An Analysis of Transformations Revisited, Rebutted

G. E. P. BOX and D. R. COX*

Transformation has long been a powerful tool in developing parsimonious representations and interpretations of data. In 1964 we examined the formal estimation of a suitable transformation. In particular, suppose that a response y is transformed to $y^{(\lambda)}$, where

$$y^{(\lambda)} = (y^{\lambda} - 1)/\lambda \quad (\lambda \neq 0)$$
$$\log y \quad (\lambda = 0),$$

and that we assume provisionally that for some unknown λ, the vector $\mathbf{y}^{(\lambda)} = (y_1^{(\lambda)}, \ldots, y_n^{(\lambda)})$ of n transformed observations satisfies a linear model

$$E(\mathbf{y}^{(\lambda)}) = \mathbf{X}\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is unknown, the errors being independently normally distributed with zero mean and constant variance σ^2. Estimation of λ, $\boldsymbol{\theta}$, and σ^2 can be by Bayesian or maximum likelihood methods.

Bickel and Doksum (1981), in a technically impressive paper, studied in particular the joint estimation of λ and $\boldsymbol{\theta}$, examining consistency and asymptotic variances. They report that the cost of not knowing λ and having to estimate it, can be severe; that ". . . the performance of all Box-Cox type procedures is unstable and highly dependent on the parameters of the model in structured models with small to moderate error variances." That is, the estimates $\hat{\lambda}$ and $\hat{\boldsymbol{\theta}}$ can be highly correlated, so that the marginal variances of the $\hat{\theta}$'s can be inflated by large factors over the conditional variances for fixed λ.

It seems to us that this general conclusion is qualitatively obvious and at the same time scientifically irrelevant.

To illustrate first the obviousness, take as a simple example the comparison of two groups of modest size, the observations $\mathbf{y}$ in group one being near 995 and those in group two being near 1005, the scatters within the two groups being roughly normal with standard deviations close to unity. A parameter θ representing the difference between groups on the y scale is quite precisely estimated to be about 10 y-units. Suppose that the possibility of transformation were contemplated. For a very wide range of λ the function $y^{(\lambda)}$ is very nearly linear in y over the span of the data, and, in particular, unless the sample sizes were very large indeed, it would be quite impossible to distinguish from the data whether y or y^{-1} gave better fit to the standard normal assumptions: if the parameter θ were to refer to a difference on the y^{-1} scale it is quite precisely estimated to be near -10^{-5} y^{-1}-units (or 10^{-5} $y^{(-1)}$-units, where $y^{(-1)} = (1/y - 1)/(-1)$). Thus if the target parameter θ is defined in terms of unknown λ in such a case as this, where λ is poorly determined, the numerical value of θ (in units of y^{λ} or $y^{(\lambda)}$) could be virtually anything.

As to the scientific implications of this, how can it be sensible scientifically to state a conclusion as a number measured on an unknown scale? Surely to know that some effect has magnitude 10 units is without content unless one knows the scale and units in which the effect is defined. To say in the above idealized example that θ, defining the difference between groups, is ill determined because the data establish a wide range of functions as virtually equivalent, seems to be very misleading.

There is, of course, no dispute with Bickel and Doksum over mathematics: the issue is one of scientific relevance. As with any procedure it is necessary to use some common sense in estimating transformations, and in particular (see, e.g., Box, Hunter, and Hunter 1978, p. 241) not to expect this to be possible or relevant when for the particular data and class of transformations in mind the transformation is essentially linear.

Of course the gross correlation effects would be avoided if, following our paper, the investigation had been conducted in terms of

$$z^{(\lambda)} = (y^{\lambda} - 1)/(\lambda \dot{y}^{(\lambda - 1)}), \quad (\lambda \neq 0)$$
$$\dot{y} \log y \quad (\lambda = 0),$$

which takes account of the Jacobian of the transformation. (For the above examples the differences in means for both $z^{(1)}$ and $z^{(-1)}$ would then have been very nearly 10 units.) However, some question of scientific relevance would still remain.

There are numerous aspects of transformations that merit further study. These include in particular the further development of simple ways of assessing *transformation potential*; that is, of providing some more formal measure of the ability of particular data to provide useful information about a class of transformations. Further, a referee has made the perceptive comment that the following issue remains unresolved. Suppose that the parameter of

* G. E. P. Box is Professor, Mathematics Research Center, University of Wisconsin, Madison, WI 53706. D. R. Cox is Professor of Statistics, Department of Mathematics, Imperial College of Science and Technology, Queensgate, London, England SW7 2AZ.

© Journal of the American Statistical Association
March 1982, Volume 77, Number 377
Theory and Methods Section

interest (difference, regression coefficient, etc.) is defined on the data-dependent scale $\hat{\lambda}$; in what circumstances do confidence intervals for these parameters calculated in the "usual" way, as if $\hat{\lambda}$ were preassigned, provide an adequate approximation?

[*Received October 1981. Revised November 1981.*]

REFERENCES

BICKEL, P.J., and DOKSUM, K.A. (1981), "An analysis of transformations revisited," *Journal of the American Statistical Association*, 76, 296–311.

BOX, G.E.P., and COX, D.R. (1964), "An analysis of transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

BOX, G.E.P., HUNTER, W.G., and HUNTER, J.S. (1978), *Statistics for Experimenters*, New York: John Wiley.