals some of whose scale points have no data. For example, on a $u = 9$-point scale (e.g., an experiment with potentially 9 equally spaced drug dosages) data may have been obtained for only $q = 5$ values of *W*: 1, 2, 4, 6, 9. We can code these as if we had a 9-point scale, using coefficients under $u = 9$ in Table 6.3.1, but, of course, omitting the coded values for points 3, 5, 7, and 8. The effect on the r_{ij} among the trend components is as before: They take on generally nonzero values, and the analysis proceeds by examining partialed effects. With q different observed values of variable *W*, in all only $(q - 1)$ trend components are required to perfectly fit the q means of the criterion, computed for each value of *W*.

Finally, using orthogonal polynomial coefficients and partialed contributions of individual trend components, we can analyze problems in which neither the given intervals nor the numbers of observations per scale are equal, by simply proceeding as in the preceding paragraph (J. Cohen, 1980).

6.3.3 Applications and Discussion

A number of circumstances in research may seem to invite the use of orthogonal polynomials. However, alternative approaches, discussed here, may be more appropriate.

Experiments

The greatest simplicity in the application of orthogonal polynomials occurs under equal n, and u equally spaced points of a variable *V*. Such data are produced by experiments where *V* is some manipulated variable (number of rewards, level of illumination, size of discussion group) and *Y* is causally dependent on this input. Typically, such data sets are characterized by relatively small number of observations per condition. These are the circumstances in which testing approach 1 would be preferred if there are few trends of interest, approach 2 if there are a large number of trends of interest. Note that throughout this chapter, the n observations at each of the u points of predictor *X* are taken to be independent. The frequently occurring case where n subjects or matched sets of subjects yield observations for each of the conditions is not analyzed as described previously. Chapter 15 addresses the analysis of repeated measures data.

Sampling from a Continuum

In some research applications cases are sampled at specific values or small subranges across a continuum. For example, in a developmental study, cases may be sampled in discrete age ranges (e.g., 3–3.5 years of age, 4–4.5 years, etc.). If so, then these discrete categories may be considered as points on a continuum, as in experiments, and orthogonal polynomials applied to the series of ordered age categories. There is argument in favor of treating the ages continuously in MR with polynomial regression applied to the actual ages. Statistical power is higher with MR applied to actual ages than to data sampled with coarse categories (Pitts & West, 2001).

Sometimes, as in surveys, continua are broken into response categories, as in age: under 20, 20–30, 30–40, . . . , over 70. Such variables may be treated provisionally as if they represented equal intervals, even if the two extreme categories are "open." The end intervals produce lateral displacements of the true curve. If the true relationship is (or is assumed to be) smooth, distortion due to inequality of the end intervals may be detected by the polynomial.

Serial Data Without Replication

Thus far we have been assuming that at each of the u points of a variable V, there are multiple Y observations or "replications," the means of which define the function to be fitted. We now consider instances in which a single individual is observed repeatedly over time, for example, ratings by a psychotherapist of each of a series of 50 consecutive weekly sessions; thus $n = u = 50$. The purpose of analysis might be to test propositions about the nature of the trend over time of Y for this patient. Observations on a single case collected repeated over time are referred to as *time series* data and are appropriately treated in *time series analysis*, discussed in Section 15.8. Time series analysis takes into account the autocorrelation among successive observations on the individual (see Sections 4.4 and 4.5 for discussions of autocorrelation).

6.4 NONLINEAR TRANSFORMATIONS

Thus far in this chapter, we have sought to assure the proper representation of a quantitatively expressed research factor X, by coding it as a set of k IVs, each representing a single aspect of X: as a set of integer powers $X, X^2, \ldots, X^k$ in *polynomial regression* (Section 6.2) or as a set of orthogonal polynomials (Section 6.3). In these treatments of the IV, we left the scale of the dependent variable Y intact. The purpose of our treatments of the IVs as polynomials or orthogonal polynomials was to permit the use of linear MR to characterize a nonlinear relationship of the IV to Y.

6.4.1 Purposes of Transformation and the Nature of Transformations

In this section we consider a wide variety of transformations that can be made on predictors X or the dependent variable Y. By transformations we mean changes in the scale or units of a variable, for example, from X to $\log X$ or to $\sqrt{X}$. From a statistical perspective there are three overarching goals for carrying out transformations.

1. *Simplify the relationship.* First, transformations are employed to simplify the relationship between X and Y. Nonlinear transformations always change the form of the relationship between X and Y. The primary goal is to select a transformation that leads to the simplest possible X–Y relationship—nearly always a linear relationship. This first goal, of simplifying the relationship, often is not merely to create a mathematical condition (linearity) but rather to create more conceptually meaningful units. For example, when economists use log(dollars) as their unit of analysis, it is partly because this function better reflects the utility of money; the use of the decibel scale for loudness and the Richter scale for earthquake intensity provide other examples (Hoaglin, 1988).

2. *Eliminate heteroscedasticity.* Transformations of Y also serve to change the structure of the variance of the residuals around the best fitting regression function. The second goal of transformations is to eliminate problems of heteroscedasticity (unequal conditional variances of residuals, see Sections 4.3.1 and 4.4.4).

3. *Normalize residuals.* Finally, nonlinear transformations serve to change the distribution of both the original variable and the residuals. The third goal of transformations is to make the distribution of the residuals closer to normal in form.

In many cases, particularly when data are highly skewed and the range of variables is wide, transformations simultaneously achieve all three goals, greatly simplifying the interpretation of the results and meeting the assumptions in linear MR. Yet this result is not inevitable. Nonlinear

transformations operate simultaneously on the form of the relationship, the variance of the residuals, and the distribution of the residuals. Occasionally, transformations may improve the regression with respect to one goal while degrading it with respect to others.

The effect of a *linear transformation* of a variable (multiplying or dividing by a constant, adding or subtracting a constant) is to stretch or contract the variable *uniformly* and/or to shift it up or down the numerical scale. Because of the nature of the product-moment correlation, particularly its standardization of the variables, linear transformation of X or Y has no effect on correlation coefficients of any order or on the proportions of variance that their squares yield. The POMP scores described in Section 5.2.2 are a linear transformation of an original scale.

The transformations we will encounter here, in contrast, are *nonlinear* transformations, such as $\log X$, or a^X, or 2 arcsin $\sqrt{X}$. These transformations stretch or contract X nonuniformly. However, they are also strictly *monotonic*, that is, as X increases, the transformed value either steadily increases or steadily decreases.

Linearizing Relationships

When the analyst is seeking to simplify the relationship between X and Y and a constant additive change in X is associated with other than a constant additive change in Y, the need for nonlinear transformations may arise. Certain kinds of variables and certain circumstances are prone to monotonic nonlinearity. For example, in learning experiments, increases in the number of trials do not generally produce uniform (linear) increases in the amount learned. As another example, it is a fundamental law of psychophysics that constant increases in the size of a physical stimulus are not associated with constant increases in the subjective sensation. As children mature, the rate of development slows down. As total length of prior hospitalization of psychiatric patients increases, scores of psychological tests and rating scales do not generally change linearly. Certain variables are more prone to give rise to nonlinear relationships than others: time-based variables such as age, length of exposure, response latency; money-based variables such as annual income, savings; variables based on counts, such as number of errors, size of family, number of hospital beds; and proportions of all kinds.

The application of nonlinear transformations arises from the utilization of a simple mathematical trick. If Y is a logarithmic function of X, then, being nonlinear, the Y–X relationship is not optimally fitted by linear correlation and regression. However, the relationship between Y and $\log X$ is linear. Similarly, if Y and X are reciprocally and hence nonlinearly related, Y and $1/X$ are linearly related. Thus, by taking nonlinear functions of X or Y that represent specific nonlinear aspects, we can *linearize* some relationships and bring them into our MR system.

Assumptions of Homoscedasticity and Normality of Residuals

In addition to the linearization of relationships, nonlinear transformation is of importance in connection with the formal statistical assumptions of regression analysis—that residuals be normally distributed and of constant variance (homoscedastic) over sets of values of the IVs (see Section 4.3.1). If data exhibit heteroscedasticity, the standard errors of regression coefficients are biased, as they are if residuals are non-normal, thus leading to less accurate inferences. If heteroscedasticity and nonnormality of residuals obtain in MR, it may be possible to find a nonlinear transformation of Y that not only linearizes the X-Y relationship but simultaneously transforms the residuals from MR predicting the transformed Y to be closer to meeting the conditions of homoscedasticity and normality.

In the remainder of Section 6.4 we consider, in turn, the linearizing of relationships and transforming to achieve homoscedasticity and normality of residuals. We also address issues in the use of transformations. Some of what we present is familiar in the behavioral sciences.

Yet other approaches to transformations presented here have rarely been employed in the behavioral sciences, although they are standardly used in statistics and other social sciences—for example, economics. We believe that part of the reason they have not been employed in the behavioral sciences is that researchers in the behavioral sciences have paid less attention to the assumptions of MR and the negative impact of violation of these assumptions on accuracy of inference.

6.4.2 The Conceptual Basis of Transformations and Model Checking Before and After Transformation—Is It Always Ideal to Transform?

It is very important to consider the conceptual basis of transformations. In many disciplines, certain transformations are considered standard procedure, arising out of long experience. Transformations should always be conducted with cognizance of other researchers' experience with similar variables. In fact, there are times when it may be unwise to carry out (or not) a transformation on a particular data set, when the circumstances of that data set suggest something other than what is standard in a research area. On the other hand, slavish adherence to historical precedents may be problematic, if an area has failed in the past to consider important aspects of the data (e.g., heteroscedasticity) that may suggest the need for transformation.

Diagnostic plots, described in Sections 4.2 and 4.4, are useful for examining whether relationships are linear and for checking assumptions on residuals. These plots may signal the need for transformation. Kernel density plots of individual variables are useful for detecting skewness (see Fig. 4.2.3). Scatterplots of residuals against predicted scores with lowess fit lines highlight nonlinearity of relationship. If plots include lowess lines one standard deviation above and below the lowess fit line, heteroscedasticity is also highlighted, as illustrated in Fig. 4.4.5. Before undertaking transformations, it is also important to use regression diagnostics to ask whether one or few influential data points are producing the nonlinearity or violations of assumptions, as in the example in Section 6.2.7, in which a single outlying case produced a cubic trend. If justified, the removal of the case may eliminate the need for transformation.

Model checking after transformation is also important, because transformation to remedy one aspect of a regression situation (e.g., nonlinearity) may lead to problems in other aspects of the regression situation. For example, in a situation in which Y is transformed to achieve linearity, there may be no outliers in the original analysis with Y as the dependent variable. However, in the revised analysis in which transformed dependent variable Y' is employed, outliers may have been produced by the transformation. Or, it is possible that in the original regression equation with Y, the residuals were approximately homoscedastic and normal; after transformation, they may not be, so that it may be more desirable to stay with the original nonlinear regression. The issue of when to transform is discussed in more detail in Section 6.4.18.

6.4.3 Logarithms and Exponents; Additive and Proportional Relationships

The transformations we employ often involve logarithms and exponents (or powers). A quick review of their workings is in order. Regression equations we estimate after variable transformation often involve combinations of variables in original form and variables transformed into logarithms; the meaning of relationships in these equations is also explored here.

A logarithm of a value X to the base m is the exponent to which the value m must be raised in order to produce the original number. We are probably most familiar with *base 10* logarithms, noted $\log_{10}$, where $m = 10$. For example, $\log_{10} 1000 = 3$, since $10^3 = 1000$.

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 $\boldsymbol{e^X}$ 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 $\hat{Y}' = \log Y = B_1X + B_0$. 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

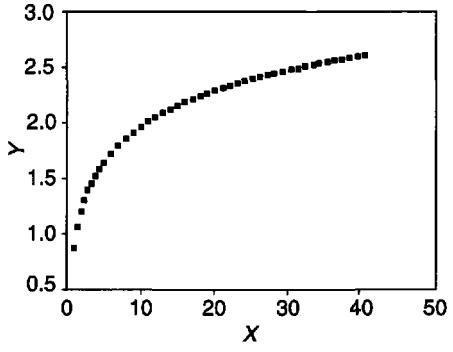
$$\log \hat{Y} = d \log X + \log c \quad \text{or} \quad \log \hat{Y} = B_1 \log X_1 + B_0. \tag{6.4.11}$$

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

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

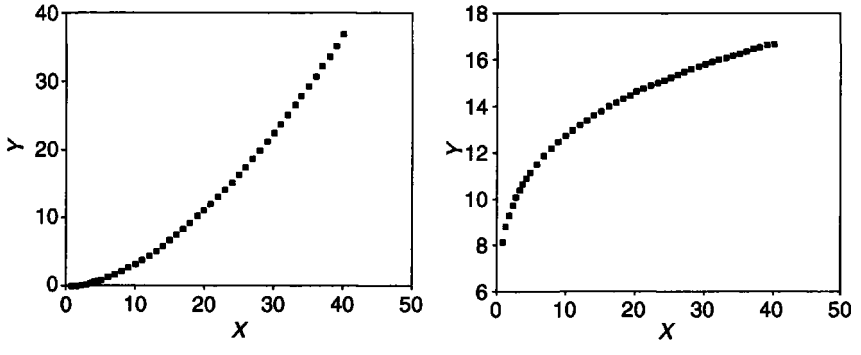
$$d = B_1. \tag{6.4.13}$$

(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 $\hat{Y}_{\log} = \hat{Y}_{\text{original units}}$.

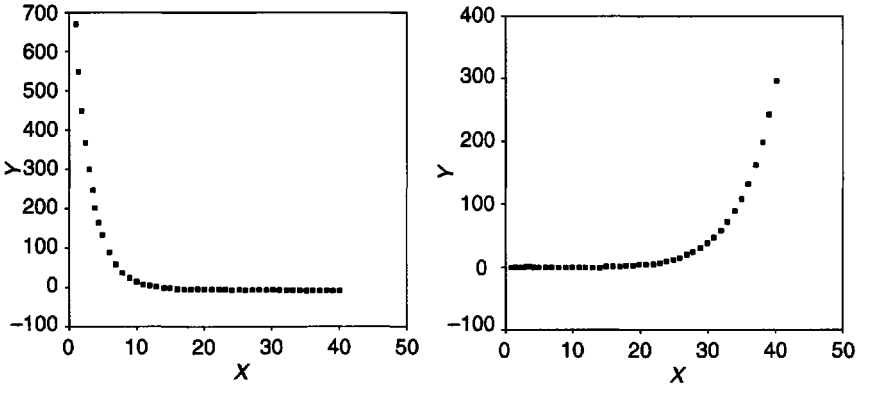
Exponential Growth Model Relationships

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

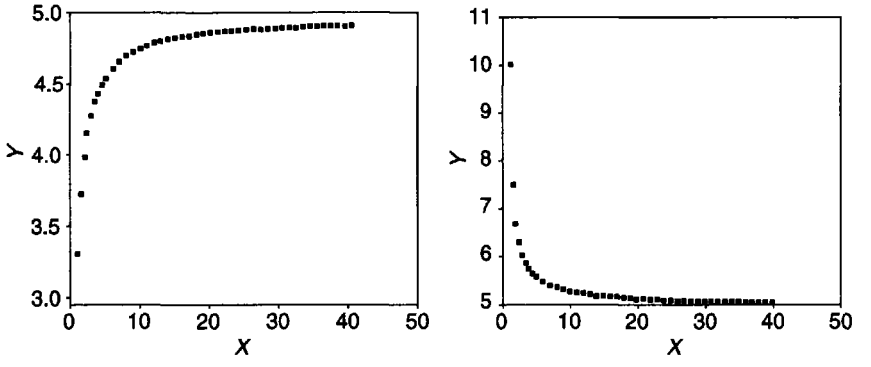
In this model, the change in Y at any point depends on the level of Y. If $d > 0$, Y grows from a starting value of c when $X = 0$, with Y rising in ever increasing amounts, for example, as in college tuition over time, referred to as *exponential growth*. Exponential growth is illustrated in

(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 $\lambda = \frac{1}{2}$). 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*.

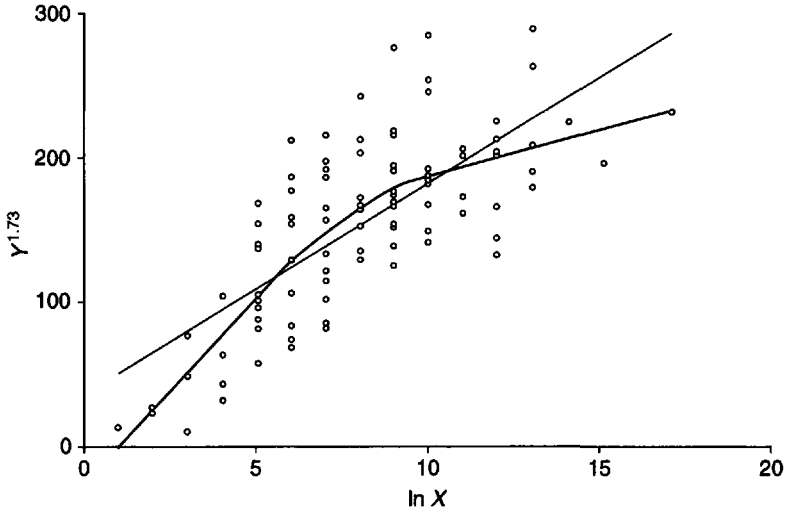
Variable Range and Power Transformations

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 $\ln(X)$.

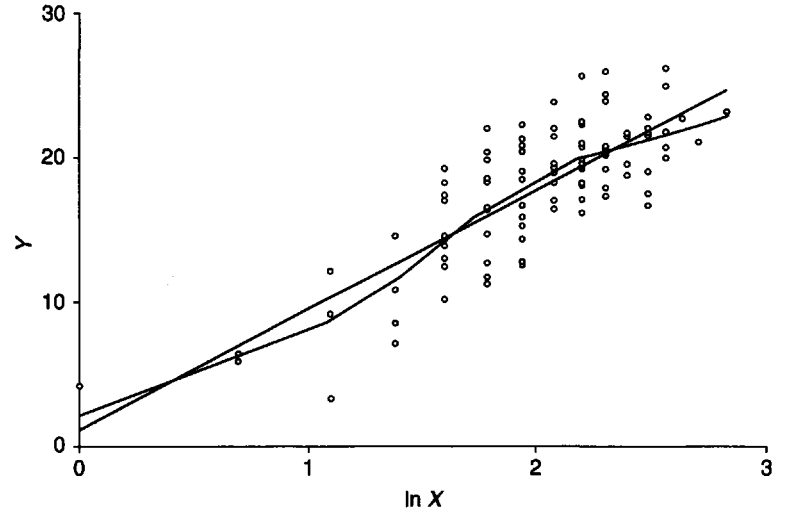


FIGURE 6.4.2 Power transformation of Y versus logarithmic transformation of X to linearize a relationship. The data are the same data as in Fig. 6.1.1 and Table 6.2.1.

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

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

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,

(6.4.39)
$$L_{P'} = -L_P$$

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):

$$R_1^2 = 1 - \frac{\sum (Y_i - \hat{Y}_{\text{original units}})^2}{\sum (Y_i - M_Y)^2}. \tag{6.4.40}$$

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