

characteristic. By use of tables<sup>48</sup> of the expected values of the ordered elements of a random sample of  $m$  from  $N(0, 1)$  such data may be transformed to behave less nonnormally.

We have seen that the simplest safeguard against bad effects from inequality of variance in complete layouts is the use of equal cell numbers. If the ratios of the variances are known we can use the weighted analysis indicated in sec. 1.5, where there is some mention of the effects of using wrong weights, and in sec. 3.8, where there is an example after (3.8.3). If extreme inequality of variance is suspected and there are several observations in each of the cells of a layout, or at each abscissa in curve fitting, we may consider performing a weighted analysis with the weights inversely proportional to estimated variances.<sup>49</sup> For layouts higher than one-way this is computationally awkward, for even with equal cell numbers, orthogonality and the resulting simplicity of computation are lost in a weighted analysis. Furthermore, there is little theoretical knowledge of the effect of replacing unknown constant weights by random variables.<sup>50</sup> Transformations to reduce inequality of variance will be treated in sec. 10.7.

Where a correlation is suspected to which might be applied the simple model we have used of serial correlation with a single coefficient  $\rho$ , one might try to estimate  $\rho$  from the data and to approximate<sup>51</sup> the effect in the inferences of a true  $\rho$  equal to the estimated  $\rho$ . Some kinds of correlation can be brought under existing multivariate theory, like the correlation within columns in the mixed model for the two-way layout in sec. 8.1. In general, of the three kinds of possible departures from assumptions we have considered, those caused by lack of independence are the most formidable to cope with.

## 10.7. TRANSFORMATIONS OF THE OBSERVATIONS

Transformations are used sometimes to reduce interactions (end of sec. 4.1) or to reduce nonnormality (see above), but most frequently to reduce inequality of variance.<sup>52</sup> Most of the commonly used transformations

<sup>48</sup> Fisher and Yates (1943), Table XXI.

<sup>49</sup> This is standard procedure in fitting curves for quantal response to dosage by the probit method; see Fisher and Yates (1943), introduction to Tables X and XI.

<sup>50</sup> For the one-way layout with  $I$  groups there are some exact results for  $I = 2$  by P. L. Hsu (1938a), and approximate results for  $I > 2$  by James (1951) and Welch (1951), which see for further references.

<sup>51</sup> Box's (1954b) exact results are available for the two-way layout.

<sup>52</sup> It is curious that if the transformation (10.7.1), derived as one to stabilize variance, is applied to the sample correlation coefficient  $r$  (when it is called Fisher's transformation) it makes the distribution behave much more like a normal one; see Cramér (1946), sec. 29.7.

are special cases or modifications<sup>53</sup> of the following general transformation: Suppose that the mean, in general unknown, of a random variable  $y$  is denoted by  $\mu$  and that the standard deviation of  $y$  can be expressed as a function of  $\mu$  only, which is either completely known or known up to an unknown constant multiplier, say  $\sigma_y = \phi(\mu)$ . This will generally be the case if the distribution of  $y$  depends on a single parameter (which may however be a function of other parameters of interest in a problem). For example, for the binomial distribution of the number  $y$  of successes in  $n$  trials with constant probability  $p$ ,  $E(y) = np$ ,  $\sigma_y = [np(1-p)]^{1/2}$ , and so  $y$  has the required property with  $\phi(\mu) = [\mu(1-\mu)]^{1/2}$ . Consider a transformation  $z = f(y)$ , which we will try to determine so that the standard deviation of  $z$  is equal, at least approximately, to a predetermined constant  $\sigma_z$ . By the approximate formula  $\sigma_z = \sigma_y f'(\mu)$ , obtained by approximating  $z$  as a linear function of  $y$  in the neighborhood of  $y = \mu$ , we get  $f'(\mu) = \sigma_z/\phi(\mu)$  and, integrating this, and switching notation,

$$(10.7.1) \quad f(y) = \sigma_z \int \frac{dy}{\phi(y)}.$$

Thus in the above example we have

$$f(y) = \sigma_z \int [y(1-y)]^{-1/2} dy = 2n^{1/2} \sigma_z \arcsin (y/n)^{1/2} + C,$$

and we take  $C = 0$ . If we choose  $\sigma_z = (4n)^{-1/2}$ , the transformation becomes the "angular transformation"  $z = \arcsin (y/n)^{1/2}$ , where  $y/n$  is the observed proportion of successes and the arcsine is in radians; if the arcsine is in degrees,  $\sigma_z = 28.6n^{-1/2}$ .

The most common transformation is to take logarithms. "Logging" the observations is appropriate for equalizing their variances in cases where their per cent error is constant; putting  $\sigma_y = c\mu$  in (10.7.1) generates this transformation.

It is important to remember that transformations transform the mean as well as the variance, so that for the transformed variable approximately  $E(z) = f(\mu)$ , and this may help or hinder the analysis: As an example where it helps, suppose that in a pilot-plant experiment to convert a certain chemical substance, which we shall call the "reactant," to a desired product by means of a catalyst we intend to vary the following four factors: (i) the kind of catalyst, (ii) the amount of the reactant, (iii) the contact time of the reactant with the catalyst, and (iv) the temperature of the reaction (controlled by a water jacket or a heating coil). Denote by  $y$  the amount of reactant converted to the product in a single run of the pilot plant. If the set of factor levels in the experiment produces a large

<sup>53</sup> Freeman and Tukey (1950).

## 122 Symptoms and remedies

*null plot*, the plot that is observed when the specified model is correct, might look something like Figure 6.1a. In residual plotting, the user is primarily interested in the shape of the plot, and often the values on the axes are not important. Since the vertical axis is for Studentized residuals, the range (+3, -3) is generally adequate. In Figure 6.1a, for example, the points tend to fall in a horizontal band, without any apparent systematic features.

**Plots against independent variables.** Like the fitted values, the  $r_i$ 's and each of the independent variables are nearly uncorrelated, and systematic features in these plots would suggest model failures that are a function of the independent variable plotted. For example, observing the right-opening megaphone, Figure 6.1b, in a plot of the  $r_i$  against an independent variable may suggest that the residual variance is an increasing function of that independent variable.

**Other plots.** In some problems, plotting  $r_i$  versus other quantities such as case number or time may be of use, as these could indicate failures of a model that are due to ordering or changes over time. These plots are of a different character than plots against  $\hat{y}_i$  or the independent variables, as the  $r_i$ 's may be highly correlated with these extra variables, and very systematic plots may be observed. Sometimes, precisely these systematic features in the plots are of interest.

## 6.2 Heterogeneity of variance

In many practical and theoretical contexts, the assumption that the error variance is constant for each data point (e.g.,  $\text{var}(e_i) = \sigma^2$ ) is uncertain. Commonly, the magnitude of  $\text{var}(e_i)$  will depend upon the magnitude of some other variable, often the response. For example, if an intrinsically positive response varies over a wide range, say from near zero into the thousands, it is intuitively clear that responses near zero usually will be less variable than responses near 1000, since the latter have more "room" to vary than do the former. This can arise in experimental situations in which the control units, without any treatments, have very small response, but addition of a treatment results in a large, but variable, response, or in a drug trial where poorly treated units die immediately, but well treated units live a long, but variable, time.

**Symptoms.** Residual plots like those of Figure 6.1b through 6.1d indicate that the error variance is a systematic function of the quantity plotted

## 6.2 Heterogeneity of variance 123

on the horizontal axis. In the situation described above, the right-opening megaphone pattern of Figure 6.1b should be expected when the horizontal axis is  $\hat{y}_i$ . Also, if  $\text{var}(e_i)$  is strongly related to any good predictor of  $Y$ , say  $X$ , then the plot of  $r_i$  against  $X$  should reveal this as well.

The left-opening megaphone (Figure 6.1c) will occur when small values of the horizontal axis imply large variability. The double outward bow, Figure 6.1d, can arise if the response is a percentage between 0 and 100%. Large or small percentages are less variable than are percentages near 50%.

Generally, residual plots will not be as clear as those in Figure 6.1. The eye will often be influenced by a few points, and heterogeneity of variance may be suspected when the problem is actually an outlier. Also, cases with extreme values of the variable plotted on the  $x$ -axis are often more carefully examined as indicators of heterogeneity. Yet these are generally the points that have the greatest influence in the estimation of parameters (Section 5.2) and will therefore tend to be overfit, that is, have residuals that are relatively small. This problem is partially solved by using the  $r_i$ 's rather than the  $\hat{e}_i$ 's, but the problem remains if the fitted model is incorrect.

**Remedies.** If heterogeneity of variance is suspected, and  $\text{var}(e_i)$  is a known function of some quantity, such as  $\text{var}(e_i) = \sigma^2 z_i$ , where  $z_i$  is known for each case ( $z_i$  may be an independent variable, or some other variable such as time), then weighted least squares is suggested (Section 4.1). As a result, best linear unbiased estimates of the parameter vector can be obtained. Often, however, the heterogeneity of variance can be removed by transforming either the response  $Y$  or one or more of the  $X$ 's or both; usually, transformations to  $Y$  will be applied. These transformations are called *variance stabilizing transformations*.

For almost any relationship between  $\text{var}(e_i)$  and a response, an appropriate variance stabilizing transformation can be found (see Scheffe (1959), Chapter 9 for technical details). In Table 6.1, the common variance stabilizing transformations are listed. The first three ( $Y^{1/2}$ ,  $\log(Y)$ ,  $1/Y$ ), as well as their forms when some of the responses are zero, are appropriate for the right- (or perhaps left-) opening megaphone form, but each is more severe than the one before it. The square-root transformation is relatively mild and is most appropriate when the  $y_i$ 's are counts following a Poisson distribution, usually the first model considered for errors in counts. The logarithm is the most commonly used transformation (the base is irrelevant). It is exactly appropriate when the error standard deviation is a percentage or a proportion of the response (i.e., the error is  $\pm 10\%$ , not  $\pm 10$  units), and  $[\text{var}(e_i)]^{1/2} \propto E(y_i)$ . More will be said about such models in the next section.

124 Symptoms and remedies

Table 6.1 Common variance stabilizers

Transformation	Situation	Comments
$\sqrt{Y}$	$\text{var}(e_i) \propto E(Y_i)$	The theoretical basis is for counts from the Poisson distribution
$\sqrt{Y} + \sqrt{Y+1}$	As above	For use when some $Y_i$ 's are zero or very small; this is called the Freeman-Tukey (1950) transformation
$\log Y$	$\text{var}(e_i) \propto [E(Y_i)]^2$	This transformation is very common; it is a good candidate if the range of $Y$ is very broad, say from 1 to several thousand; all $Y_i$ must be strictly positive
$\log(Y+1)$	As above	Used if $Y_i = 0$ for some cases
$1/Y$	$\text{var}(e_i) \propto [E(Y_i)]^4$	Appropriate when responses are "bunched" near zero, but, in markedly decreasing numbers, large responses do occur; e.g., if the response is a latency or response time for a treatment or a drug, some subjects may respond quickly while a few take much longer; the reciprocal transformation changes the scale of time per response to the rate of response, response per unit time; all $Y_i$ must be positive
$1/(Y+1)$	As above	Used if $Y_i = 0$ for some cases
$\sin^{-1}(\sqrt{Y})$	$\text{var}(e_i) \propto E(Y_i)(1 - E(Y_i))$	For binomial proportions ( $0 < Y_i < 1$ )

*cf example of popl in Tukey*

The reciprocal (or inverse) transformation is often applied when the response is a time of waiting, healing, survival, and so on. Taking reciprocals changes the scale from time per response to responses per unit time. This latter is a rate and it is often more interesting on theoretical grounds than the original measurement.

When repeated measurements are made at each of several values of  $x$ , additional information concerning nonconstant variance is available, since an estimate of a variance at each  $x$  can then be computed. For example, consider again the apple shoot data in Example 4.2. There, for each of the

6.2 Heterogeneity of variance 125

sample days, the mean and standard deviation of the number of stem units observed were recorded. They are plotted in Figure 6.2. From the plot, it appears that the standard deviation is an increasing function of the mean, suggesting the need for transformation of  $Y$ . From Table 6.1, the choices of  $\log(Y)$  or  $Y^{1/2}$  appear to be good candidates. Unfortunately, the original data are not available, so we cannot actually check the transformations and see how well they work.

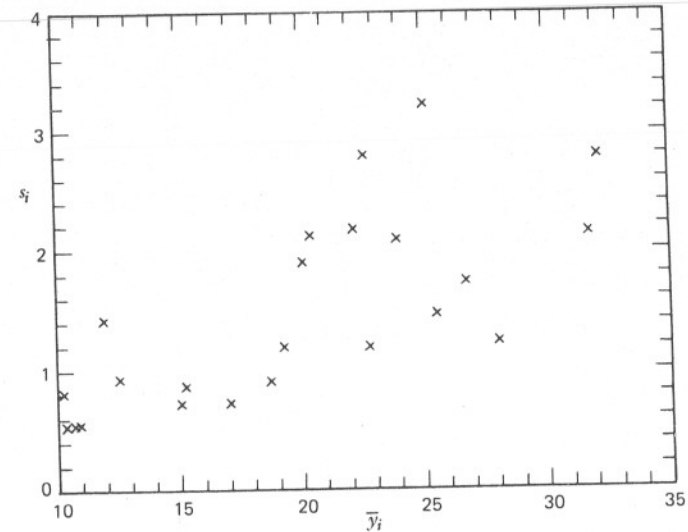


Figure 6.2 Day averages versus standard deviation for apple shoot data.

Table 6.2 Linearizing transformations

Transformation	Simple Regression Form	Multiple Regression Form
$\log Y$	$\log X$	$Y = \alpha X_1^{\beta_1} X_2^{\beta_2} \dots X_p^{\beta_p}$
$\log Y$	$X$	$Y = \alpha e^{\beta X}$
$Y$	$\log X$	$Y = \alpha + \sum \beta_j \log(X_j)$
$\frac{1}{Y}$	$\frac{1}{X}$	$Y = \frac{1}{\alpha + \sum (\beta_j / X_j)}$
$\frac{1}{Y}$	$X$	$Y = \frac{1}{\alpha + \sum \beta_j X_j}$
$Y$	$\frac{1}{X}$	$Y = \alpha + \sum \beta_j \left(\frac{1}{X_j}\right)$