# CHAPTER 13

# Transformation of the Response Variable

## 13.1. INTRODUCTION AND PRELIMINARY REMARKS

### Simplifying Models Via Transformation

Suppose we had $n$ data values $(f_i, p_i, Y_i)$, $i = 1, 2, \ldots, n$, that could be well explained by a quadratic equation, say,

$$Y = \beta_0 + \beta_1 f + \beta_2 p + \beta_{11} f^2 + \beta_{22} p^2 + \beta_{12} fp + \epsilon. \tag{13.1.1}$$

We might be perfectly happy with such a fit. Later, however, we might be told that it was customary with this particular type of data to use $\ln Y$, the natural[1] logarithm of $Y$ (logarithm to the base $e$, or ln) instead of $Y$. Armed with this knowledge, we might find that a simpler planar equation fit of

$$\ln Y = \beta_0 + \beta_1 f + \beta_2 p + \epsilon \tag{13.1.2}$$

gave a better, as good, or almost as good an explanation of the variation in the data. Essentially we would have used the *response transformation* $\ln Y$ to "flatten out" our original six-parameter quadratic surface to be a simpler three-parameter plane. We would also have changed the assumption about the error structure. If the errors in Eq. (13.1.1) were independent $N(0, \sigma^2)$ errors, the errors in Eq. (13.1.2) would not be; and vice versa. So we must give some thought to what is being assumed about the error structure when we transform a response.

### Thinking About the Error Structure

It is sometimes reasonable to believe that a model function might be multiplicative rather than additive. Suppose we think that

$$\eta = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} \tag{13.1.3}$$

is a sensible model function for a certain set of data. Let us take natural logarithms in (13.1.3). Then

[1] The difference between taking natural logarithms or logarithms to another base (10, say) is a constant multiple. Suppose a given number $A = e^b = 10^c$, say. Then $\ln A = b$ and $\log_{10} A = c$ and their ratio $b/c = \ln 10 = (\log_{10} e)^{-1} = 2.302585$.

$$\ln \eta = \ln \alpha + \beta \ln X_1 + \gamma \ln X_2 + \delta \ln X_3. \tag{13.1.4}$$

One would then be led to fitting by least squares the model

$$\ln Y = \ln \eta + \epsilon, \tag{13.1.5}$$

where $\ln \eta$ is given by (13.1.4). In doing this, we would assume that the errors are $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$. Now let us work backward from (13.1.5). By noting that $\epsilon = \ln (e^{\epsilon})$, where $e$ here is the natural logarithm base 2.718282, and exponentiating (13.1.5), we get

$$Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} e^{\epsilon} = \eta e^{\epsilon}. \tag{13.1.6}$$

This model does not have *additive* errors (i.e., we do not have $Y = \eta$ + error) but *multiplicative* ones; the model function is *multiplied* by the error. Thus a fit of (13.1.5) is appropriate only if we "believe" that (13.1.6) is a suitable model. Transforming the $Y$ into $\ln Y$ has altered the error structure. If we really believed that the errors were additive and that

$$Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} + \text{error}, \tag{13.1.7}$$

we could not take logarithms and use least squares. We would have to use the methods of nonlinear estimation instead. [However, (13.1.5) could then be fitted to give some *initial estimates*; see Chapter 24.] We say that (13.1.7) is *intrinsically nonlinear*, whereas (13.1.6) is *intrinsically linear*.

To take another example, if we decided to fit the model

$$\frac{1}{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{13.1.8}$$

then we would "believe" that an appropriate model was

$$Y = 1/(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon). \tag{13.1.9}$$

Now, in fact, most people do not think about models in this sequence. In practice, it is simpler to decide on a transformation, fit it, and then examine the residuals *in the metric of the transformed variable* to see if they are reasonably well behaved. If they are, the error specifications in the transformed response space are assumed to be all right. Note that, for (13.1.5), the residuals to examine are of the form $\ln Y_i - (\widehat{\ln Y})_i$, while for (13.1.8) they are $Y_i^{-1} - (\widehat{Y^{-1}})_i$. All tests and confidence statements must be made in the transformed space also.

### Predictions in *Y*-Space

Some scientists are unhappy about working in a transformed $Y$-space, but a return to the original $Y$-space can be made after the model has been fitted. Suppose we have fitted a model to $\ln Y$, and we make a prediction $\widehat{\ln Y}$ at a certain set of $X$'s. We can, if we wish, evaluate $\hat{Y} = \exp\{\widehat{\ln Y}\}$ and predict in the original space. Also, a confidence statement on $E(\ln Y)$ with interval $(a, b)$ can be translated into a confidence statement with interval $(e^a, e^b)$ in the $Y$-space. It will not be symmetric about the predicted value $\hat{Y}$, of course. We can also evaluate residuals $Y_i - \hat{Y}_i$ at the data points, if we wish. These residuals are *not*, however, checked; these are *not* the residuals that should satisfy the residuals checks for normality, and so on.

**TABLE 13.1. Values of Certain Power Functions for Five Benchmark Powers**

| $\lambda$ | $Y^\lambda$ | $W = (Y^\lambda - 1)/\lambda$ | $V = (Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1})$ |
|---|---|---|---|
| 1 | $Y$ | $Y - 1$ | $Y - 1$ |
| $\frac{1}{2}$ | $Y^{1/2}$ | $2(Y^{1/2} - 1)$ | $2\dot{Y}^{1/2}(Y^{1/2} - 1)$ |
| 0 | 1(?) | $\ln Y$ | $\dot{Y} \ln Y$ |
| $-\frac{1}{2}$ | $Y^{-1/2}$ | $2(1 - Y^{-1/2})$ | $2\dot{Y}^{3/2}(1 - Y^{-1/2})$ |
| $-1$ | $Y^{-1}$ | $1 - Y^{-1}$ | $\dot{Y}^2(1 - Y^{-1})$ |

## Preliminary Remarks on the Power Family of Transformations

One extremely useful way of picking a transformation is to assume that a member of the power family will be appropriate, and then to estimate the best power by maximum likelihood. This is often called the "Box–Cox method" in honor of the authors of the seminal paper on this topic, written in 1964. We describe this in the next section. There is a particular difficulty in thinking about powers $Y^\lambda$, because as $\lambda$ approaches zero, $Y^\lambda$ approaches 1. This would clearly be a senseless transformation! We shall see soon that a zero power is associated with a ln $Y$ (or log $Y$) transformation. To make the calculations for choosing the best $\lambda$ value run smoothly as $\lambda$ approaches zero, we must perform the Box–Cox calculations using not $Y^\lambda$, which gives problems at $\lambda = 0$, but with either $W = (Y^\lambda - 1)/\lambda$, now out of fashion, or (better) $V = W/\dot{Y}^{\lambda-1} = (Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1})$. ($\dot{Y}$ is the geometric mean of the $Y_i$ in the data set.) Note that for a power transformation to be applicable, all the $Y$'s must be positive. Table 13.1 will help prepare the reader for the fuller discussion of Section 13.2. It shows what the functions $Y^\lambda$, $W = (Y^\lambda - 1)/\lambda$, and $V = W/\dot{Y}^{\lambda-1}$ look like for five benchmark values of $\lambda$, namely, $\lambda = 1, \frac{1}{2}, 0, -\frac{1}{2}, -1$. The query next to the one in the second column of Table 13.1 denotes bewilderment at the possibility of a transformation $Y^\lambda$ when $\lambda = 0$, because all the data would revert to 1's. However, when $\lambda$ approaches zero, $W$ approaches ln $Y$ and $V$ thus approaches $\dot{Y}$ ln $Y$.

## Points to Keep in Mind

In general, when we make a transformation, it is impossible to relate the parameters of the model used for the transformed data to the parameters in a model initially intended for the untransformed data. Usually, there is no mathematical equivalence except in an approximate sense via a Taylor series expansion. For example, if instead of fitting $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$ we fit $Y^\lambda = \alpha_0 + \alpha_1 X + \epsilon$, the relationship of $\beta_0, \beta_1, \beta_{11}$ to $\lambda, \alpha_0, \alpha_1$ is not clear. An attempt to find such a relationship is usually not fruitful.

When several sets of data arise from similar experimental situations, it may not be necessary to carry out complete analyses on all the sets to determine appropriate transformations. Quite often, the same transformation will work for all.

The fact that a general analysis exists for finding transformations does not mean that it should always be used. Often, informal plots of the data will clearly reveal the need for a transformation of an obvious kind (such as ln $Y$ or $1/Y$). In such a case, the more formal analysis may be viewed as a useful check procedure to hold in reserve.

## 13.2 POWER FAMILY OF TRANSFORMATIONS ON THE RESPONSE: BOX–COX METHOD

Suppose we have data $(Y_1, Y_2, \ldots, Y_n)$ on a response variable $Y$ that is always positive. (Other cases will be discussed later.) If the ratio of the largest observed $Y$ to the smallest is "considerable," say, 10 or higher, we might consider the possibility of transforming $Y$. There are many possible types of transformations. A useful idea in many applications is to consider powers, $Y^\lambda$, say, and to try to find the best value of $\lambda$ to use. A snag soon becomes apparent; when $\lambda = 0$, $Y^0 = 1$—making all the data equal! However, if we were to try working with

$$W = \begin{cases} (Y^\lambda - 1)/\lambda, & \text{for } \lambda \neq 0, \\ \ln Y, & \text{for } \lambda = 0, \end{cases} \tag{13.2.1}$$

the problem at $\lambda = 0$ would be overcome, because $\ln Y$ is the appropriate limit, as $\lambda$ tends to zero, of $(Y^\lambda - 1)/\lambda$, and so the family is now continuous in $\lambda$. A disadvantage of (13.2.1) is that, as $\lambda$ varies, the sizes of the $W$'s can change enormously, leading to minor problems in the analysis and requiring a special program to get the best $\lambda$ value. For that reason, it is preferable to use the alternative form

$$V = \begin{cases} (Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1}), & \text{for } \lambda \neq 0, \\ \dot{Y} \ln Y, & \text{for } \lambda = 0, \end{cases} \tag{13.2.2}$$

where the additional divisor $(\dot{Y}^{\lambda-1})$ in (13.2.2), compared with (13.2.1), is the $n$th power of the appropriate *Jacobian* of the transformation, which converts the set of $Y_i$ into the set of $W_i$. This ensures that unit volume is preserved in moving from the set of $Y_i$ to the set of $V_i$ in (13.2.2). (To appreciate this remark fully, some knowledge of calculus is needed, but we can proceed without that.)

The quantity $\dot{Y}$ is the geometric mean of the $Y_i$,

$$\dot{Y} = (Y_1 Y_2 \cdots Y_n)^{1/n}. \tag{13.2.3}$$

$\dot{Y}$ is a constant and it would be evaluated at the beginning of the calculation procedure, usually by antilogging (exponentiating) the formula

$$\ln \dot{Y} = n^{-1} \sum_{i=1}^{n} \ln Y_i. \tag{13.2.4}$$

When formula (13.2.2) is applied to each $Y_i$ we create a vector $\mathbf{V} = (V_1, V_2, \ldots, V_n)'$ and use it to fit a linear model

$$\mathbf{V} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{13.2.5}$$

by least squares for any specified value of $\lambda$. More generally, we have to estimate $\lambda$ as well as $\boldsymbol{\beta}$. We do this by invoking the principle of maximum likelihood under the assumption that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\,\sigma^2)$ for the proper choice of $\lambda$. This method (and also a Bayesian equivalent of it), applicable to any family of transformations, including the one above, is discussed by Box and Cox (1964). The basic idea is that, if an appropriate $\lambda$ could be found, an additive model with normally distributed, independent, and homogeneous error structure could be fitted by the maximum likelihood method. We do not have to understand maximum likelihood, Bayesian statistics, or the Jacobian to perform the procedure. The necessary steps are as follows.

### Maximum Likelihood Method of Estimating $\lambda$

1. Choose a value of $\lambda$ from a selected range. Usually we look at $\lambda$'s in the range $(-1, 1)$, or perhaps even $(-2, 2)$, at first, and extend the range later if necessary. We would usually cover the selected range with about 11–21 values of $\lambda$. We can always divide up a portion of the interval more finely later if we need the additional detail, but this is often unnecessary—see (3) below.

2. For each chosen $\lambda$ value, evaluate $\mathbf{V}$ via (13.2.2). Remember to use $V = \dot{Y} \ln Y$ when $\lambda = 0$. Or else avoid using $\lambda = 0$ *exactly*, in covering the selected range of $\lambda$. Now fit (13.2.5) and record $S(\lambda, \mathbf{V})$, the residual sum of squares for the regression. Any ordinary least squares regression program can be used for this calculation.

3. Plot $S(\lambda, \mathbf{V})$ versus $\lambda$. [Some workers prefer to plot $\ln S(\lambda, \mathbf{V})$ versus $\lambda$; make your own choice depending on how big the numbers are.] Draw a smooth curve through the plotted points, and find at what value of $\lambda$ the lowest point of the curve lies. That value, $\hat{\lambda}$, is the maximum likelihood estimate of $\lambda$. Typically, we would not use this precise value of $\lambda$ in subsequent calculations, but would use instead the nearest convenient value in the sequence, $\ldots, -2, -1\frac{1}{2}, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 1\frac{1}{2}, 2, \ldots$ after first checking that such a value lay within a selected confidence interval (see below). For example, if $\hat{\lambda}$ came out to be about 0.11, we would probably use $\lambda = 0$, If $\hat{\lambda}$ were about 0.94, we would use $\lambda = 1$, and so on. (There is, however, considerable leeway for a personal decision in the choice of $\lambda$, after the calculations have been examined. In some situations, the values $\frac{1}{3}, \frac{2}{3}$ might be appropriate. Some workers prefer to round to the nearest quarter, rather than the nearest half; others feel unhappy with any rounding and proceed using $\hat{\lambda}$ instead.) We then analyze the transformed data—transformed via whatever value of $\lambda$ was finally selected—and report the results.

### Some Conversations on How to Proceed

The last sentence of (3) needs some further explanation. Once we have chosen a $\lambda$, how do we actually transform the data? Do we use the form (13.2.2) exactly as it is given but with the selected $\lambda$? We can do this if we wish. Alternatively, if a nonzero $\lambda$ is selected to transform the data, we can carry out our analysis on $Y^{\lambda}$ if we wish, rather than on the first line of (13.2.2). Similarly, if $\lambda = 0$ is the value of $\lambda$ actually chosen to transform the data, we can use either $\ln Y$ (natural logarithms) or $\log Y$ (logarithms to any other base, such as 10, for example). These logarithms differ only by a constant factor, and so only the scale of the numbers involved is affected, not the basic nature of the subsequent analysis. Most people would choose the simplest representation possible ($Y^{\lambda}$ or $\ln Y$). We do that in our example. Equation (13.2.2) is then used only for the analysis that determines $\lambda$. This form (13.2.2) has several advantages for this purpose. It is conceptually simple (see Box and Cox, 1964, p. 216), it provides better computational accuracy, especially for large $\lambda$, and the calculations can be performed using any standard regression program. Also, it allows direct comparison of the residual sums of squares, because the scale factor divisor $\dot{Y}^{\lambda-1}$ essentially reconverts the $W_i$ back to comparable units.

Use of the (13.2.2) form for the final analysis is also acceptable. Only a scale difference and an origin shift are involved and the basic nature of the subsequent analysis is unaffected by these, for a linear model.

Two points relevant to the regression analysis after choice of $\lambda$ should be noted:

1. The fact that the "best $\lambda$" has been selected does not necessarily guarantee an

equation useful in practice. The final equation must be evaluated in the usual ways on its own merits.

2. To allow for the fact that $\lambda$ has been estimated, some workers remove one df for $\hat{\lambda}$ in the analysis of variance table in the subsequent regression analysis. This reduces the total degrees of freedom from $n$ to $(n - 1)$ and the residual df are adjusted accordingly. This reduction is optional. (Note that no sum of squares is removed. Although we do not make this reduction in our examples below, we are not opposed to it. For large $n$, it makes little difference, of course.)

### Approximate Confidence Interval for $\lambda$

The maximum likelihood equations, after simplification, result in an estimation of $\lambda$ by choosing the $\lambda$ that minimizes the residual sum of squares function $S(\lambda, \mathbf{V})$. One step back from this point is the equivalent criterion: maximize $L(\lambda)$, where

$$L(\lambda) = -\tfrac{1}{2}n \ln\{S(\lambda, \mathbf{V})/n\}. \tag{13.2.6}$$

Obviously we do not need to plot this form because $n$ is fixed; that is why the simpler plot of $S(\lambda, \mathbf{V})$ versus $\lambda$ is used. However, an approximate $100(1 - \alpha)\%$ confidence interval for $\lambda$ consists of those values of $\lambda$ that satisfy the inequality

$$L(\hat{\lambda}) - L(\lambda) \leq \tfrac{1}{2}\chi_1^2(1 - \alpha), \tag{13.2.7}$$

where $\chi_1^2(1 - \alpha)$ is the percentage point of the chi-squared distribution with one degree of freedom, which leaves an area of $\alpha$ in the upper tail of the distribution. Some of these values are as follows:

| $\alpha$ | 0.10 | 0.05 | 0.025 | 0.01 | 0.001 |
|---|---|---|---|---|---|
| $\chi_1^2(1 - \alpha)$ | 2.71 | 3.84 | 5.02 | 6.63 | 10.83 |

(13.2.8)

To implement Eq. (13.2.7), we could draw, on a plot of $L(\lambda)$ versus $\lambda$, a horizontal line at the level

$$L(\hat{\lambda}) - \tfrac{1}{2}\chi_1^2(1 - \alpha) \tag{13.2.9}$$

of the vertical scale. This would cut the curve at two values of $\lambda$, and these would be the end points of the approximate confidence interval. Translating this via (13.2.6), we see that we must cut across at heights of

$$S(\lambda, \mathbf{V}) = S(\hat{\lambda}, \mathbf{V})e^{\chi_1^2(1-\alpha)/n} \qquad \text{for } S(\lambda, \mathbf{V}) \tag{13.2.10}$$

or

$$\ln S(\lambda, \mathbf{V}) = \ln S(\hat{\lambda}, \mathbf{V}) + \chi_1^2(1 - \alpha)/n \qquad \text{for } \ln S(\lambda, \mathbf{V}) \tag{13.2.11}$$

according to which plot is used. In both cases we cut across the plot somewhat above the minimum level. $S(\hat{\lambda}, \mathbf{V})$ is the minimum sum of squares value that occurs at $\lambda = \hat{\lambda}$.

### The Confidence Statement Has Several Forms

Readers who consult various sources on this matter will perhaps be confused when they read, elsewhere, that instead of the factor $\exp\{\chi_1^2(1 - \alpha)/n\}$ recommended on the right of (13.2.10), they are told to use $1 + t_\nu^2/\nu$, or $1 + z^2/\nu$, or $1 + \chi_1^2(1 - \alpha)/\nu$, or $1 + \chi_1^2(1 - \alpha)/n$, or $1 + z^2/n$, where $t_\nu$ and $z$ are the two-tailed percentage points

**TABLE 13.2. Mooney Viscosity $MS_4$ at 100°C as Function of Filler and Oil Levels in SBR-1500**[a]

| Naphthenic Oil,[b] phr, $p$ | Filler, phr, $f$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 12 | 24 | 36 | 48 | 60 |
| 0 | 26 | 38 | 50 | 76 | 108 | 157 |
| 10 | 17 | 26 | 37 | 53 | 83 | 124 |
| 20 | 13 | 20 | 27 | 37 | 57 | 87 |
| 30 | — | 15 | 22 | 27 | 41 | 63 |

[a] Phillips Petroleum Co.
[b] Cirolite Process Oil, Sun Oil Co.

of a $t$-variable with $\nu$ df [the df of the residual SS, $S(\lambda, \mathbf{V})$] and a unit normal variable, respectively. [For (13.2.11), take natural logarithms, ln, of the quantities listed.] All of these alternatives are based on the points that (i) $\exp(x) = 1 + x + x^2/(2!) + x^3/(3!) + \cdots$, which, cut to $1 + x$ as a first approximation, accounts for the "1+" portion; (ii) $\chi_1^2 = z^2 \approx t_\nu^2$ unless $\nu$ is small; and (iii) one can argue about whether $n$ df or $\nu$ df should be used. For most problems, all these intervals will be more or less the same; thus agonizing about which to use is usually a waste of time. If in doubt, use the most conservative one, the one that gives the largest confidence interval. We recommend (13.2.10) or (13.2.11) in general, however.

***Example 1.*** The data in Table 13.2 are part of a more extensive set given by Derringer (1974). This paper has been adapted with permission of John Wiley & Sons, Inc. We wish to find a transformation of the form $(Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1})$ for $\lambda \neq 0$, or $\dot{Y} \ln Y$ for $\lambda = 0$, which will provide a good first-order fit to the data and leave satisfactory residuals. Our model form Eq. (13.2.5) is

$$V = \beta_0 + \beta_1 f + \beta_2 p + \epsilon, \tag{13.2.12}$$

where $f$ is the filler level and $p$ is the plasticizer level (the latter is indicated in the first column of Table 13.2).

Note that the response data range from 157 to 13, a ratio of 157/13 = 12.1. When the ratio of the largest response value to the smallest is, or exceeds, about an order of magnitude (i.e., about 10), a transformation on $Y$ is likely to be effective. The geometric mean is $\dot{Y} = 41.5461$.

Table 13.3 shows selected values of $S(\lambda, \mathbf{V})$ for various $\lambda$. [An initial set of calcula-

**TABLE 13.3. Values of $S(\lambda, V)$ for Selected Values of $\lambda$ for the Viscosity Data**

| $\lambda$ | $S(\lambda, V)$ | $\lambda$ | $S(\lambda, V)$ |
|---|---|---|---|
| −1.0 | 2456 | −0.04 | 83.5 |
| −0.8 | 1453 | −0.02 | 85.5 |
| −0.6 | 779.1 | 0.00 | 89.3 |
| −0.4 | 354.7 | 0.05 | 106.7 |
| −0.2 | 131.7 | 0.10 | 135.9 |
| −0.15 | 104.5 | 0.2 | 231.1 |
| −0.10 | 88.3 | 0.4 | 588.0 |
| −0.08 | 84.9 | 0.6 | 1222 |
| −0.06 | 83.3 | 0.8 | 2243 |
| −0.05 | 83.2 | 1.0 | 3821 |

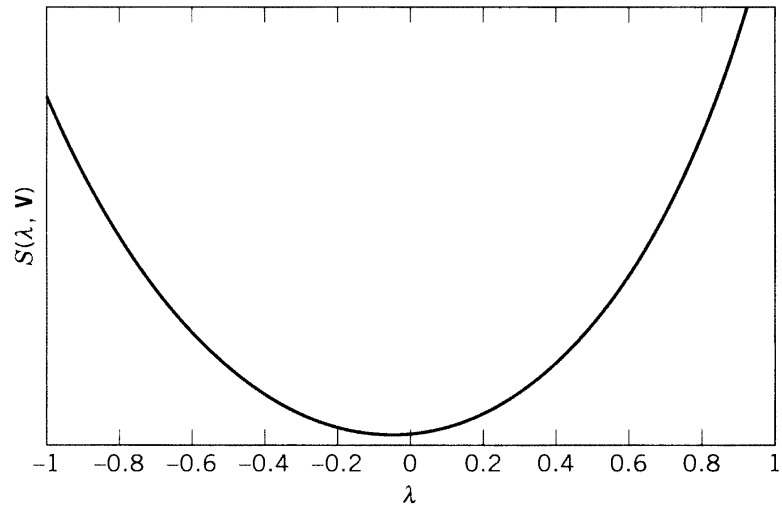


**Figure 13.1.** Plot of $S(\lambda, \mathbf{V})$ versus $\lambda$ for the viscosity data.

tions for $\lambda = 2(0.1)2$ was followed by a finer division $\lambda = -0.2(0.01)0.1$ near the bottom of the curve.] A smooth curve through these points is plotted in Figure 13.1. We see that the minimum ($S(\lambda, \mathbf{V})$ occurs at about $\lambda = -0.05$. This is close to zero, suggesting that the transformation

$$V = \dot{Y} \ln Y, \tag{13.2.13}$$

or more simply $\ln Y$, might be a suitable one for this set of data. The approximate 95% confidence interval obtained via Eq. (13.2.10) at a level of $S(\lambda, \mathbf{V}) = 98.3$ is given

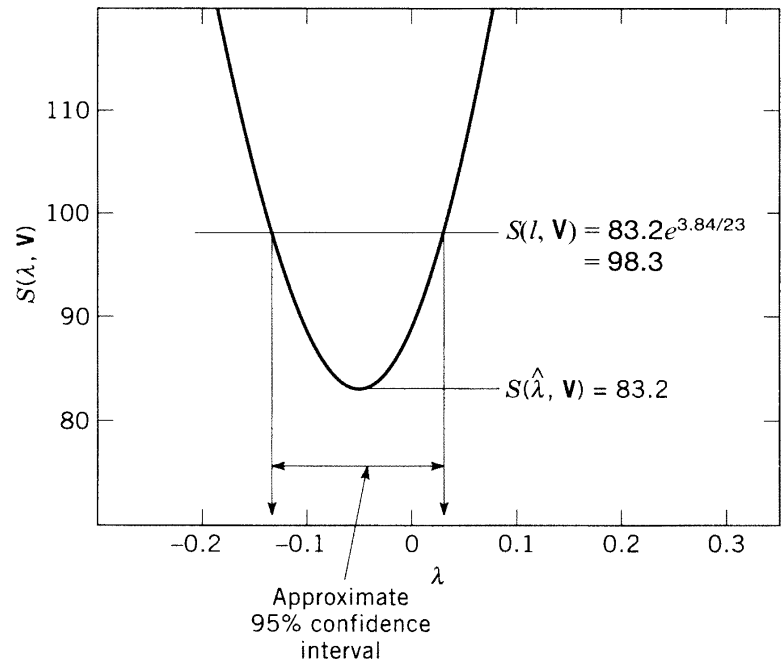


**Figure 13.2.** Obtaining an approximate 95% confidence interval for $\lambda$ using the viscosity data.

by $-0.13 \leq \lambda \leq 0.03$. An enlargement of the bottom of the $S(\lambda, \mathbf{V})$ curve is drawn in Figure 13.2 to show this calculation more clearly. We see that the validity of using $\lambda = 0$ is confirmed by this calculation, and that the transformation is well estimated. Values of $\lambda$ such as $\lambda = 1$ (no transformation at all), $\lambda = \frac{1}{2}$ (the square root transformation), $\lambda = -1$ (the inverse transformation), and many others are completely excluded as possibilities by the data. If the alternative factors for determining approximate 95% confidence intervals are used, the "cuts" across Figure 13.2 would take place at heights of 101.3, or 99.2, or 99.2, or 97.1, or 97.1, respectively. Clearly, whichever is used, there is no practical difference in the conclusion reached.

*Note*: A wide confidence interval that included two or more of the benchmark levels of $\lambda = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$ would indicate that $\lambda$ is not crisply estimated and would imply that it made little difference which of a wide range of possibilities for $\lambda$ was used. While, in one sense, this seems advantageous, it also may mean that the resulting fitted equation will not predict effectively. If a wide confidence interval includes $\lambda = 1$, the implication is that it may not be worthwhile to transform $Y$ at all.

Application of the natural logarithm transformation to the original data gives us the transformed data of Table 13.4. The best plane, fitted to these transformed data by least squares, is now

$$\widehat{\ln Y} = 3.212 + 0.03088f - 0.03152p. \tag{13.2.14}$$

The corresponding analysis of variance table is shown as Table 13.5. Of the variation about the mean, $100R^2 = 99.51\%$ is explained by the three-parameter model and the $F$-statistic for overall regression. $F = 2045$ is very significant indeed. Clearly, an excellent fit has been attained.

If we had fitted a first-order model to the *untransformed* data we would have obtained

$$\hat{Y} = 28.184 + 1.55f - 1.717p, \tag{13.2.15}$$

with a $100R^2$ value of 87.93% and an overall $F = 72.9$ (see Table 13.6). This, in itself, is an excellent fit, but the improvement when $\ln Y$ is used in quite dramatic. (In other examples, the initial fit can be quite poor, and the proper transformation enables a significant fit to be achieved; sometimes, the transformation enables a lower degree of polynomial to be fitted than would otherwise be possible. This is true here, too, as we explain below.)

***Coding the Predictors.*** To avoid complicating our example with additional steps, we have used the two predictor variables $f$ and $p$ in the units in which they were given. In a case like the above, where the levels of $f$ and $p$ are equally spaced, the codings

$$x_1 = (f - 30)/6, \qquad x_2 = (p - 15)/5 \tag{13.2.16}$$

would provide coded levels of $x_1 = -5, -3, -1, 1, 3, 5$ and $x_2 = -3, -1, 1, 3$, a slight

**TABLE 13.4. Transformed Values $W = \ln Y$ of the Data in Table 13.2**

| Naphthenic Oil, phr, $p$ | Filler, phr, $f$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 12 | 24 | 36 | 48 | 60 |
| 0 | 3.258 | 3.638 | 3.912 | 4.331 | 4.682 | 5.056 |
| 10 | 2.833 | 3.258 | 3.611 | 3.970 | 4.419 | 4.820 |
| 20 | 2.565 | 2.996 | 3.296 | 3.611 | 4.043 | 4.466 |
| 30 | — | 2.708 | 3.091 | 3.296 | 3.714 | 4.143 |

**TABLE 13.5. Analysis of Variance of First-Order Model in $f$ and $p$ Fitted to Logged Viscosity Data**

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| $b_0$ | 1 | 319.44855 | — | |
| $b_1, b_2 \mid b_0$ | 2 | 10.55167 | 5.27583 | 2045 |
| Residual | 20 | 0.05171 | 0.00258 | |
| Total | 23 | 330.05193 | | |

numerical simplification. Note that this sort of simple coding of the predictors has no effect whatsoever on the estimation of $\lambda$. However, in some problems, proper coding will simplify the regression calculations. For example, if the $f = 0, p = 30$ observation were not missing in Table 13.2, the coding as shown in Eq. (13.2.16) would make the $x_1$ and $x_2$ columns orthogonal to each other, and to the column of 1's in the $\mathbf{X}$ matrix. (Note, however, that *transformation* of the predictors, say, to $x_1 = f^{\alpha_1}$, $x_2 = p^{\alpha_2}$ will alter the problem completely and *will* affect the estimation of $\lambda$.)

### Importance of Checking Residuals

Transformations on the response variable affect the distribution of errors. Our assumption is that, *after* the transformation, the errors in the transformed response will be $N(\mathbf{0}, \mathbf{I}\sigma^2)$. Thus it is important to examine the residuals from the model finally fitted, to see if those assumptions appear to be violated. The residuals from the first-order fit Eq. (13.2.14) are given in Table 13.7. We leave their examination as an exercise for the reader.

## 13.3. A SECOND METHOD FOR ESTIMATING $\lambda$

In the second method of estimation, we choose $\lambda$ to minimize some quantity that we desire to be small and/or maximize some quantity that we desire to be large. For example, suppose that the original response $Y$ could reasonably be fitted by a second-order model in $X_1$ and $X_2$,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon, \tag{13.3.1}$$

and that the idea behind transforming from $Y$ to $V$ via Eq. (13.2.2) is to attempt to represent the transformed response by a first-order model $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. We could fit Eq. (13.3.1) to $V$ by least squares for a selected set of values of $\lambda$ and choose, as best for our purposes, the value of $\lambda$ that minimized an appropriate statistic. Possible choices are the $F$-value connected with the extra sum of squares $\mathrm{SS}(b_{11}, b_{22}, b_{12} \mid b_0, b_1, b_2)$ or the ratio of mean squares arising from second- and first-order fitted parame-

**TABLE 13.6. Analysis of Variance of First-Order Model in $f$ and $p$ Fitted to Untransformed Viscosity Data**

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| $b_1 b_2 \mid b_0$ | 2 | 27,842.62 | 13,921.31 | 72.9 |
| Residual | 20 | 3,820.60 | 191.03 | |
| Total, corrected | 22 | 31,663.22 | | |

**TABLE 13.7. Residuals Multiplied by 1000, from First-Order Model Fitted to Logged Viscosity Data**

| Naphthenic Oil, phr, $p$ | Filler, phr, $f$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 12 | 24 | 36 | 48 | 60 |
| 0 | 46 | 55 | −41 | 7 | −13 | −9 |
| 10 | −64 | −10 | −27 | −39 | 39 | 70 |
| 20 | −17 | 43 | −27 | −83 | −21 | 31 |
| 30 | — | 71 | 83 | −83 | −36 | 23 |

ters. For our idea to be successful, second-order terms would have to be nonsignificant for the value of λ finally selected.

***Example 2.*** We again use the viscosity data of Table 13.2. We wish to find a transformation of the form $V = (Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1})$ for $\lambda \neq 0$, or $V = \dot{Y} \ln Y$ for $\lambda = 0$, that will allow a good first-order fit without need for second-order terms. We first fit the model

$$V = \beta_0 + \beta_1 f + \beta_2 p + \beta_{11} f^2 + \beta_{22} p^2 + \beta_{12} pf + \epsilon, \tag{13.3.2}$$

where, as before, $f$ is the filler level and $p$ is the plasticizer level, for a series of chosen values of λ. [Fitting in coded variables such as those given in Eq. (13.2.16) could also be done without affecting the basic results.] For each λ we evaluate

$$\begin{aligned} \text{MS}_1 &= \text{mean square arising from SS}(b_1, b_2 \mid b_0)/2, \\ \text{MS}_2 &= \text{mean square arising from SS}(b_{11}, b_{22}, b_{12} \mid b_0, b_1, b_2)/3, \\ \gamma &= \text{MS}_2/\text{MS}_1, \end{aligned} \tag{13.3.3}$$

and we plot γ against λ to give Figure 13.3. The numbers needed to create this plot are given in Table 13.8.

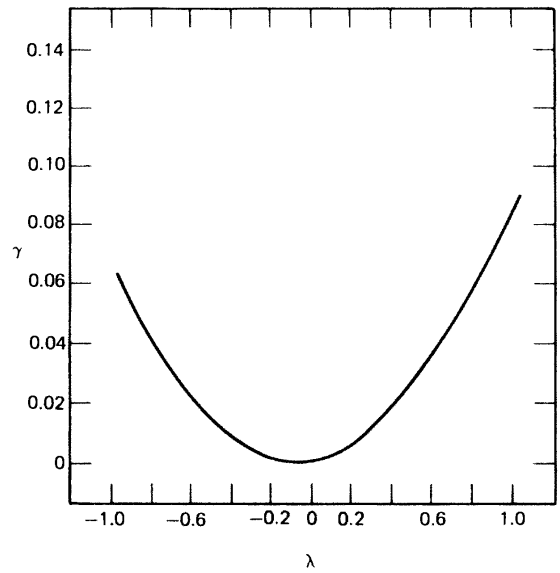


**Figure 13.3.** Fitting a second-order model to transformed viscosity data: plot of $\gamma = \text{MS}_2/\text{MS}_1$ versus λ.

**TABLE 13.8. Values of $MS_1$, $MS_2$, and $\gamma = MS_2/MS_1$ for Selected Values of $\lambda$ for the Viscosity Data**

| $\lambda$ | $MS_1$ | $MS_2$ | $\gamma = MS_2/MS_1$ | $\lambda$ | $MS_1$ | $MS_2$ | $\gamma = MS_2/MS_1$ |
|---|---|---|---|---|---|---|---|
| −1.0 | 11162 | 724 | 0.0649 | 0.025 | 9133 | 7 | 0.0008 |
| −0.8 | 10218 | 424 | 0.0415 | 0.05 | 9162 | 11 | 0.0012 |
| −0.6 | 9572 | 218 | 0.0228 | 0.1 | 9233 | 21 | 0.0023 |
| −0.4 | 9183 | 85 | 0.0093 | 0.2 | 9421 | 54 | 0.0057 |
| −0.2 | 9030 | 16 | 0.0018 | 0.4 | 10001 | 173 | 0.0173 |
| −0.1 | 9040 | 3 | 0.0003 | 0.6 | 10891 | 382 | 0.0351 |
| −0.05 | 9066 | 2 | 0.0002 | 0.8 | 12162 | 711 | 0.0584 |
| −0.025 | 9084 | 3 | 0.0003 | 1.0 | 13921 | 1207 | 0.0867 |
| 0.0 | 9107 | 4 | 0.0005 | | | | |

We see that the minimum $\gamma$ is attained at about $\lambda = -0.05$, indicating that use of $\lambda = 0$, the logarithmic transformation, is sensible, exactly as we found via the previous method. (A disadvantage[2] of the present procedure, however, is that we cannot easily obtain a confidence interval for $\lambda$.) This transformed response leads to the fitted second-order equation

$$\widehat{\ln Y} = 3.231 + 0.02861f - 0.03346p + 0.00004416f^2 + 0.00011207p^2 - 0.00003718fp. \tag{13.3.4}$$

The corresponding analysis of variance table is shown in Table 13.9. It is clear that the transformation is a successful one, that the full second-order model is not needed, and that the first-order model Eq. (13.2.14) is perfectly adequate. By comparison, the fitted second-order equation for the untransformed data is

$$\hat{Y} = 24.067 + 0.57387f - 0.82628p + 0.02639f^2 + 0.02752p^2 - 0.04930fp, \tag{13.3.5}$$

with analysis of variance table as in Table 13.10. Thus when no transformation is carried out, significant second-order curvature is present in the data.

### Advantages of the Likelihood Method

Of the two methods given for estimating transformation parameters, we would favor the likelihood method in practice for most problems. Through it, we can always obtain

**TABLE 13.9. Analysis of Variance of Second-Order Model in *f* and *p* Fitted to Logged Viscosity Data**

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| $b_1, b_2 \mid b_0$ | 2 | 10.55167 | 5.27583 | 2037.0 |
| $b_{11}, b_{22}, b_{12} \mid b_0, b_1, b_2$ | 3 | 0.00776 | 0.00259 | 1.0 |
| Residual | 17 | 0.04395 | 0.00259 | |
| Total, corrected | 22 | 10.60338 | | |

[2] This disadvantage exists in this example. However, if the method is used on an $F$-statistic, an approximate confidence interval can be calculated using $F(\nu_1, \nu_2, 1 - \alpha)$ as described in "Transformations: some examples revisited," by N. R. Draper and W. G. Hunter, *Technometrics*, **11**, 1969, 23–40.

**TABLE 13.10. Analysis of Variance of Second-Order Model in *f* and *p* Fitted to Original Viscosity Data**

| Source | df | SS | MS | *F* |
|---|---|---|---|---|
| $b_1, b_2 \mid b_0$ | 2 | 27,842.616 | 13,921.308 | 1179.7 |
| $b_{11}, b_{22}, b_{12} \mid b_0, b_1, b_2$ | 3 | 3,619.987 | 1,206.662 | 102.3 |
| Residual | 17 | 200.615 | 11.801 | |
| Total, corrected | 22 | 31,663.217[a] | | |

[a] Rounding discrepancy of 0.001.

an approximate confidence interval or region, and we have only to fit the model we are interested in, not a more complicated one, as is usually required in the second method. (In some problems, in fact, the data may be inadequate to fit the desired higher-order alternative.) The second method can be useful, however, when it is desired to examine a variety of criteria. The various plots of the criteria versus $\lambda$ can be viewed simultaneously, and a compromise value of $\lambda$ can be selected from these plots.

## 13.4. RESPONSE TRANSFORMATIONS: OTHER INTERESTING AND SOMETIMES USEFUL PLOTS

The procedures described in Sections 13.2 and 13.3 for estimating a transformation are the basic ones. A number of other plots (not illustrated) provide useful ancillary information when needed.

1. If there are genuine repeat runs in the data set, a pure error sum of squares can be calculated for each $\lambda$ used and a plot of the lack of fit $F$-value versus $\lambda$ can be constructed. A horizontal line can be drawn at the height of some chosen significance level taken from the $F$-tables, and acceptable values of $\lambda$ are those for which the curve falls below the significance line. Typically (but not always) the acceptable range of $\lambda$'s includes $\hat{\lambda}$ and most or all of the confidence interval for $\lambda$. When this does not happen, deeper investigation is called for. For examples where this technique can be used, see Exercises 13D and 13E in "Exercises for Chapter 13."

2. The regression coefficients can be plotted versus $\lambda$.

3. The $t$-values for each of the fitted coefficients can be plotted versus $\lambda$. Alternatively, $t$-values regarded as significant can be incorporated into the regression coefficient plot (2) by using, for example, solid lines for those values of the coefficients regarded as significantly different from zero (or from whatever other test value is decided upon), and dashed lines otherwise. (Of course, for those $\lambda$-values for which the errors are non-normal, a comparison with the $t$-distribution is not, strictly speaking, valid. Also, for $\lambda$-values at which the variance is nonhomogeneous, the $t$-values can be misleading.)

4. $R^2$ may be plotted versus $\lambda$ to see, for which $\lambda$-values, acceptable expanatory levels are to be found. Note that this plot will not be simply a reflection of the $S(\lambda, \mathbf{V})$ plot. It is true that

$$S(\lambda, \mathbf{V}) = (\mathbf{V}'\mathbf{V} - n\overline{V}^2)(1 - R^2), \tag{13.4.1}$$

so that

$$R^2 = 1 - S(\lambda, \mathbf{V})/(\mathbf{V}'\mathbf{V} - n\overline{V}^2), \tag{13.4.2}$$

but the divisor on the right-hand side is not constant as $\lambda$ varies. Two such plots could be drawn for the first- and second-order models of Example 2. It is where these two $R^2$ curves differ the *least* that interesting values of $\lambda$ lie. (In fact, it would be possible to construct an interval of $\lambda$-values for which the difference between the $R^2$ values does not exceed some desired amount.)

## 13.5. OTHER TYPES OF RESPONSE TRANSFORMATIONS

### A Two-Parameter Family of Response Transformations

By adding an additional parameter, we can extend the range of possible transformations beyond those discussed above. Consider the two-parameter family

$$V = \begin{cases} \dfrac{(Y + \lambda_2)^{\lambda_1} - 1}{\lambda_1 \dot{Y} + \lambda_2)^{\lambda_1 - 1}}, & \text{for } \lambda_1 \neq 0, \\ (\dot{Y} + \lambda_2) \ln (Y + \lambda_2), & \text{for } \lambda_1 = 0, \end{cases} \tag{13.5.1}$$

where

$$\dot{Y} + \lambda_2 = \{(Y_1 + \lambda_2)(Y_2 + \lambda_2) \cdots (Y_n + \lambda_2)\}^{1/n} \tag{13.5.2}$$

is the geometric mean of the $(Y_i + \lambda_2)$ terms, and where, necessarily, $Y > -\lambda_2$. The methods used for the one-parameter family (in which $\lambda_2 = 0$) can be extended for this case. We now have $Y + \lambda_2$, where $Y$ was previously, and must search over a two-dimensional $(\lambda_1, \lambda_2)$ grid to find the maximum likelihood values $(\hat{\lambda}_1, \hat{\lambda}_2)$. Also, in the calculation to obtain a confidence region for $(\lambda_1, \lambda_2)$, we now use $\chi^2$ with *two* degrees of freedom, rather than one, because there are now two transformation parameters. Exactly the same ideas are involved, but the calculations are more complicated. An example is discussed briefly on pp. 225–226 of the 1964 Box and Cox paper. The $\chi^2_2$ $(1 - \alpha)$ percentage points needed for 2 df are these:

| $\alpha$ | 0.10 | 0.05 | 0.025 | 0.01 | 0.001 |
|---|---|---|---|---|---|
| $\chi^2_2(1 - \alpha)$ | 4.61 | 5.99 | 7.38 | 9.21 | 13.82 |

Note that, because $Y > -\lambda_2$, we must avoid searching $\lambda_2$ values such that $\lambda_2 \leq -Y_{\min}$. Although a singularity occurs at this $\lambda_2$ boundary, it is not relevant; the local maximum satisfying the constraint is what is needed.

### A Modulus Family of Response Transformations

When a residuals plot indicates a fairly symmetric but non-normal error distribution, the one-parameter *modulus* power family

$$W = \begin{cases} (\text{sign of } Y)[\{|Y| + 1\}^{\lambda} - 1]/\lambda, & \lambda \neq 0 \\ (\text{sign of } Y) \ln \{|Y| + 1\}, & \lambda = 0 \end{cases} \tag{13.5.3}$$

may be useful. For an illustration, see John and Draper (1980).

### Transforming Both Sides of the Model

An extremely versatile method of transformation is to transform the model function as well as the response variable, while leaving the error additive. For example, a model

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \epsilon \tag{13.5.4}$$

could be transformed and fitted as

$$Y^{\lambda} = \{f(\mathbf{X}, \boldsymbol{\beta})\}^{\lambda} + \epsilon, \tag{13.5.5}$$

where the best $\lambda$-value needs to be selected. Alternatively, $Y^{\lambda}$ and $\{f(\mathbf{X}, \boldsymbol{\beta})\}^{\lambda}$ could be replaced by their respective $W$ or $V$ forms discussed in Section 13.2 This method requires a great deal of computation but promises to be more rewarding in its effects in situations where the model (13.5.4) is reasonable one to fit, but the residuals show signs that they may be non-normal or heteroscedastic. For more on this see, for example, Chapter 4 of Carroll and Ruppert (1988).

### A Power Family for Proportions

Exactly the same ideas as in Section 13.2 can be used to estimate the parameter $\lambda$ in the family

$$P = \{p^{\lambda} - (1 - p)^{\lambda}\}/\lambda, \tag{13.5.6}$$

where $p$ is the observed proportion of times a stated event takes place. The observed values of $p$ would, in general, depend on the values of a number of predictor variables $X_1, X_2, \ldots$, and a model of general form

$$P = f(X_1, X_2, \ldots, \lambda, \boldsymbol{\beta}) + \epsilon \tag{13.5.7}$$

would be postulated where $\boldsymbol{\beta}$ is a vector of parameters. The value of $\lambda$ would be chosen to give the best fit to these data under the assumption that $\epsilon \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$.

The power transformation (13.5.6) was suggested by Tukey. For work on the statistical distribution of $P$ when $p$ has a uniform distribution, good initial reading is Joiner and Rosenblatt (1971); see, also, the references mentioned in the paper.

Two specific transformations for proportions are given in the next section. One is a special case of the above family, and the second is an approximation to a special case.

## 13.6. RESPONSE TRANSFORMATIONS CHOSEN TO STABILIZE VARIANCE

In situations where the transformed data are to be analyzed by least squares, it is important that the variance of the response to be fitted be independent of its mean value. Where it is known or where it has been found empirically that the standard deviation of the untransformed resonse $Y$, say, $\sigma_Y$, is a particular function $f(\eta)$ of the mean value, $\eta = E(Y)$, we can obtain an appropriate transformation immediately by using the transformed variable $h(y)$, where

$$\frac{\partial h(Y)}{\partial Y} \propto \frac{1}{f(Y)}. \tag{13.6.1}$$

In other words, we obtain $h(Y)$ by integrating $1/f(Y)$ with respect to $Y$. Some well-

**TABLE 13.11. Appropriate Variance Stabilizing Transformation when $\sigma_Y = f(\eta)$**

| Nature of Dependence $\sigma_y = f(\eta)$ | | Variance Stabilizing Transformation[a] |
|---|---|---|
| $\sigma_Y \propto \eta^k$ and in particular | $(Y \geq 0)$ | $Y^{1-k}$ |
| $\sigma_Y \propto \eta^{1/2}$ (Poisson) | $(Y \geq 0)$ | $Y^{1/2}$ |
| $\sigma_Y \propto \eta$ | $(Y \geq 0)$ | $\ln Y$ |
| $\sigma_Y \propto \eta^2$ | $(Y \geq 0)$ | $Y^{-1}$ |
| $\sigma_Y \propto \eta^{1/2}(1 - \eta)^{1/2}$ (binomial) | $(0 \leq Y \leq 1)$ | $\sin^{-1}(Y^{1/2})$ |
| $\sigma_Y \propto (1 - \eta)^{1/2}/\eta$ | $(0 \leq Y \leq 1)$ | $(1 - Y)^{1/2} - (1 - Y)^{3/2}/3$ |
| $\sigma_Y \propto (1 - \eta^2)^{-2}$ | $(-1 \leq Y \leq 1)$ | $\ln\{(1 + Y)/(1 - Y)\}$ |

[a] Modifications for the Poisson and binomial cases have been suggested by Freeman and Tukey (1950). These modifications are discussed in the text.

known transformations that arise in this way are shown in Table 13.11. Note that some of these are members of the power family.

### Estimation of $k$ in Table 13.11

If the data contain $m$ sets of replicate runs, it is possible to get an estimate of $k$ fairly quickly by finding the slope of the "best line" through the points (abscissa, ordinate) = $(\ln \overline{Y}_j, \ln s_j)$, $j = 1, \ldots, m$, where $\overline{Y}_j$ and $s_j^2$ are the sample mean and variance of the $j$th set of repeat runs. This method is based on the idea that, if $\sigma_Y$ is proportional to the $k$th power of $\eta$, so that $\sigma_Y \propto \eta^k$, it follows that

$$\ln \sigma_Y = k_0 + k \ln \eta. \tag{13.6.2}$$

The pairs $(\ln s_j, \ln \overline{Y}_j)$ provide us with some data on $(\ln \sigma_Y, \ln \eta)$ from which we can estimate $k$. This slope estimation is often done simply by eye. A least squares fit calculation would also provide us with "se($\hat{k}$)," which would enable us to judge, at least roughly, how precise our estimation of $k$ appears to be.

When there are no replicate runs, the residuals versus $\hat{Y}$ plot should be examined. At a particular $\hat{Y}_j$ value the width of the band of residuals could be taken as roughly $4s_j$. A few values of $(\ln s_j, \ln \hat{Y}_j)$ can then be plotted and the slope $k$ estimated as in the foregoing paragraph. After transformation, a new residuals versus $\hat{Y}$ plot can be done to see if the transformation was effective or needs readjustment.

### Transformations for Responses That Are Proportions

Many types of response data occur as proportions, $0 \leq Y_i \leq 1$, obtained as the number of times a "success" (however that may be defined) happens in a larger number of "trials." For example, six rats of ten complete a given task for a $Y_i = 0.60$. Proportion-type data typically do not have a uniform variance pattern because $V(Y_i) = \pi_i(1 - \pi_i)/m_i$, where $E(Y_i) = \pi_i$ and $m_i$ is the number of trials. Two popular transformations for such data are the following.

1. *The log odds transformation.* (This does *not* stabilize variance.) We set

$$W_i = \ln\{Y_i/(1 - Y_i)\}. \tag{13.6.3}$$

Thus $W_i$ is the natural logarithm of the "odds ratio" $Y_i/(1 - Y_i)$, the ratio of the proportion of successes to the proportion of failures. To fit the model

$$W_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{p-1} X_{p-1,i} + \epsilon_i \tag{13.6.4}$$

to data $(W_i, X_{1i}, \ldots, X_{p-1,i})$, $i = 1, 2, \ldots, n$, we use weighted least squares because

$$V(W_i) = 1/\{\pi_i(1 - \pi_i)m_i\}, \tag{13.6.5}$$

approximately and is not constant. To show this, we use the fact that $\ln(1 + x) = x - x^2/2 + x^3/3 - \cdots \simeq x$ for small $x$. Then, dropping the subscript $i$ for the moment, we see that (all results are approximate)

$$\ln Y = \ln \pi + (Y - \pi)/\pi \tag{13.6.6}$$

so that

$$E(\ln Y) = \ln \pi \quad \text{and} \quad V(\ln Y) = (1 - \pi)/(\pi m). \tag{13.6.7}$$

Similarly,

$$E\{\ln(1 - Y)\} = \ln(1 - \pi) \quad \text{and} \quad V\{\ln(1 - Y)\} = \pi/\{(1 - \pi)m\} \tag{13.6.8}$$

and

$$\operatorname{cov}\{\ln Y, \ln(1 - Y)\} = -1/m. \tag{13.6.9}$$

It follows that $V(W) = \{(1 - \pi)/\pi + \pi/(1 - \pi) - 2(-1)\}/m = 1/\{\pi(1 - \pi)m\}$. These $n$ variances $V(W_i) = 1/\{\pi_i(1 - \pi_i)m_i\}$ are, of course, unknown but are estimated by the corresponding values $s_i^2 = 1/\{Y_i(1 - Y_i)m_i\}$ for $i = 1, 2, \ldots, n$, and the estimation proceeds as in Section 9.2. In this case, the matrix $\mathbf{V}$ is diagonal with estimated entries $s_1^2, s_2^2, \ldots, s_n^2$. An alternative analysis for the log odds ratio is to use generalized linear models; see Chapter 18.

2. *The arc-sine transformation.* (Use radians or degrees.) As indicated in Table 13.11, the transformation $U = \sin^{-1}Y^{1/2}$ will stablize the variance over a range of $Y$-values, if the samples that determine the observations $Y_i$ are all of the same size $m$, say. In fact, $W = 2\sin^{-1}Y^{1/2}$ is slightly preferred because it has the uniform theoretical variance $1/m$ (for radians; multiply by $\{360/(2\pi)\}^2$ for degrees). Note that, if $m$ is not constant throughout the data, $Z_i = 2m_i^{1/2}\sin^{-1}Y_i^{1/2}$ is needed, where $Y_i$ is determined through $m_i$ trials. Data in the middle range of proportions (say, 0.30–0.70) will not be much affected by these transformations, because of the approximate linearity of the transformation in that range. The range of $Y$-values over which the arc-sine transformation produces a flat variance function depends on the sample size $m$. For smaller samples, the flat range does not cover true values of the binomial parameter $E(Y)$ that are close to zero or one. Freeman and Tukey (1950) have suggested the improved transformation

$$U^* = \tfrac{1}{2}[\sin^{-1}\{m_iY_i/(m_i + 1)\}^{1/2} + \sin^{-1}\{(m_iY_i + 1)/(m_i + 1)\}^{1/2}]$$

(instead of $U$), which extends the range of flatness of the variance function. See Bisgaard and Fuller (1994–95, Figure 2) for examples for $m_i = 20, 50$.

In general, we must keep in mind that there is no guarantee that use of these transformations will necessarily be better than analyzing the proportions directly; much depends on the data. The effectiveness of a transformation is best assessed by trying it on the data and then checking the fit of the model and the pattern of residuals that results.

Basic information on the reasons for some transformations that can be made on the response variable can be found in Bartlett (1947).

### Transformations for Responses that Are Poisson Counts

Table 13.11 suggests the square root transformation for Poisson data. Again, a flatter variance profile is achieved by the Freeman and Tukey (1950) suggestion:

$$\tfrac{1}{2}\{Y^{1/2} + (Y + 1)^{1/2}\},$$

which improves the situation for small values of the Poisson parameter $\lambda = E(Y) = V(Y)$. See Bisgaard and Fuller (1994–95, Figure 3).

### References

Atkinson (1985); Bartlett (1947); Bisgaard and Fuller (1994–1995); Box and Cox (1964); Carroll and Ruppert (1988); Freeman and Tukey (1950).

## EXERCISES FOR CHAPTER 13

**A.** Consider the following representative data:

| Year, $X$ | Speed mph, $Y$ | Means of Attaining Speed |
|---|---|---|
| 1830 | 30 | Railroad |
| 1905 | 130 | Railroad |
| 1930 | 400 | Airplane |
| 1947 | 760 | Airplane |
| 1952 | 1,500 | Airplane |
| 1969 | 25,000 | Spaceship |

1. Plot these $(X, Y)$ data points. Do you feel this plot is an informative one or not? Why?
2. Transform the data by $Z = \log Y$, and plot the $(X, Z)$ points. Is this plot preferable to the previous plot or not? Why?
3. Can you find a reasonably simple transformation $U = f(Y)$ that will produce a (more or less) straight line plot for the points $(X, U)$?
4. Whatever you conclude in (3), plot the points $(X, V)$, where $V = \log(\log Y)$. Fit a straight line $V = \beta_0 + \beta_1 X + \epsilon$ to these points using least squares. Draw the fitted line on your $(X, V)$ plot. Find the residuals and comment on them.
5. Set down the appropriate analysis of variance table for (4), test for overall regression, and find $R^2$. Comment appropriately.
6. Use the fitted straight line from (4) to predict when humans will attain the speed of light (186,000 miles per second: note, per *second*).
7. Discuss the reasonability or otherwise of your prediction. On what assumptions does it depend? Whether you feel your prediction is realistic or unrealistic, set out your reasons carefully but succinctly.

**B.** (*Source*: "An empirical model for viscosity of filled and plasticized elastomer compounds," by G. C. Derringer, *Journal of Applied Polymer Science*, **18,** 1974, 1083–1101.) Two sets of response data (see footnotes *c* and *d*) are shown in Table B. For each of the sets, find the best transformation of the form $V = (Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1})$ for $\lambda \neq 0$, and $V = \dot{Y} \ln Y$ for $\lambda = 0$, that will allow a useful model of the form $V = \beta_0 + \beta_1 f + \beta_2 p + \epsilon$ to be developed via least squares, where $f$ is the filler level and $p$ is the naphthenic oil level. Perform all the usual regression analyses for your best $\hat{\lambda}$, including examination of residuals. (Note that the coding discussed in Section 13.2 will be especially useful for the second set of data.)

**TABLE B. Mooney Viscosity $MS_4$ at 100°C as Function of Filter and Oil Levels in SBR-1500[a]**

| Naphthenic Oil,[b] phr, $p$ | Filler, phr, $f$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 12 | 24 | 36 | 48 | 60 |
| 0 | 26[c] | 28 | 30 | 32 | 34 | 37 |
| | 25[d] | 30 | 35 | 40 | 50 | 60 |
| 10 | 18 | 19 | 20 | 21 | 24 | 24 |
| | 18 | 21 | 24 | 28 | 33 | 41 |
| 20 | 12 | 14 | 14 | 16 | 17 | 17 |
| | 13 | 15 | 17 | 20 | 24 | 29 |
| 30 | — | 12 | 12 | 13 | 14 | 14 |
| | 11 | 14 | 15 | 17 | 18 | 25 |

[a] SBR 1500, Phillips Petroleum Co.
[b] Circolite Process Oil, Sun Oil Co.
[c] N990, Cabot Corp.
[d] Silica B, Hi-Sil EP, PPG Industries.

**C.** (*Source*: "An empirical model for viscosity of filled and plasticized elastomer compounds," by G. C. Derringer, *Journal of Applied Polymer Science*, **18**, 1974, 1083–1101.)

**TABLE C. Mooney Viscosities $Y_1(ML_4)$ and $Y_2(MS_4)$ at 100°C for Various Combinations of Coded Variables[a] $x_1$, $x_2$, $x_3$, and $x_4$**

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $Y_1$ | $Y_2$ |
|---|---|---|---|---|---|
| −3 | −3 | −3 | −3 | 51 | 29 |
| −1 | −1 | −1 | −1 | 61 | 34 |
| −1 | −1 | −1 | −1 | 64 | 35 |
| −3 | −1 | −1 | 1 | 36 | 20 |
| −3 | 1 | 1 | −1 | 39 | 21 |
| −1 | −3 | −1 | −1 | 55 | 30 |
| 1 | −3 | 1 | 1 | 50 | 27 |
| −1 | 1 | −3 | 1 | 88 | 49 |
| 1 | −1 | −3 | −1 | 124 | 68 |
| −1 | −1 | 1 | −3 | 54 | 30 |
| 1 | 1 | −1 | −3 | 133 | 74 |

[a] $x_i = (X_i - 22\frac{1}{2})/7\frac{1}{2}$, $i = 1, 2$, $x_j = (X_j - 15)/5$, $j = 3, 4$, where $X_1$ = level of Silica A, Hi-Sil 233, PPG Industries, $X_2$ = level of N330, Cabot Corp., $X_3$ = Naphthenic oil, Circolite Process Oil, Sun Oil Co., and $X_4$ = Cumarone indene resin, Camar MH $2\frac{1}{2}$, Allied Chemical Corp.

For each response individually choose the best value of $\lambda$ in the transformation $V = (Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1})$ for $\lambda \neq 0$, $V = \dot{Y} \ln Y$ for $\lambda = 0$, to enable the model $V = \beta_0 + \Sigma \beta_i x_i + \epsilon$ to be well fitted to the data by least squares. After choosing the best transformation parameter estimate $\hat{\lambda}$ in each case, carry out all the usual regression analyses including examination of residuals.

**D.** (*Source*: "Third order rotatable designs in three factors: analysis," by N. R. Draper, *Technometrics*, **4**, 1962, 219–234.) The data in the accompanying table are constructed data, generated from a third-order polynomial model to illustrate use of a third-order sequential design. Find the best value of $\lambda$ in the transformation $V = (Y^\lambda - 1)/(\lambda \dot{Y}^{\lambda-1})$ for $\lambda \neq 0$, and $V = \dot{Y} \ln Y$ for $\lambda = 0$, to enable a full second-order ten-parameter model in $x_1, x_2, x_3$ to be fitted. After choosing the best transformation parameter estimate $\lambda$, carry out all the usual regression analyses and state your conclusions.

| $x_1$ | $x_2$ | $x_3$ | $y$ |
|---|---|---|---|
| −1 | −1 | −1 | 34.727 |
| 1 | −1 | −1 | 38.917 |
| −1 | 1 | −1 | 44.907 |
| 1 | 1 | −1 | 24.641 |
| −1 | −1 | 1 | 24.658 |
| 1 | −1 | 1 | 45.636 |
| −1 | 1 | 1 | 33.702 |
| 1 | 1 | 1 | 5.374 |
| $2^{1/2}$ | 0 | 0 | 33.414, 34.453 |
| $-2^{1/2}$ | 0 | 0 | 38.540, 39.201 |
| 0 | $2^{1/2}$ | 0 | 40.393, 38.335 |
| 0 | $-2^{1/2}$ | 0 | 40.687, 40.092 |
| 0 | 0 | $2^{1/2}$ | 23.869, 25.823 |
| 0 | 0 | $-2^{1/2}$ | 33.727, 33.068 |
| 0 | 0 | 0 | 43.832, 44.562 |
| 0 | 0 | 0 | 42.165, 41.187 |

**E.** (*Source*: Ice crystal growth data from B. F. Ryan, E. R. Wishart, and D. E. Shaw, Commonwealth Scientific and Industrial Research Organisation (C.S.I.R.O.), Australia. Related journal reference: "The growth rates and densities of ice crystals between −3°C and −21°C," *Journal of the Atmospheric Sciences*, **33**, 1976, 842–850.) Ice crystals are introduced into a chamber, the interior of which is maintained at a fixed temperature (−5°C) and a fixed level of saturation of air with water. The growth of the crystals with time is observed. The 43 sets of measurement presented here are of the mass of the crystals ($M$) in nanograms for times ($T$) of 50–180 seconds from the introduction of the crystals. Each measurement represents a single complete experiment; the experiments were conducted over a number of days and were randomized as to observation time. (The actual order in which they were conducted is not available.) It was desired to connect the response $M$ to the predictor $T$ by a simple fitted relationship. [The possibility that $E(M) = \alpha T^{\beta}$ was suggested.] Perform the following calculations.

1. Fit the model $V = \gamma + \beta \ln T$, where $V = (M^{\lambda} - 1)/(\lambda \dot{M}^{\lambda-1})$ for $\lambda \neq 0$, and $V = \dot{M} \ln M$ for $\lambda = 0$, for a suitable set of values of $\lambda$, and so pick the best transformation for $M$, using the methods described in Section 13.2.
2. Use the selected values of $\lambda$ to carry out all the details of the usual least squares analysis of the data and state your conclusions. In particular, do the residuals bear any suggestion that the variance structure of the errors might not be stable, as assumed?

| $T$ | $M$ | $T$ | $M$ |
|---|---|---|---|
| 50 | 11.5 | 125 | 47.7 |
| 60 | 8.2, 11.5 | 130 | 92.0, 87.2 |
| 70 | 14.1, 17.2 | 135 | 58.0, 47.7 |
| 80 | 33.5, 28.8 | 140 | 73.2, 58.0 |
| 90 | 15.6, 24.4, 33.5 | 145 | 47.7 |
| 95 | 38.8 | 150 | 118.9, 58.0 |
| 100 | 47.7, 58.0, 36.1 | 155 | 143.9, 87.2 |
| 105 | 47.7, 65.5 | 160 | 143.9, 73.2, 73.7 |
| 110 | 58.0, 47.7, 33.5 | 165 | 97.0 |
| 115 | 69.5, 69.5, 47.7 | 170 | 112.3 |
| 120 | 87.2, 51.0, 33.5 | 180 | 113.2 |

**F.** It is believed that a response relationship $\eta = \alpha X_1^{\beta} X_2^{\gamma}$ is responsible for producing the data below; how experimental errors enter into the situation is not known. Fit the model $\log Y = \log \alpha + \beta \log X_1 + \gamma \log X_2 + \epsilon$ to the data by least squares, examine the resulting fit by whatever methods are available to you, and provide conclusions that will cast some light on the situation. (Use logarithms to the base 10.)

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 10 | 10 | 2,040 |
| 100 | 10 | 7,350 |
| 1,000 | 10 | 12,210 |
| 10,000 | 10 | 23,580 |
| 10 | 100 | 18,200 |
| 100 | 100 | 10 |
| 1,000 | 100 | 2,960 |
| 10,000 | 100 | 108,040 |
| 10 | 1,000 | 10,370 |
| 100 | 1,000 | 1,150 |
| 1,000 | 1,000 | 23,580 |
| 10,000 | 1,000 | 296,120 |
| 10 | 10,000 | 9,040 |
| 100 | 10,000 | 1,960 |
| 1,000 | 10,000 | 96,980 |
| 10,000 | 10,000 | 1,004,020 |

**G.** The following data were collected on spray congealing:

**Values for the Experimental Operating Variables and Average Particle Sizes**

| Run | Feed Rate per Unit Whetted Wheel Periphery (gm/sec/cm) ($X_1$) | Peripheral Wheel Velocity (cm/sec) ($X_2$) | Feed Viscosity (poise) ($X_3$) | Mean Surface–Volume Particle Size of Product ($\mu$) ($Y$) |
|---|---|---|---|---|
| 1 | 0.0174 | 5300 | 0.108 | 25.4 |
| 2 | 0.0630 | 5400 | 0.107 | 31.6 |
| 3 | 0.0622 | 8300 | 0.107 | 25.7 |
| 4 | 0.0118 | 10800 | 0.106 | 17.4 |
| 5 | 0.1040 | 4600 | 0.102 | 38.2 |
| 6 | 0.0118 | 11300 | 0.105 | 18.2 |
| 7 | 0.0122 | 5800 | 0.105 | 26.5 |
| 8 | 0.0122 | 8000 | 0.100 | 19.3 |
| 9 | 0.0408 | 10000 | 0.106 | 22.3 |
| 10 | 0.0408 | 6600 | 0.105 | 26.4 |
| 11 | 0.0630 | 8700 | 0.104 | 25.8 |
| 12 | 0.0408 | 4400 | 0.104 | 32.2 |
| 13 | 0.0415 | 7600 | 0.106 | 25.1 |
| 14 | 0.1010 | 4800 | 0.106 | 39.7 |
| 15 | 0.0170 | 3100 | 0.106 | 35.6 |
| 16 | 0.0412 | 9300 | 0.105 | 23.5 |
| 17 | 0.0170 | 7700 | 0.098 | 22.1 |
| 18 | 0.0170 | 5300 | 0.099 | 26.5 |
| 19 | 0.1010 | 5700 | 0.098 | 39.7 |
| 20 | 0.0622 | 6200 | 0.102 | 31.5 |
| 21 | 0.0622 | 7700 | 0.102 | 26.9 |
| 22 | 0.0170 | 10200 | 0.100 | 18.1 |
| 23 | 0.0118 | 4800 | 0.102 | 28.4 |
| 24 | 0.0408 | 6600 | 0.102 | 27.3 |
| 25 | 0.0622 | 8300 | 0.102 | 25.8 |
| 26 | 0.0170 | 7700 | 0.102 | 23.1 |
| 27 | 0.0408 | 9000 | 0.613 | 23.4 |
| 28 | 0.0170 | 10100 | 0.619 | 18.1 |
| 29 | 0.0408 | 5300 | 0.671 | 30.9 |
| 30 | 0.0622 | 8000 | 0.624 | 25.7 |
| 31 | 0.1010 | 7300 | 0.613 | 29.0 |
| 32 | 0.0118 | 6400 | 0.328 | 22.0 |

(*continued*)

| Run | Feed Rate per Unit Whetted Wheel Periphery (gm/sec/cm) ($X_1$) | Peripheral Wheel Velocity (cm/sec) ($X_2$) | Feed Viscosity (poise) ($X_3$) | Mean Surface–Volume Particle Size of Product ($\mu$) ($Y$) |
|---|---|---|---|---|
| 33 | 0.0170 | 8000 | 0.341 | 18.8 |
| 34 | 0.0118 | 9700 | 1.845 | 17.9 |
| 35 | 0.0408 | 6300 | 1.940 | 28.4 |

*Source*: "Spray congealing: particle size relationships using a centrifugal wheel atomizer," by M. W. Scott, M. J. Robinson, J. F. Pauls, and R. J. Lantz, *Journal of Pharmaceutical Sciences*, **53**(6), June 1964, 670–675. Reproduced with permission of the copyright owner.

A proposed model, based on theoretical considerations, is

$$Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} \epsilon.$$

***Requirements.*** After transformation, fit the proposed model by least squares. State which predictor variable appears most important and check all coefficients for statistical significance (Take $\alpha = 0.05$). Is the model a satisfactory one?

**H.** Apply the $\sin^{-1}(Y^{1/2})$ transformation to the response observations below, and fit a planar model in all four $X$'s. Compare the root mean square of your analysis, which estimates the standard deviation of the response used, to the theoretical value $(4m)^{-1/2} = 0.05$.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | Fraction Defective |
|---|---|---|---|---|
| −1 | −1 | −1 | −1 | 0.16 |
| −1 | 1 | 1 | 1 | 0.17 |
| −1 | −1 | 1 | 1 | 0.12 |
| −1 | 1 | −1 | −1 | 0.06 |
| 1 | 1 | −1 | 1 | 0.06 |
| 1 | −1 | 1 | −1 | 0.68 |
| 1 | 1 | 1 | −1 | 0.42 |
| 1 | −1 | −1 | 1 | 0.26 |