4 Transforming Data

"Classical" statistical models, for example, linear least-squares regression, make strong assumptions about the structure of data—assumptions which, more often than not, fail to hold in practice. One solution is to abandon classical methods in favor of more flexible alternatives, such as nonparametric regression analysis. These newer methods are valuable, and I expect that they will be used with increasing frequency, but they are more complex and have their own limitations, as we saw in Chapter 2.[1]

It is, alternatively, often feasible to transform the data so that they conform more closely to the restrictive assumptions of classical statistical models. In addition, and as we will discover in this chapter, transformations can often assist in the examination of data, even in the absence of a statistical model. The chapter introduces two general families of transformations and shows how they can be used to make distributions symmetric, to make the relationship between two variables linear, and to equalize variation across groups.

> Transformations can often facilitate the examination and statistical modeling of data.

4.1 The Family of Powers and Roots

There is literally an infinite variety of functions $f(x)$ that could be used to transform a quantitative variable X. In practice, of course, it helps to be more restrictive, and a particularly useful group of transformations is the "family" of powers and roots:

$$X \rightarrow X^p \tag{4.1}$$

where the arrow indicates that we intend to replace X with the transformed variable X^p. If p is negative, then the transformation is an inverse power: For example, $X^{-1} = 1/X$ (i.e., inverse), and $X^{-2} = 1/X^2$ (inverse square). If p is a fraction, then the transformation represents a root: For example, $X^{1/3} = \sqrt[3]{X}$ (cube root) and $X^{-1/2} = 1/\sqrt{X}$ (inverse square root).

For some purposes, it is convenient to define the family of power transformations in a slightly more complex manner, called the *Box-Cox family* of transformations (Box and Cox, 1964):[2]

$$X \rightarrow X^{(p)} \equiv \frac{X^p - 1}{p} \tag{4.2}$$

We use the parenthetical superscript (p) to distinguish this definition from the more straightforward one in Equation 4.1. Because $X^{(p)}$ is a linear function of X^p, the two transformations have

[1] Also see Chapter 18.

[2] In addition to revealing the relative effect of different power transformations, the Box-Cox formulation is useful for estimating a transformation as a parameter, as in Section 4.6.

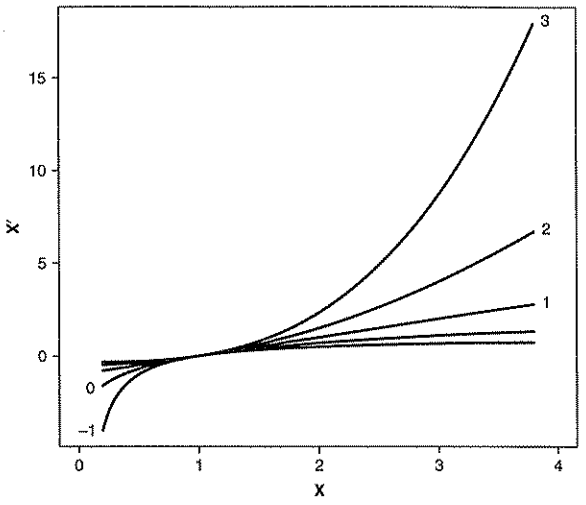

Figure 4.1 The family of power transformations X' of X. The curve labeled p is the transformation $X^{(p)}$, that is, $(X_p-1)/p$; $X^{(0)}$ is $\log_e X$.

the same essential effect on the data, but, as is apparent in Figure 4.1, the definition in Equation 4.2 reveals more transparently the essential unity of the family of powers and roots:[3]

- Dividing by p preserves the direction of X, which otherwise would be reversed when p is negative, as illustrated in the following example:

X	X^{-1}	$\frac{X^{-1}}{-1}$	$\frac{X^{-1}-1}{-1}$
1	1	−1	0
2	1/2	−1/2	1/2
3	1/3	−1/3	2/3
4	1/4	−1/4	3/4

  Note that subtracting 1 from the numerator does not affect *differences* between adjacent transformed values in the table.
- The transformations $X^{(p)}$ are "matched" above $X = 1$ both in level and in slope: (1) $1^{(p)} = 0$, for all values of p; and (2) each transformation has a slope of 1 at $X = 1$.[4]
- Matching the transformations facilitates comparisons among them and highlights their relative effects. In particular, descending the "ladder" of powers and roots toward $X^{(-1)}$ compresses the large values of X and spreads out the small ones; ascending the ladder of powers and roots toward $X^{(2)}$ has the opposite effect.[5] As p moves further from $p = 1$ (i.e., no transformation) in either direction, the transformation grows more powerful.
- The power transformation X^0 is useless because it changes all values to 1, but we can think of the log (i.e., logarithm) transformation as a kind of "zeroth" power: As p gets very

[3] See Exercise 4.1.

[4] *That is, the derivative of $X^{(p)}$ at $X = 1$ is 1; see Exercise 4.2.

[5] The heuristic characterization of the family of powers and roots as a "ladder" follows Tukey (1977).

close to 0, the log function more and more closely approximates $X^{(p)}$.[6] Because the log transformation is so useful, we will, by convention, take $X^{(0)} \equiv \log_e X$, where $e \approx 2.718$ is the base of the natural logarithms.[7]

In practice, it is generally more convenient to use logs to the base 10 or base 2, which are more easily interpreted than logs to the base e: For example, increasing $\log_{10} X$ by 1 is equivalent to multiplying X by 10; increasing $\log_2 X$ by 1 is equivalent to doubling X. Selection of a base for the log transformation is essentially arbitrary and inconsequential, however, because changing bases is equivalent to multiplying by a constant; for example,

$$\log_{10} X = \log_{10} e \times \log_e X \approx 0.4343 \times \log_e X$$

Likewise, because of its relative simplicity, we usually use X^p in applied work in preference to $X^{(p)}$ when $p \neq 0$. Transformations such as log, square root, square, and inverse have a long history of use in data analysis, often without reference to each other; thinking about these transformations as members of a family facilitates their systematic application, as illustrated later in this chapter.

The powers and roots are a particularly useful family of transformations: $X \to X^p$. When $p = 0$, we employ the log transformation in place of X^0.

The effects of the various power transformations are apparent in Figure 4.1 and in the following simple examples (in which the numbers by the braces give *differences* between adjacent values):

$-1/X$	$\log_2 X$	X	X^2	X^3
-1	0	1	1	1
$\frac{1}{2}$	1	1	3	7
$-1/2$	1	2	4	8
$\frac{1}{6}$	0.59	1	5	19
$-1/3$	1.59	3	9	27
$\frac{1}{12}$	0.41	1	7	37
$-1/4$	2	4	16	64

Power transformations are sensible only when all the values of X are positive. First of all, some of the transformations, such as log and square root, are undefined for negative or zero values. Second, even when they are defined, the power transformations are not monotone—that is, not order preserving—if there are both positive and negative values in the data; for example,

[6]*More formally,

$$\lim_{p \to 0} \frac{X^p - 1}{p} = \log_e X$$

[7]Powers and logarithms are reviewed in Appendix C.

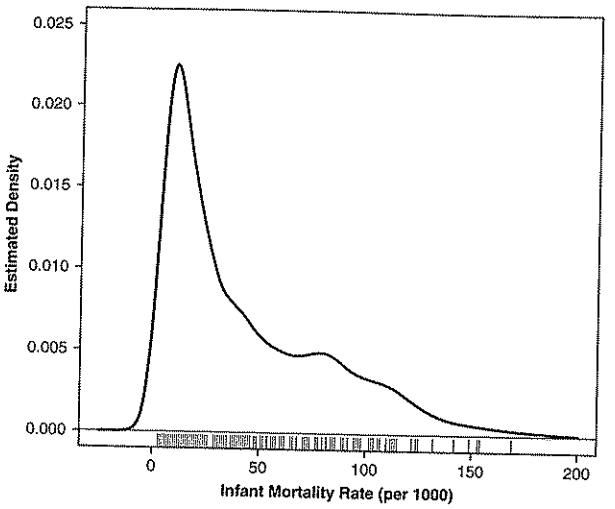

Figure 4.2 Adaptive-kernel density estimate for the distribution of infant mortality rates of 193 nations of the world. The data values are displayed in the rug-plot at the bottom of the figure.

suitable start—an adequate power transformation can usually be found in the range $-2 \leq p \leq 3$. We usually select integer values of p or simple fractions such as $\frac{1}{2}$ or $\frac{1}{3}$.

> Power transformations preserve the order of the data only when all values are positive and are effective only when the ratio of the largest to the smallest data values is itself large. When these conditions do not hold, we can impose them by adding a positive or negative start to all the data values.

4.2 Transforming Skewness

Power transformations can make a skewed distribution more symmetric. But why should we bother?

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data. Recall from the previous chapter, for example, the distribution of infant mortality rates, redisplayed in Figure 4.2.[8]
- Apparently outlying values in the direction of the skew are brought in toward the main body of the data when the distribution is made more symmetric. In contrast, unusual values in the direction opposite to the skew can be hidden prior to transforming the data.
- Some of the most common statistical methods summarize distributions using means. Least-squares regression, which traces the mean of Y conditional on X, comes immediately to mind.[9] The mean of a skewed distribution is not, however, a good summary of its center.

[8] Adapting Figure 3.5.
[9] See Chapter 5.

The following simple example illustrates how a power transformation can eliminate a positive skew:

	X	$\log_{10} X$	
9	1	0	1
	10	1	
90	100	2	1
900	1000	3	1

Descending the ladder of powers to log X makes the distribution more symmetric by pulling in the right tail. Ascending the ladder of powers (toward X^2 and X^3) can, similarly, "correct" a negative skew.

An effective transformation can be selected analytically or by trial and error.[10] Examining the median and the hinges, moreover, can provide some guidance to trial and error. A convenient property of order statistics—including the median and hinges—is that they are preserved under nonlinear monotone transformations of the data, such as powers and roots; that is, if $X' = X^{(p)}$, then $X'_{(i)} = [X_{(i)}]^{(p)}$, and thus median$(X')$ = [median$(X)]^{(p)}$.[11] This is not the case for the mean and standard deviation.

In a symmetric distribution, the median is midway between the hinges, and consequently, the ratio

$$\frac{\text{Upper hinge} - \text{Median}}{\text{Median} - \text{Lower hinge}}$$

is approximately 1. In contrast, a positive skew is reflected in a ratio that exceeds 1 and a negative skew in a ratio that is smaller than 1. Trial and error can begin, therefore, with a transformation that makes this ratio close to 1.

Some statistical software allows the transformation p to be selected interactively using a "slider," while a graph of the distribution—for example, a density plot—is updated when the value of p changes. This is an especially convenient and effective approach. A static alternative is to show parallel boxplots for various transformations, as in Figure 4.3 for the infant mortality data.

For the distribution of infant mortality rates, we have

Transformation	H_U	*Median*	H_L	$\frac{H_U - \text{Median}}{\text{Median} - H_L}$
X	68	30	13	2.23
$\sqrt{X}$	8.246	5.477	3.605	1.48
$\log_{10} X$	1.833	1.477	1.114	0.98
$-1/\sqrt{X}$	−0.1213	−0.1825	−0.2773	0.65
$-1/X$	−0.01471	−0.03333	−0.07692	0.43

[10] See Sections 4.6 and 12.5 for analytic methods for selecting transformations.

[11] There is some slippage here because the median and hinges sometimes require *averaging* adjacent order statistics. The two averaged values are seldom very far apart, however, and therefore the distinction between the median of the transformed values and the transformation of the median is almost always trivial. The same is true for the hinges. The results presented for the example give the median and hinges of the transformed data.

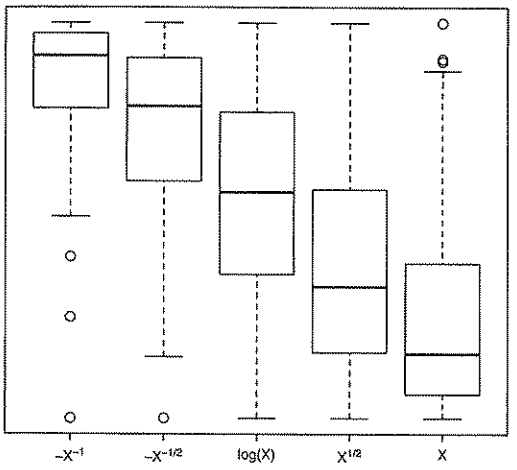

Figure 4.3 Boxplots for various power transformations of infant mortality; because the distribution of infant mortality is positively skewed, only transformations "down" the ladder of powers and roots are considered.

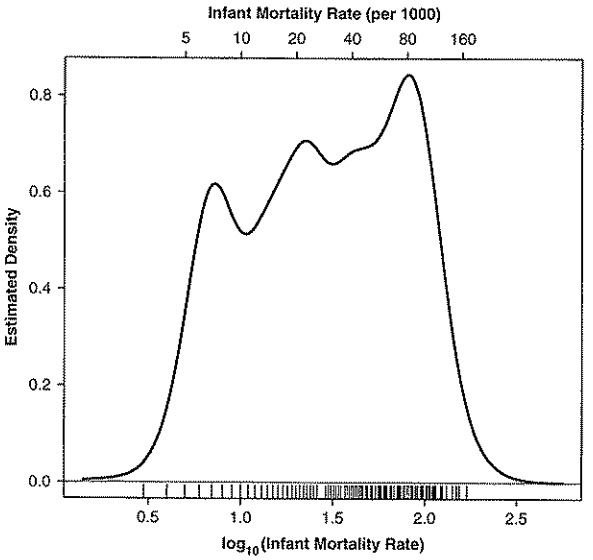

Figure 4.4 Adaptive-kernel density estimate for the distribution of $\log_{10}$ infant mortality. The window half-width for the adaptive-kernel estimator is $h = 0.1$ (on the $\log_{10}$ infant mortality scale). A rug-plot of the data values appears at the bottom of the graph and the original infant mortality scale at the top.

This table and the boxplots in Figure 4.3 suggest the log transformation of infant mortality, and the result of transforming the data is shown in Figure 4.4. Not only is the distribution much more symmetric than before, but three modes are clearly resolved (and there is the suggestion of a fourth); the modes at infant mortality rates of about 7 and 20 were not distinguishable in the untransformed data.

Note the untransformed scale for infant mortality at the top of the graph: These values, which are equally spaced on the log scale, represent doubling of infant mortality rates. This is, in my experience, an effective device for presenting the results of a statistical analysis in which the familiar scale of a variable is lost through a transformation.

Although it is not the case here, where the log transformation is clearly indicated, we often have a choice between transformations that perform roughly equally well. Although we should try to avoid distorting the data, we may prefer one transformation to another because of interpretability. I have already mentioned that the log transformation has a convenient multiplicative interpretation. In certain contexts, other transformations may have specific substantive meanings. Here are a few common examples: The inverse of the time (say, in hours) required to travel a given distance (a kilometer) is speed (kilometers per hour); the inverse of response latency (say, in milliseconds, as in a psychophysical experiment) is response frequency (responses per 1,000 seconds); the square root of a measure of area (say, in square meters) is a linear measure of size (in meters); and the cube of a linear measure of size (say in centimeters) can be interpreted as a volume (cubic centimeters).

We generally prefer interpretable transformations when variables are measured on familiar and meaningful scales. Conversely, because the rating "scales" that are ubiquitous in social research are not really measurements, there is typically no reason to prefer the original scores to a monotone transformation of them.[12]

> Descending the ladder of powers (e.g., to $\log X$) tends to correct a positive skew; ascending the ladder of powers (e.g., to X^2) tends to correct a negative skew.

4.3 Transforming Nonlinearity

Power transformations can also be used to make many nonlinear relationships more nearly linear. Again, we ask, why bother?

- Linear relationships—expressible in the form $\widehat{Y} = A + BX$—are particularly simple. Recall that this equation specifies that the average value of the response variable Y is a linear function of the explanatory variable X, with intercept A and slope B. Linearity implies that a unit *increase* in X—regardless of the *level* of X—is associated, on average, with a change of B units in Y.[13] Fitting a linear equation to data makes it relatively easy to answer certain questions about the data: If B is positive, for example, then Y tends to increase with X.
- Especially when there are several explanatory variables, the alternative of nonparametric regression may not be feasible because of the sparseness of the data. Even if we can fit a nonparametric regression with several Xs, it may be difficult to visualize the multidimensional result.[14]

[12] Rating scales are composed, for example, of items with response categories labeled *strongly agree*, *agree*, *disagree*, *strongly disagree*. A scale is constructed by assigning arbitrary numbers to the categories (e.g., 1–4) and adding or averaging the items. See Coombs, Dawes, and Tversky (1970, Chapters 2 and 3) for an elementary treatment of measurement issues in the social sciences and Duncan (1984) for an interesting account of the history and practice of social measurement. I believe that social scientists should pay more attention to measurement issues (employing, e.g., the methods of item-response theory; e.g., Baker & Kim, 2004). It is unproductive, however, simply to discard rating scales and similar "measurements by fiat" (a felicitous term borrowed from Torgerson, 1958): There is a prima facie reasonableness to many rating scales, and to refuse to use them without adequate substitutes would be foolish.

[13] I use the terms "increase" and "change" loosely here as a shorthand for static comparisons between average values of Y for X-values that differ by one unit: *Literal* change is not necessarily implied.

[14] See, however, the additive regression models discussed in Section 18.2.2, which overcome this deficiency.

- There is a simple and elegant statistical theory for linear models, which we explore in subsequent chapters. If these models are reasonable for the data, then their use is convenient.
- There are certain technical advantages to having linear relationships among the *explanatory* variables in a regression analysis.[15]

The following simple example suggests how a power transformation can serve to straighten a nonlinear relationship: Suppose that $Y = \frac{1}{5}X^2$ (with no residual) and that X takes on successive integer values between 1 and 5:

X	Y
1	0.2
2	0.8
3	1.8
4	3.2
5	5.0

These "data" are graphed in panel (a) of Figure 4.5, where the nonlinearity of the relationship between Y and X is apparent. Because of the manner in which the example was constructed, it is obvious that there are two simple ways to transform the data to achieve linearity:

1. We could replace Y by $Y' = \sqrt{Y}$, in which case $Y' = \sqrt{\frac{1}{5}}X$.
2. We could replace X by $X' = X^2$, in which case $Y = \frac{1}{5}X'$.

In either event, the relationship is rendered perfectly linear, as shown graphically in panels (b) and (c) of Figure 4.5. To achieve an intuitive understanding of this process, imagine that the original plot in panel (a) is drawn on a rubber sheet: Transforming Y "down" the ladder of powers to square root differentially stretches the rubber sheet vertically so that small values are spread out relative to large ones, stretching the curve in (a) into the straight line in (b). Likewise, transforming X "up" the ladder of powers spreads out the large values relative to the small ones, stretching the curve into the straight line in (c).

A power transformation works here because the relationship between Y and X is smooth, monotone (in this instance, strictly increasing), and simple. What I mean by "simple" in this context is that the direction of curvature of the function relating Y to X does not change (i.e., there is no point of inflection). Figure 4.6 seeks to clarify these distinctions: The relationship in panel (a) is simple and monotone; the relationship in panel (b) is monotone but not simple; and the relationship in panel (c) is simple but not monotone. I like to use the term "curvilinear" for cases such as (c), to distinguish nonmonotone from monotone nonlinearity, but this is not standard terminology. In panel (c), no power transformation of Y or X can straighten the relationship between them, but we could capture this relationship with a quadratic model of the form $\hat{Y} = A + B_1X + B_2X^2$.[16]

Like transformations to reduce skewness, a transformation to correct nonlinearity can be selected analytically or by guided trial and error.[17] Figure 4.7 introduces Mosteller and Tukey's (1977) "bulging rule" for selecting a transformation: If the "bulge" points *down* and to the *right*, for example, we need to transform Y *down* the ladder of powers or X *up* (or both). This case corresponds to the example in Figure 4.5, and the general justification of the rule follows from

[15]This point is developed in Section 12.3.3.

[16]Quadratic and other polynomial regression models are discussed in Section 17.1.

[17]See Sections 4.6 and 12.5 for analytic methods of selecting linearizing transformations.

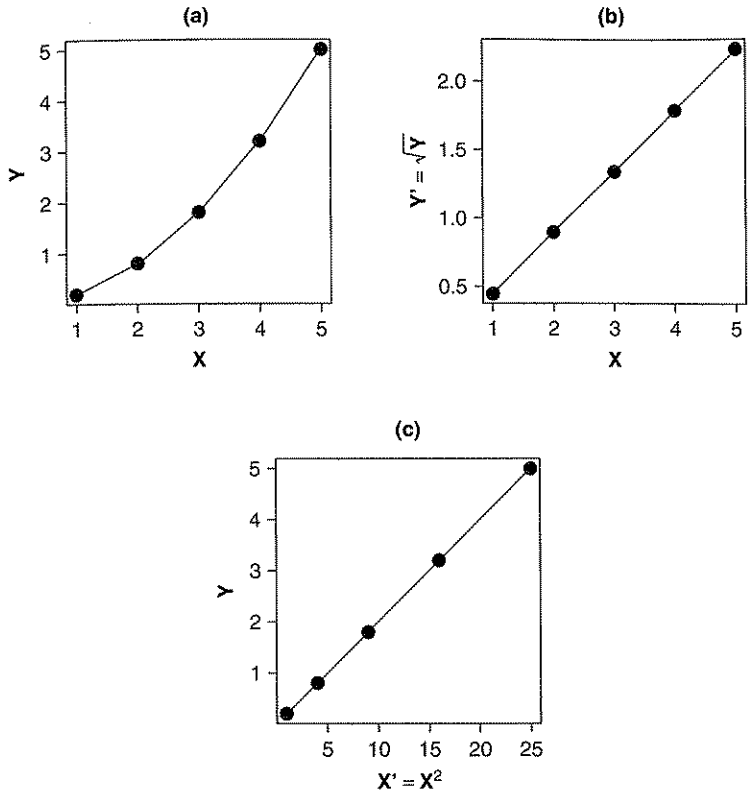

Figure 4.5 How a power transformation of Y or X can make a simple monotone nonlinear relationship linear. Panel (a) shows the relationship $Y = \frac{1}{5}X^2$. In panel (b), Y is replaced by the transformed value $Y' = Y^{1/2}$. In panel (c), X is replaced by the transformed value $X' = X^2$.

the need to stretch an axis differentially to transform the curve into a straight line. Trial and error is simplest with software that provides "sliders" for the power transformations of X and Y, immediately displaying the effect of a change in either power on the scatterplot relating the two variables, but we can in any event examine a series of scatterplots for different transformations.

> Simple monotone nonlinearity can often be corrected by a power transformation of X, of Y, or of both variables. Mosteller and Tukey's bulging rule assists in selecting linearizing transformations.

Now, we reexamine, in the light of this discussion, the relationship between prestige and income for the 102 Canadian occupations first encountered in Chapter 2 and shown in Figure 4.8.[18] The relationship between prestige and income is clearly monotone and nonlinear: Prestige rises with income, but the slope is steeper at the left of the plot, where income is low, than at the right, where it is high. The change in slope appears fairly abrupt rather than smooth, however, and we

[18] Repeating Figure 2.10.

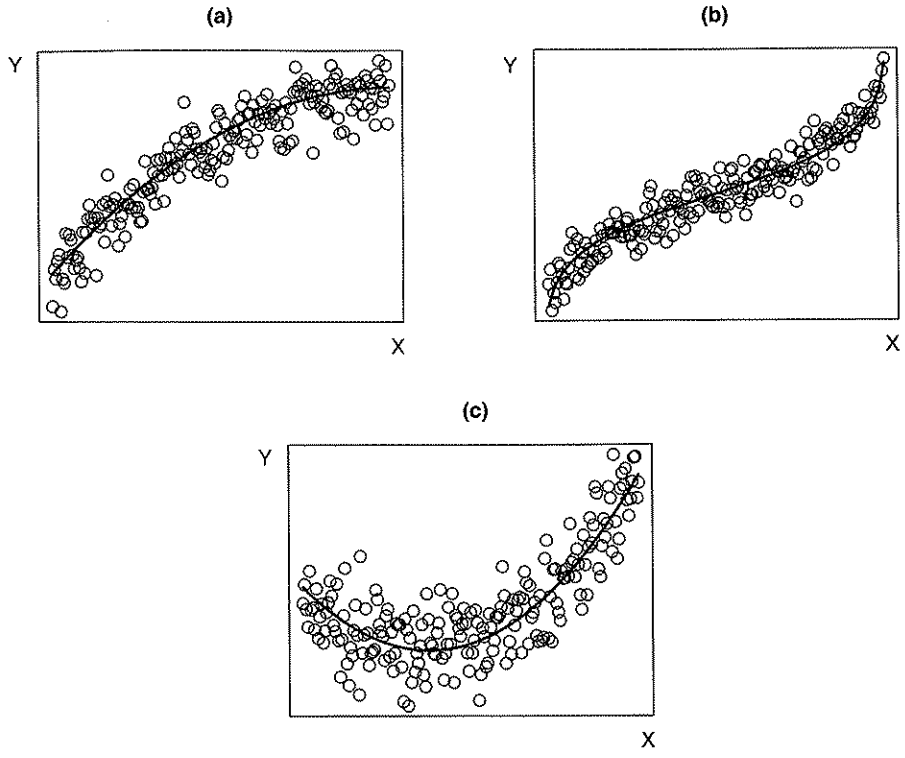

Figure 4.6 (a) A simple monotone relationship between *Y* and *X*; (b) a monotone relationship that is not simple; (c) a relationship that is simple but not monotone. A power transformation of *Y* or *X* can straighten (a) but not (b) or (c).

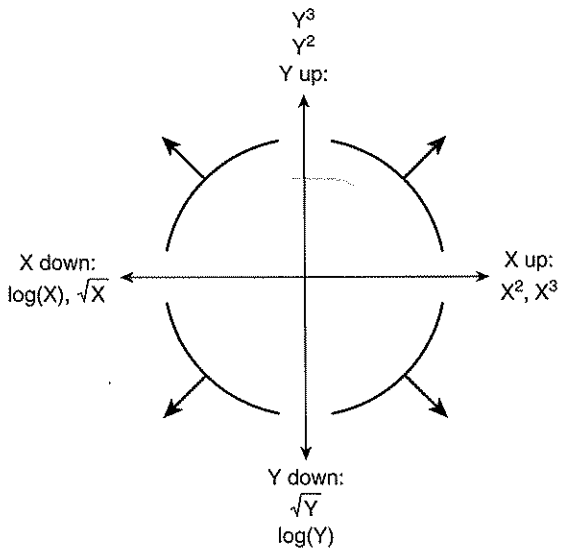

Figure 4.7 Tukey and Mosteller's bulging rule: The direction of the bulge indicates the direction of the power transformation of *Y* and/or *X* to straighten the relationship between them.

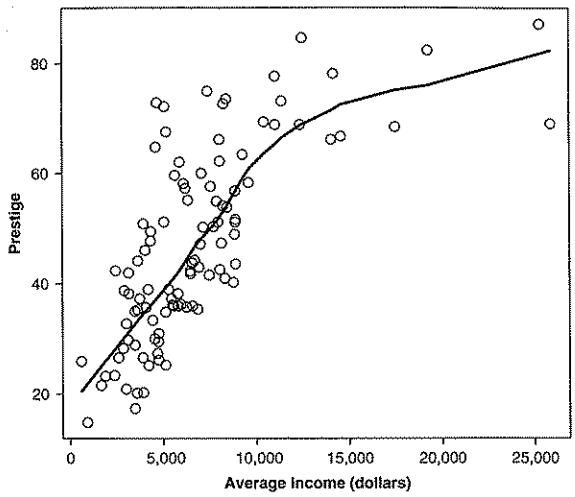

Figure 4.8 The relationship between prestige and income for the Canadian occupational prestige data. The nonparametric regression line on the plot is computed by lowess, with a span of 0.6.

might do better to model the relationship with two straight lines (one for relatively small values of income, one for relatively large ones) than simply to transform prestige or income.[19]

Nevertheless, the bulge points up and to the left, and so we can try transforming prestige up the ladder of powers or income down. Because the income distribution is positively skewed, I prefer to transform income rather than prestige, which is more symmetrically distributed. As shown in Figure 4.9, the cube-root transformation of income works reasonably well here. Some nonlinearity remains, but it is not simple, and the linear regression of prestige on income no longer *grossly* distorts the relationship between the two variables. I would have preferred to use the log transformation, which makes the income distribution more symmetric and which is simpler to interpret, but this transformation "overcorrects" the nonlinearity in the relationship between prestige and income.

For a more extreme, and ultimately more successful, example, consider the relationship between infant mortality and GDP per capita shown in Figure 4.10 and first discussed in Chapter 3.[20] As I pointed out previously, both variables are highly positively skewed and, consequently, most of the data are confined to a small region at the lower left of the plot.

The skewness of infant mortality and income in Figure 4.10 makes the scatterplot difficult to interpret; nevertheless, the nonparametric regression shown on the plot reveals a nonlinear but monotone relationship between infant mortality and income. The bulging rule suggests that infant mortality or income should be transformed down the ladder of powers and roots. In this case, transforming both variables by taking logs makes the relationship nearly linear (as shown in Figure 4.11). Moreover, although several countries still stand out as having relatively high infant mortality for their GDP, others now are revealed to have relatively *low* infant mortality in comparison to countries with similar GDP.

The least-squares regression line in Figure 4.11 has the equation

$$\log_{10} \widehat{\text{Infant mortality}} = 3.06 - 0.493 \times \log_{10} \text{GDP}$$

[19] For an alternative interpretation of the relationship between prestige and income, plot the data using different symbols for different types of occupations. (The data set distinguishes among blue-collar, white-collar, and professional and managerial occupations.)

[20] Repeating Figure 3.13. This example is motivated by a discussion of similar data in Leinhardt and Wasserman (1979).

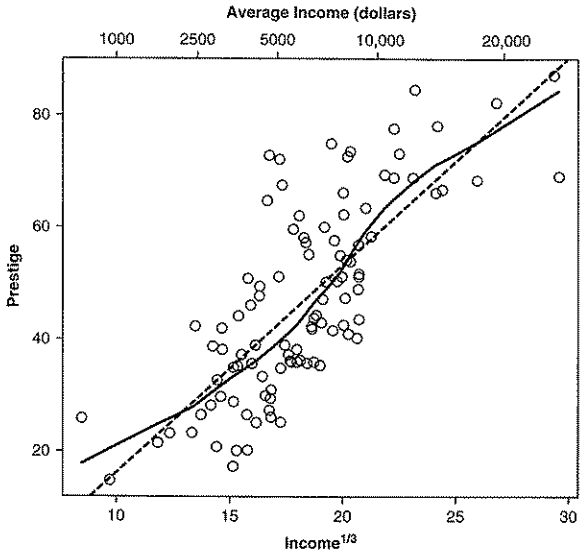

Figure 4.9 Scatterplot of prestige versus income$^{1/3}$. The broken line shows the linear least-squares regression, while the solid line shows the lowess smooth, with a span of 0.6. The original income scale is shown at the top of the graph.

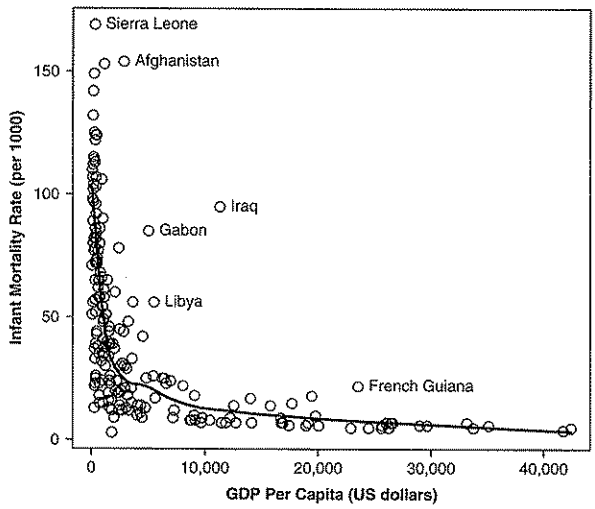

Figure 4.10 Scatterplot for infant mortality and GDP per capita for 193 nations. The line is for a lowess smooth with a span of 1/2. Several nations with high infant mortality for their levels of GDP are identified.

Because both variables are expressed on log scales to the same base, the slope of this relationship has a simple interpretation: A 1% increase in per-capita income is associated, on average, with an approximate 0.49% decline in the infant mortality rate. Economists call this type of coefficient an "elasticity."[21]

[21]Increasing X by 1% is equivalent to multiplying it by 1.01, which in turn implies that the log of X increases by $\log_{10} 1.01 = 0.00432$. The corresponding change in log Y is then $B \times 0.00432 = -0.493 \times 0.00432 = -0.00213$.

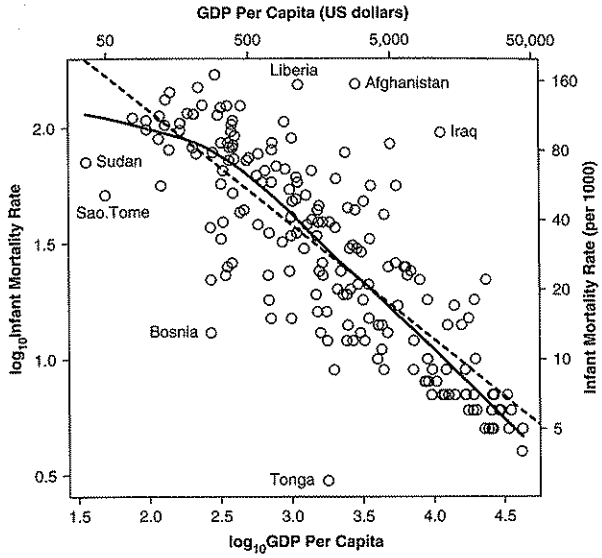

Figure 4.11 Scatterplot of $\log_{10}$ infant mortality rate versus $\log_{10}$ per-capita GDP. The broken line was calculated by linear least-squares linear regression and the solid line by lowess with a span of 1/2. The original scales of the variables appear at the top and to the right.

4.4 Transforming Nonconstant Spread

When a variable has very different degrees of variation in different groups, it becomes difficult to examine the data and to compare differences in level across the groups. We encountered this problem in the preceding chapter, where we compared the distribution of number of interlocking directorships by nation of control, employing Ornstein's data on 248 dominant Canadian corporations, shown in Figure 4.12.[22]

Differences in spread are often systematically related to differences in level: Groups with higher levels tend to have higher spreads. Using the median and hinge-spread as indices of level and spread, respectively, the following table shows that there is indeed an association, if only an imperfect one, between spread and level for Ornstein's data:

Nation of Control	*Lower Hinge*	*Median*	*Upper Hinge*	*Hinge Spread*
Other	3	14.5	23	20
Canada	5	12.0	29	24
United Kingdom	3	8.0	13	10
United States	1	5.0	12	11

Subtracting 0.00213 from $\log Y$ is equivalent to multiplying Y by $10^{-0.00213} = 0.99511$, that is, decreasing Y by $100 \times (1 - 0.99511) = 0.489 \approx B$. The approximation holds because the log function is nearly linear across the small domain of X-values between log 1 and log 1.01.

[22] Repeating Figure 3.16.

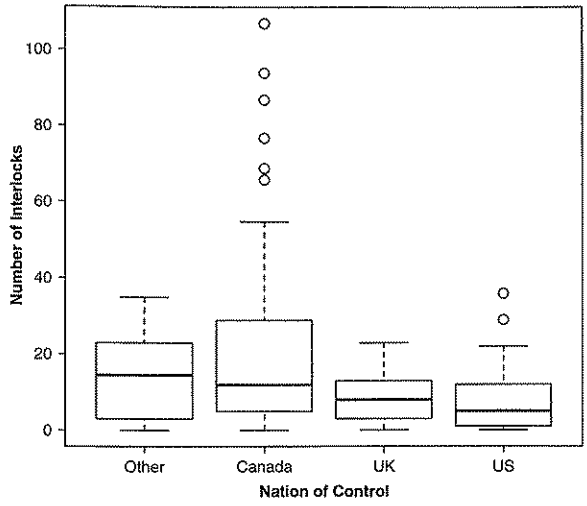

Figure 4.12 Number of interlocking directorate and executive positions by nation of control, for 248 dominant Canadian firms.

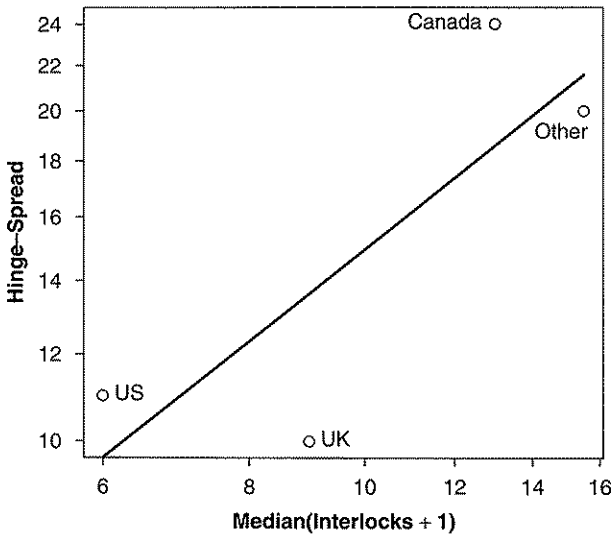

Figure 4.13 Spread (log hinge-spread) versus level [log(median+1)]. The plot is for Ornstein's interlocking-directorate data, with groups defined by nation of control. The line on the plot was fit by least squares.

Tukey (1977) suggests graphing the log hinge-spread against the log median, as shown in Figure 4.13. Because some firms maintained 0 interlocks, I used a start of 1 to construct this graph—adding 1 to each median but leaving the hinge-spreads unchanged.

The slope of the linear "trend," if any, in the spread-level plot can be used to suggest a spread-stabilizing power transformation of the data: Express the linear fit as

$$\log \text{spread} \approx a + b \log \text{level}$$

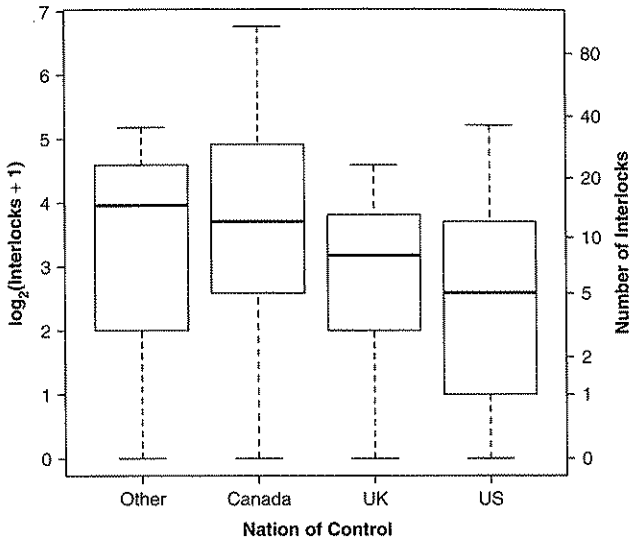

Figure 4.14 Parallel boxplots of number of interlocks by nation of control, transforming interlocks+1 to the $\log_2$ scale. Compare this plot with Figure 4.12, where number of interlocks is not transformed. The original scale for number of interlocks is shown at the right.

Then the corresponding spread-stabilizing transformation uses the power $p = 1 - b$. When spread is positively related to level (i.e., $b > 0$), therefore, we select a transformation *down* the ladder of powers and roots.

Starting with this transformation, it is convenient to employ statistical software that connects a "slider" for the power p to the spread-by-level plot and the parallel boxplots. Changing the value of p via the slider immediately updates the plots, allowing us to assess the relative effects of different transformations.

> When there is a positive association between the level of a variable in different groups and its spread, the spreads can be made more constant by descending the ladder of powers. A negative association between level and spread is less common but can be corrected by ascending the ladder of powers.

In Figure 4.13, a line was fit by least squares to the spread-level plot for the interlocking directorate data. The slope of this line, $b = 0.85$, suggests the power transformation $p = 1 - 0.85 = 0.15 \approx 0$. I decided, therefore, to try a log transformation. Figure 4.14 shows the result, employing logs to the base 2.[23] The spreads of the several groups are now much more similar, and differences in level are easier to discern. The within-group distributions are more symmetric as well.

The problems of unequal spread and skewness commonly occur together because they often have a common origin. When, as here, the data represent frequency counts (*number* of interlocks), the impossibility of obtaining a negative count tends to produce positive skewness, together

[23] Recall that increasing $\log_2 X$ by 1 represents doubling X (where, here, X is the number of interlocks plus 1).

```
1 | 2: represents 12
 leaf unit: 1
        n: 102

  32     0* | 00000000000000111111222233334444
  44     0. | 555566777899
 (8)     1* | 01111333
  50     1. | 5557779
  43     2* | 1344
  39     2. | 57
  37     3* | 01334
  32     3. | 99
         4* |
  30     4. | 678
  27     5* | 224
  24     5. | 67
  22     6* | 3
  21     6. | 789
  18     7* | 024
  15     7. | 5667
  11     8* | 233
         8. |
   8     9* | 012
```

Figure 4.15 Stem-and-leaf display of percentage of women in each of 102 Canadian occupations in 1970. Note how the data "stack up" against both boundaries.

with a tendency for larger levels to be associated with larger spreads. The same is true of other types of variables that are bounded below (e.g., wage and salary income). Likewise, variables that are bounded above but not below (e.g., grades on a very simple exam) tend both to be negatively skewed and to show a negative association between spread and level. In the latter event, a transformation "up" the ladder of powers (e.g., to X^2) usually provides a remedy.[24]

4.5 Transforming Proportions

Power transformations are often unhelpful for proportions because these quantities are bounded below by 0 and above by 1. Of course, if the data values do not approach the two boundaries, then proportions can be handled much like other sorts of data.

Percentages and many sorts of rates (e.g., infant mortality rate per 1,000 live births) are simply rescaled proportions and, therefore, are similarly affected. It is, moreover, common to encounter "disguised" proportions, such as the number of questions correct on an exam of fixed length or the number of affirmative responses to a series of dichotomous attitude questions.

An example, drawn from the Canadian occupational prestige data, is shown in the stem-and-leaf display in Figure 4.15. The distribution is for the percentage of women among the incumbents of each of 102 occupations. There are many occupations with no women or a very small percentage of women, but the distribution is not simply positively skewed, because there are also occupations that are predominantly female. In contrast, relatively few occupations are balanced with respect to their gender composition.

Several transformations are commonly employed for proportions, P, including the following:

- The *logit* transformation,

$$P \rightarrow \text{logit}(P) = \log_e \frac{P}{1-P}$$

[24] Plotting log spread against log level to select a spread-stabilizing transformation is quite a general idea. In Section 12.2, for example, we will use a version of the spread-level plot to find a variance-stabilizing transformation in regression analysis.

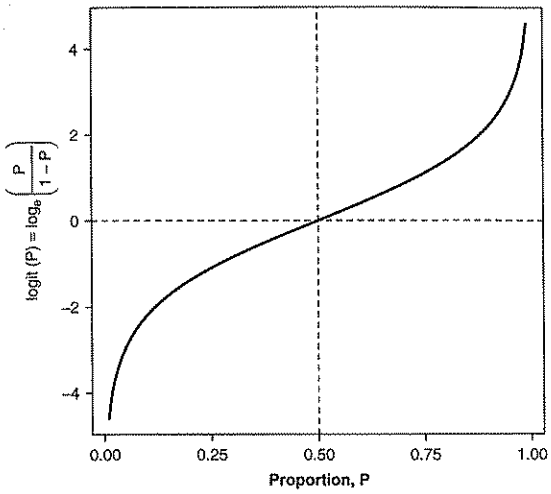

Figure 4.16 The logit transformation $\log_e[P/(1-P)]$ of a proportion P.

The logit transformation is the log of the "odds," $P/(1-P)$. The "trick" of the logit transformation is to remove the upper and lower boundaries of the scale, spreading out the tails of the distribution and making the resulting quantities symmetric about 0; for example,

P	$\frac{P}{1-P}$	*logit*
.01	1/99	−4.59
.05	1/19	−2.94
.1	1/9	−2.20
.3	3/7	−0.85
.5	1	0
.7	7/3	0.85
.9	9/1	2.20
.95	19/1	2.94
.99	99/1	4.59

A graph of the logit transformation, shown in Figure 4.16, reveals that the transformation is nearly linear in its center, between about $P = .2$ and $P = .8$.

- The *probit* transformation,

$$P \rightarrow \text{probit}(P) = \Phi^{-1}(P)$$

where Φ^{-1} is the inverse distribution function (i.e., the quantile function) for the standard normal distribution. Once their scales are equated, the logit and probit transformations are, for practical purposes, indistinguishable: logit $\approx (\pi/\sqrt{3}) \times$ probit.[25]

- The *arcsine-square-root* transformation also has a similar shape:

$$P \rightarrow \sin^{-1}\sqrt{P}$$

[25] We will encounter the logit and probit functions again in a different context when we take up the analysis of categorical data in Chapter 14.

Tukey (1977) has embedded these common transformations for proportions into the family of "folded" powers and roots, indexed by the power q, which takes on values between 0 and 1:

$$P \rightarrow P^q - (1 - P)^q$$

When $q = 0$, we take the natural log, producing the logit transformation. Setting $q = 0.14$ yields (to a very close approximation) a multiple of the probit transformation. Setting $q = 0.41$ produces (again, to a close approximation) a multiple of the arcsine-square-root transformation. When $q = 1$, the transformation is just twice the "plurality" (i.e., the difference between P and $\frac{1}{2}$), leaving the shape of the distribution of P unaltered:

$$P \rightarrow P - (1 - P) = 2(P - \tfrac{1}{2})$$

> Power transformations are ineffective for proportions P that simultaneously push the boundaries of 0 and 1 and for other variables (e.g., percentages, rates, disguised proportions) that are bounded both below and above. The folded powers $P \rightarrow P^q - (1 - P)^q$ are often effective in this context; for $q = 0$, we employ the logit transformation, $P \rightarrow \log_e[P/(1 - P)]$.

The logit and probit transformations cannot be applied to proportions of exactly 0 or 1. If, however, we have access to the original counts on which the proportions were based, then we can avoid this embarrassment by employing

$$P' = \frac{F + \frac{1}{2}}{N + 1}$$

in place of P. Here, F is the frequency count in the focal category (e.g., number of women) and N is the total count (total number of occupational incumbents, women plus men). If the original counts are not available, then we can use the expedient of mapping the proportions to an interval that excludes 0 and 1. For example, $P' = .005 + .99 \times P$ maps proportions to the interval [.005, .995].

Employing the latter strategy for the Canadian occupational data produces the distribution for $\text{logit}(P'_{\text{women}})$ that appears in Figure 4.17. Spreading out the tails of the distribution has improved its behavior considerably, although there is still some stacking up of low and high values.

4.6 Estimating Transformations as Parameters*

If we lived in a world in which the joint distribution of all quantitative data were multivariate-normal then statistical analysis would be simple indeed: Outliers would be rare, all variables would be symmetrically distributed, all regressions would be linear, and least-squares regression would be a fine method of estimation. Making data as close to multivariate-normal as possible by transformation, therefore, can facilitate their analysis.

If the vector random variable $\mathbf{x} = (X_1, X_2, \ldots, X_p)'$ with population mean vector $\underset{(p\times 1)}{\boldsymbol{\mu}}$ and covariance matrix $\underset{(p\times p)}{\boldsymbol{\Sigma}}$ is multivariate-normal, then its probability density function is[26]

[26] See Appendix D on probability and estimation.

```
1 | 2: represents 1.2
 leaf unit: 0.1
        n: 102

   5     -5* | 22222
   8     -4. | 555
  16     -4* | 44332111
  21     -3. | 98875
  31     -3* | 4432111000
  39     -2. | 98887655
  48     -2* | 443220000
 (10)    -1. | 9888666555
  44     -1* | 331110
  38     -0. | 987666
  32     -0* | 44110
  27      0* | 00122
  22      0. | 577889
  16      1* | 01111
  11      1. | 556
   8      2* | 23
   6      2. | 5
   5      3* | 00014
```

Figure 4.17 Stem-and-leaf display for the logit transformation of proportion of women in each of 102 Canadian occupations. Because some occupations have no women, the proportions were mapped to the interval .005 to .995 prior to calculating the logits.

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

In shorthand, $\mathbf{x} \sim \mathbf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For a sample of n observations, $\underset{(n\times p)}{\mathbf{X}}$, we have

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left[\frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Sigma}}}\right]^n \exp\left\{\sum_{i=1}^{n}\left[-\tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]\right\}$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. The log-likelihood for the parameters is, therefore,[27]

$$\log_e L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = -\frac{np}{2}\log_e(2\pi) - \frac{n}{2}\log_e \det \boldsymbol{\Sigma} - \tfrac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]$$

The maximum-likelihood estimators (MLEs) of the mean and covariance matrix are, then,[28]

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}} = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)'$$

$$\widehat{\boldsymbol{\Sigma}} = \{\widehat{\sigma}_{jj'}\} = \left\{\frac{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ij'} - \overline{X}_{j'})}{n}\right\}$$

Now, suppose that $\mathbf{x}$ is not multivariate-normal, but that it can be made so by a power transformation of its elements.[29] It is convenient to use the Box-Cox family of power transformations

[27] The likelihood function and maximum-likelihood estimation are described in Appendix D on probability and estimation.

[28] Note that the MLEs of the covariances have n rather than $n - 1$ in the denominator, and consequently will be biased in small samples.

[29] This cannot be strictly correct, because Box-Cox transformations are only applicable when the elements of $\mathbf{x}$ are positive and normal distributions are unbounded, but it can be true to a close-enough approximation. There is no guarantee, however, that $\mathbf{x}$ can be made normal by a power transformation of its elements.

(Equation 4.2) because they are continuous at $p = 0$. Rather than thinking about these powers informally, let us instead consider them as additional parameters,[30] $\boldsymbol{\lambda} \equiv (\lambda_1, \lambda_2, \ldots, \lambda_p)'$, one for each element of $\mathbf{x}$, so that

$$\mathbf{x}^{(\boldsymbol{\lambda})} \equiv \left[x_1^{(\lambda_1)}, x_2^{(\lambda_2)}, \ldots, x_p^{(\lambda_p)}\right]'$$

Then,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\tfrac{1}{2}(\mathbf{x}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})\right] \prod_{j=1}^{p} X_j^{\lambda_j - 1} \tag{4.3}$$

where now $\boldsymbol{\mu} = E\left[\mathbf{x}^{(\boldsymbol{\lambda})}\right]$ and $\boldsymbol{\Sigma} = V\left[\mathbf{x}^{(\boldsymbol{\lambda})}\right]$ are the mean vector and covariance matrix of the transformed variables, and $\prod_{j=1}^{p} X_j^{\lambda_j - 1}$ is the Jacobian of the transformation from $\mathbf{x}^{(\boldsymbol{\lambda})}$ to $\mathbf{x}$.[31]

The log-likelihood for the model is

$$\log_e L(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = -\frac{np}{2}\log_e(2\pi) - \frac{n}{2}\log_e \det \boldsymbol{\Sigma} - \tfrac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})\right]$$
$$+ \sum_{j=1}^{p}(\lambda_j - 1)\sum_{i=1}^{n}\log_e X_{ij}$$

There is no closed-form solution for the MLEs of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$, but we can find the MLEs by numerical methods. Standard errors for the estimated transformations are available in the usual manner from the inverse of the information matrix, and both Wald and likelihood-ratio tests can be formulated for the transformation parameters.

Moreover, because our real interest lies in the transformation parameters $\boldsymbol{\lambda}$, the means $\boldsymbol{\mu}$ and covariances $\boldsymbol{\Sigma}$ are "nuisance" parameters; indeed, given $\widehat{\boldsymbol{\lambda}}$, the MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are just the sample mean vector and covariance matrix of $\mathbf{x}^{(\widehat{\boldsymbol{\lambda}})}$. Let us define the *modified Box-Cox family* of transformations as follows:

$$X^{[\lambda]} = \begin{cases} \widetilde{X}^{1-\lambda}\dfrac{X^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \widetilde{X}\log_e X & \text{for } \lambda = 0 \end{cases}$$

where

$$\widetilde{X} \equiv \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$

is the *geometric mean* of X. Multiplication by $\widetilde{X}^{1-\lambda}$ is a kind of standardization, equating the scales of different power transformations of X. Let $\mathbf{V}^{[\boldsymbol{\lambda}]}$ represent the sample covariance matrix of

$$\mathbf{x}^{[\boldsymbol{\lambda}]} \equiv \left[x_1^{[\lambda_1]}, x_2^{[\lambda_2]}, \ldots, x_p^{[\lambda_p]}\right]'$$

Velilla (1993) shows that the the MLEs of $\boldsymbol{\lambda}$ in Equation 4.3 are the values that minimize the determinant of $\mathbf{V}^{[\boldsymbol{\lambda}]}$.

[30] We will encounter this general approach again in Section 12.5 in the context of the linear regression model.

[31] See Appendix D on probability and estimation.

Applying this approach to the joint distribution of infant mortality and GDP per capita produces the following results:

	$\widehat{\lambda}_j$	$\mathrm{SE}(\widehat{\lambda}_j)$	$z_0 = \dfrac{\widehat{\lambda}_j - 1}{SE(\widehat{\lambda}_j)}$	p
Infant mortality	−0.0009	0.0655	−15.28	$\ll .0001$
GDP per capita	0.0456	0.0365	−26.14	$\ll .0001$

The first column in this table gives the the MLE of each transformation parameter; the second column gives the asymptotic standard error of the transformation; the third column gives the Wald statistic for testing the hypothesis H_0: $\lambda_j = 1$ (i.e., that no transformation is required); and the final column gives the two-sided p-value for this test. In this case, evidence for the need to transform the two variables is very strong. Moreover, both estimated transformations are very close to 0—that is, the log transformation. A likelihood-ratio test for the hypothesis H_0: $\lambda_1 = \lambda_2 = 1$ yields the chi-square test statistic $G_0^2 = 680.25$ on 2 degrees of freedom, which is also wildly statistically significant. In contrast, testing the hypothesis that H_0: $\lambda_1 = \lambda_2 = 0$ produces $G_0^2 = 1.649$ on 2 degrees of freedom, for which $p = .44$, supporting the use of the log transformations of infant mortality and GDP. We know from our previous work that these transformations make the distribution of the two variables symmetric and linearize their relationship.

Finally, we can also apply this method to individual variables to attempt to normalize their *univariate* distributions. For the current example, the individual MLEs of the power transformation parameters λ for infant mortality and GDP are similar to those reported above:

	$\widehat{\lambda}$	$\mathrm{SE}(\widehat{\lambda})$	$z_0 = \dfrac{\widehat{\lambda} - 1}{SE(\widehat{\lambda})}$	p
Infant mortality	0.0984	0.0786	−11.47	$\ll .0001$
GDP per capita	−0.0115	0.0440	−23.00	$\ll .0001$

> The method of maximum likelihood can be used to estimate normalizing power transformations of variables.

Exercises

Exercise 4.1. Create a graph like Figure 4.1, but for the *ordinary* power transformations $X \to X^p$ for $p = -1, 0, 1, 2, 3$. (When $p = 0$, however, use the log transformation.) Compare your graph to Figure 4.1, and comment on the similarities and differences between the two families of transformations X^p and $X^{(p)}$.

Exercise 4.2. *Show that the derivative of $f(X) = (X^p - 1)/p$ is equal to 1 at $X = 1$ regardless of the value of p.

Exercise 4.3. *We considered starts for transformations informally to ensure that all data values are positive and that the ratio of the largest to the smallest data values is sufficiently large. An

alternative is to think of the start as a parameter to be estimated along with the transformation power to make the distribution of the variable as normal as possible. This approach defines a *two-parameter Box-Cox family*:

$$X^{(\alpha,\lambda)} \equiv \frac{(X-\alpha)^{\lambda}}{\lambda}$$

(a) Develop the MLEs of α and λ for the two-parameter Box-Cox family.
(b) Attempt to apply the estimator to data. Do you encounter any obstacles? [*Hint*: Examine the correlation between the parameter estimates $\widehat{\alpha}$ and $\widehat{\lambda}$.]

Summary

- Transformations can often facilitate the examination and statistical modeling of data.
- The powers and roots are a particularly useful family of transformations: $X \to X^p$. When $p = 0$, we employ the log transformation in place of X^0.
- Power transformations preserve the order of the data only when all values are positive and are effective only when the ratio of largest to smallest data values is itself large. When these conditions do not hold, we can impose them by adding a positive or negative start to all the data values.
- Descending the ladder of powers (e.g., to $\log X$) tends to correct a positive skew; ascending the ladder of powers (e.g., to X^2) tends to correct a negative skew.
- Simple monotone nonlinearity can often be corrected by a power transformation of X, of Y, or of both variables. Mosteller and Tukey's bulging rule assists in selecting linearizing transformations.
- When there is a positive association between the level of a variable in different groups and its spread, the spreads can be made more constant by descending the ladder of powers. A negative association between level and spread is less common but can be corrected by ascending the ladder of powers.
- Power transformations are ineffective for proportions, P, that simultaneously push the boundaries of 0 and 1 and for other variables (e.g., percentages, rates, disguised proportions) that are bounded both below and above. The folded powers $P \to P^q - (1-P)^q$ are often effective in this context; for $q = 0$, we employ the logit transformation, $P \to \log_e[P/(1-P)]$.
- The method of maximum likelihood can be used to estimate normalizing power transformations of variables.

Recommended Reading

Because examination and transformation of data are closely related topics, most of the readings here were also listed at the end of the previous chapter.

- Tukey's important text on exploratory data analysis (Tukey, 1977) and the companion volume by Mosteller and Tukey (1977) on regression analysis have a great deal of interesting information and many examples. As mentioned in the previous chapter, however, Tukey's writing style is opaque. Velleman and Hoaglin (1981) is easier to digest, but it is not as rich in material on transformations.

- Several papers in a volume edited by Hoaglin, Mosteller, and Tukey (1983) have valuable material on the family of power transformations, including a general paper by Emerson and Stoto, an extended discussion of the spread-versus-level plot in a paper on boxplots by Emerson and Strenio, and a more difficult paper by Emerson on the mathematics of transformations.
- The tools provided by the Lisp-Stat statistical computing environment (described in Tierney, 1990)—including the ability to associate a transformation with a slider and to link different plots—are especially helpful in selecting transformations. Cook and Weisberg (1994, 1999) have developed a system for data analysis and regression based on Lisp-Stat that includes these capabilities. Similar facilities are built into some statistical packages and can be implemented in other statistical computing environments (such as R).