

# Data Transformations to Equalize Variances

Consider observations  $Y_{ij} = \mu_i + \epsilon_{ij} \sim (\mu_i, \sigma_i^2)$  with independent errors  $\epsilon_{ij}$ . Thus, the variance  $\sigma_i^2$  of the random errors  $\epsilon_{ij}$  depends on the treatment  $i$  (or the regressor  $x, \dots$ ), and there is *replication* for the  $i$ th treatment if  $r_i > 1$ , where  $j = 1, \dots, r_i$ .

Finding a transformation of data to equalize the variances of the error variables involves finding some function  $h(y_{ij})$  of the response variables so that  $h(Y_{ij}) = \mu_i^* + \epsilon_{ij}^*$  with  $\epsilon_{ij}^* \sim (0, \sigma^2)$ —namely, so the error variance is constant across treatments.

An appropriate transformation can generally be found if there is a clear relationship between the error variance  $\text{Var}(\epsilon_i) = \sigma_i^2$  and the mean response  $E[Y_{ij}] = \mu_i$ . If the variance and the mean increase together, as suggested by a megaphone-shaped residual plot (or if one increases as the other decreases), then the relationship between  $\sigma_i^2$  and  $\mu_i$  is often of the form

$$\sigma_i^2 = k(\mu_i)^q \quad (1)$$

where  $k$  and  $q$  are constants. In this case, the function  $h(y_{ij})$  should be chosen to be

$$h(y_{ij}) = \begin{cases} (y_{ij})^{1-(q/2)}, & \text{if } q \neq 2 \\ \ln(y_{ij}), & \text{if } q = 2 \text{ and all } y_{ij} \text{'s are non-zero} \\ \ln(y_{ij} + 1), & \text{if } q = 2 \text{ and some } y_{ij} \text{'s are zero.} \end{cases} \quad (2)$$

Here “ln” denotes the natural logarithm, which is the log to the base e, though the log base 10 could also be used. (For more details, see Box, Hunter and Hunter, 1978, page 233).

Usually the value of  $q$  is not known, but a reasonable approximation can be obtained empirically as follows. Substituting the least squares estimates  $\hat{\mu}_i = \bar{y}_i$  and  $\hat{\sigma}_i^2 = s_i^2$  for the parameters  $\mu_i$  and  $\sigma_i^2$ , respectively, in equation (1) and taking the natural log of both sides gives

$$\ln(s_i^2) = \ln(k) + q(\ln(\bar{y}_i)).$$

So, the slope of the line obtained by plotting  $\ln(s_i^2)$  against  $\ln(\bar{y}_i)$  gives an estimate for  $q$ . (This approach requires replication. See KNNL section 3.9 for a numerical search approach which does not.)

The value of  $q$  is sometimes suggested by theoretical considerations. For example, if the normal distribution assumed in the model is actually an approximation to the Poisson distribution, then the variance would be equal to the mean. Then  $q = 1$ , and the square-root transformation  $h(y_{ij}) = (y_{ij})^{1/2}$  is used. The binomial distribution provides another commonly occurring case for which an appropriate transformation can be obtained theoretically. If the  $Y_{ij}$  have a binomial distribution with mean  $mp$  and variances  $mp(1-p)$ , then a variance-stabilizing transformation is

$$h(y_{ij}) = \sin^{-1} \sqrt{y_{ij}/m} = \arcsin \left( \sqrt{y_{ij}/m} \right).$$

When a transformation is found that equalizes the variances, then it is necessary to check or recheck the other model assumptions, since a transformation which cures one problem could cause others. If there are no problems with the other model assumptions, then analysis can proceed as usual, but using the transformed data  $h(y_{ij})$ .