IDS/TP 9

NONLINEAR TRANSFORMATIONS AND THE TREATMENT OF STRICTLY NON-POSITIVE VALUES OF VARIABLES

by

T.C.I. Ryan

TECHNICAL PAPER NO. 9

(b) INSTITUTE FOR DEVELOPMENT STUDIES
(a) UNIVERSITY OF NAIROBI
P.O. Box 30197
Nairobi, Kenya

NOVEMBER, 1974

NONLINEAR TRANSFORMATIONS AND THE TREATMENT OF STRICTLY NONPOSITIVE VALUES OF VARIABLES

INTRODUCTION

This brief paper aims to answer the common question: If I have observations on variables (either dependent or independent) that have values less than or equal to zero and I want to do logrithmic transformations, what should I do?

CURVES

Usually the researcher looks at a scatter diagram of his data points and selects a rectilinear or curvilinear function for fittings according to the demonstrated pattern of points. He is constrained in his choice of potential curves by the need to linearise his function if he is to use ordinary least squares.[1] The usual choice is a quadradic, which obviously has limitations if extrapolation is desired and the turning point is in the relevant range. It is often desireable to get a curve that approaches some asymptote such as an exponential curve or an hyperbola. Furthermore transformed data often save degrees of freedom for curvilinear fits.

For easy reference the general shapes of the curves to be considered are given below together with their functional form.

1. Hyperbolas

$$y = \frac{a + bx}{c + dx}$$

equivalently

$$y = \frac{b}{d} + \frac{a - \frac{bc}{d}}{c + dx}$$

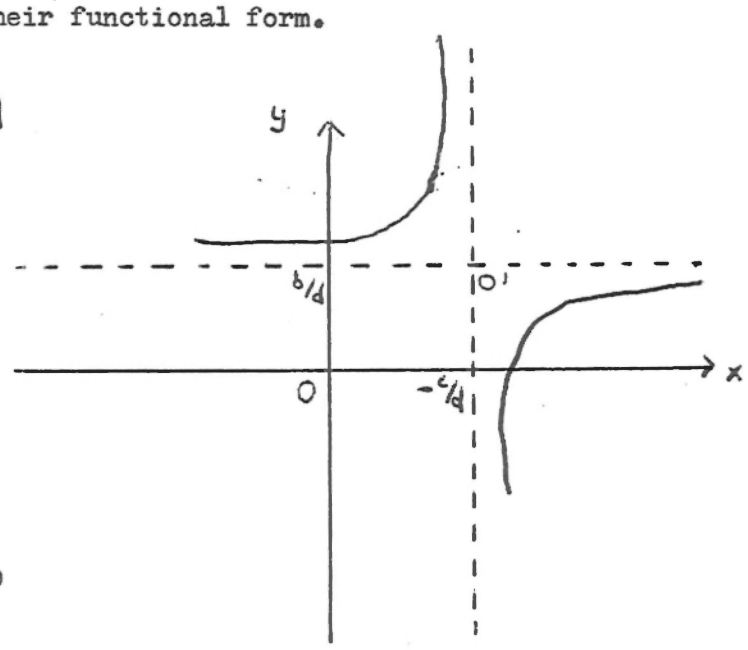
or generally

$$y = k + \frac{m}{c + dx}$$

e.g. $R = PQ$ $\quad a = R$, $c, b = 0$

or $P = \frac{R}{Q}$ $\quad d = 1$

Fig. 1



The branches of the hyperbola are arranged round the axes of symmetry, origin O. The quadrants they are in depend on the signs and sizes of a, b, c and d. For example, if they are all positive and $a > \frac{bc}{d}$ then the curves will be as shown in Figure 1.

---

1. This paper is not intended to cover the econometrics of data transformation, but it is well to note that OLSQ estimators may be both biased and inconsistent if error terms cannot be carefully specified, such that the Gauss-Markov conditions hold on the transformed error terms.

2. Power Curves

$y = k + a(x+c)^b$

e.g. $y = x^2$ c, k = 0, a = 1, b = 2

Fig. 2

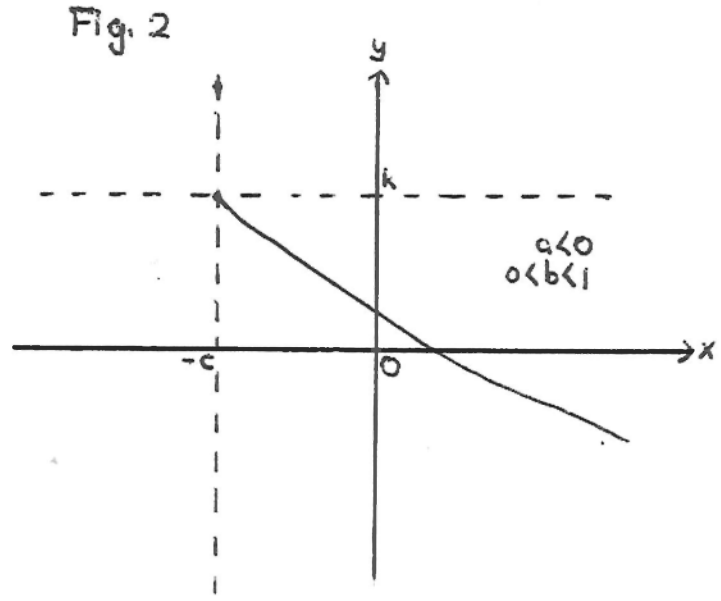


Fig. 3

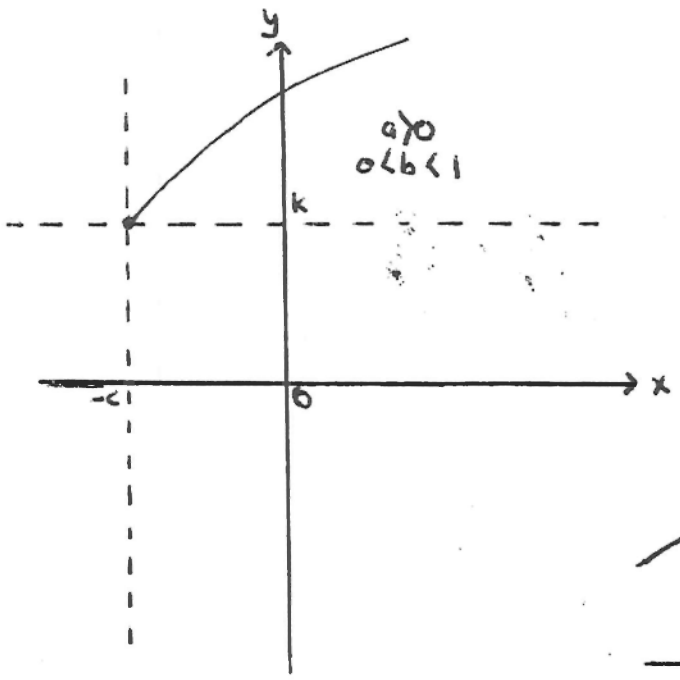
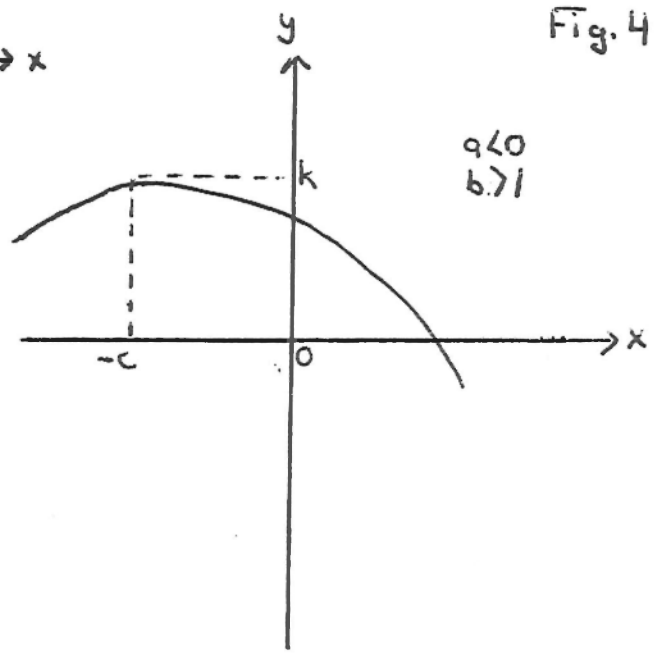


Fig. 4



(x, y) value of the turning point is (-c, k)

3. Exponential Curves

$y = k + ab^{(x+c)}$

alternatively:

$y = k + ae^{b(x+c)}$

e.g. $L_t = L_o e^{nt}$ $\quad c,k = 0$

$a = L_o$

$b = n$

Fig. 5

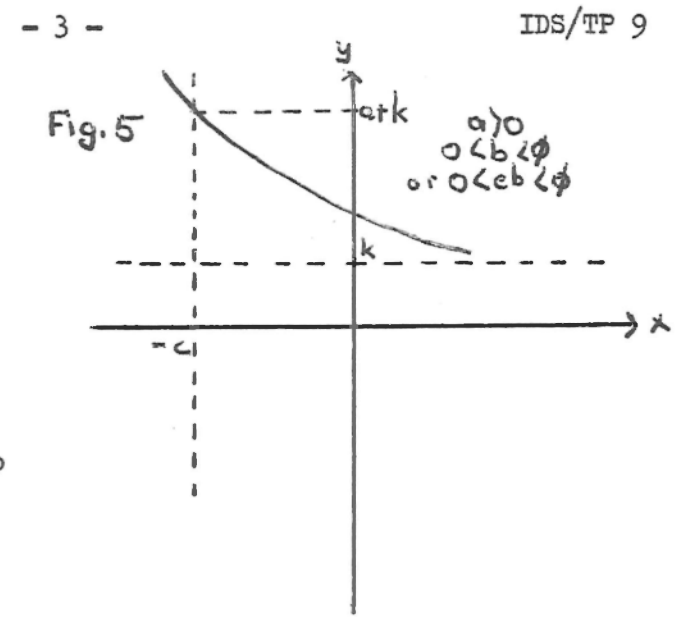


Fig. 6

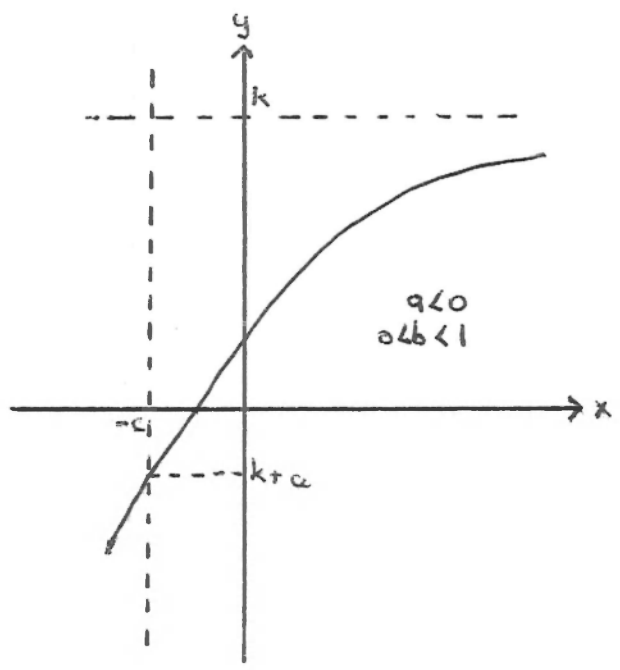
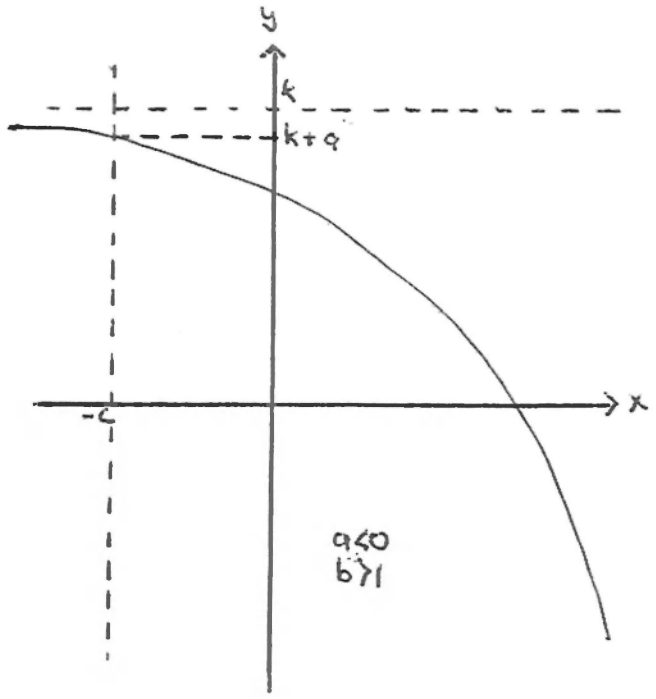


Fig. 7



Each curve has an asymptote bounding it in y.

4. Selecting a functional form

(a) Data points should be plotted on a scatter diagram and the best curve to approximate their pattern should be chosen - ignoring the location of the axes.

(b) Next transform the function for that curve into a rectilinear function:

i) $y = k + \dfrac{m}{c + dx}$ yields $\dfrac{1}{y - k} = \dfrac{c}{m} + \dfrac{d}{m}x$,

ii) $y = k + a(x+c)^b$ yields $\ln(y-k) = \ln a + b \ln(x+c)$,

iii) $y = k + ab^{(x+c)}$ yields $\ln(y-k) = \ln a + (x+c) \ln b$,

$y = k + ae^{b(x+c)}$ yields $\ln(y-k) = \ln a + b(x+c)$.

(c) Taking the values of c and k in equations (ii) and (iii) to be zero, plot the appropriately transformed data. If it appears as approximately a straight line then the selected function is the right one to estimate. Time can be saved if untransformed data of (ii) is plotted on double-log paper and the points for (iii) on semi-log paper - again the points should turn out to be more or less a straight line if those curves are to be used with c and k = 0. (Before doing d), e) or f) read "Treatment of Negative Values" below.)

(d) If the scatter diagram leads you to believe that either equations (ii) or (iii) look to yield the right shapes then experiment with a couple of c's and k's and replot the data; again plotting on log paper can save time - and also help in choosing c and k.

(e) If equation (i) is appropriate, and the horizontal asymptote is not identical with the x axis, then choose a non-zero k ($k<y$), generate a work variable (y-k) and form the transformed variable $\frac{1}{y-k}$, then proceed normally: with this as the dependent variable regress x on it.

(f) If equations (ii) or (iii) are chosen, generate work variables (y-k) and (x+c), do the necessary logrithmic transformations then proceed with the normal ordinary least squares regression analysis on the transformed variables.

THE STATISTICS OF INCREMENTS

It must have been noted that the choice of the increments c and k has been arbitrary so the reasonable question is: What new bias has been introduced into the OLSQ estimators by doing this? The answer is none. In fact the conventional transformations, e.g. $y = ax^b$ yielding $\ln y = \ln a + b \ln x$, are merely special cases where - equally arbitrarily - the increments have been chosen to equal zero.

Since OLSQ estimators aim at minimising the sum of the squared deviations, the exercise is to solve a system of equations derived from partially differentiating - say -

$$G = \Sigma(\ln y - \ln a - b \ln x)^2$$

with respect to ln a and b, to get the usual normal equations

$$\Sigma \ln y = n \ln a + b \Sigma \ln x$$

$$\Sigma \ln y \ln x = \ln a \Sigma \ln x + b \Sigma (\ln x)^2.$$

These yield the estimators $\hat{A}$ for $A \equiv \ln a$ and $\hat{b}$ for b. These can be

proved to be BLUE[2] if the Gauss-Markov conditions hold.

By introducing an increment, say k, in $y=k+ab^x$, then

$$G = \sum(\ln(y-k) - \ln a - x\ln b)^2.$$

This gives rise to the following normal equations when partially differentiated with respect to k, lna and lnb.

$$\sum \ln(y-k) = n \ln a + \ln b \sum x$$

$$\sum x\ln(y-k) = \ln a \sum x + \ln b \sum x^2$$

$$\sum \frac{\ln(y-k)}{y-k} + \ln a \sum \frac{1}{y-k} + \ln b \sum \frac{x}{y-k}$$

This system does not have a unique solution. From the first two equations the usual estimators for lna and lnb ($\hat{A}$ and $\hat{B}$) can be derived and these will always satisfy the third equation, irrespective of what value of k is chosen within the permissable range $k < y$.

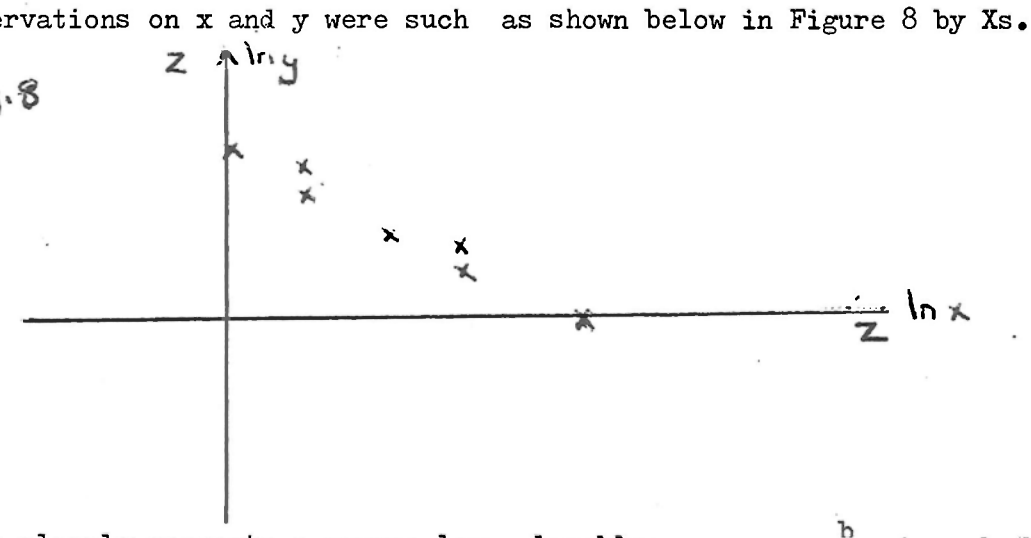
We can therefore conclude that OLSQ estimators of the parameters are LUE when the function is left in its general form, but once a specific k (or c or both) has been chosen then the estimators are BLUE. (As in the case mentioned earlier where both k and c are chosen to equal zero.)

TREATMENT OF ZERO VALUES

Unfortunately some researchers, not wanting to lose some of their data, add arbitrary increments to any observations that are zero since these can neither be transformed in logs or reciprocals.

The bias in the estimated parameters that is introduced by this practice is best demonstrated graphically. Suppose the non-zero observations on x and y were such as shown below in Figure 8 by Xs.

Fig.8



This clearly suggests a curve: $\ln y = \ln a + b\ln x$ or $y = ax^b$, where $b < 0$.

---

2. BLUE, i.e. best linear unbiased estimators. Strictly, in the cases to be considered, these are log-linear.

Should there be zero observations on either variable, no matter how small the increment, it must pull the line towards the 'biased' point, and the better the fit for the other points the more distorting the effect. The points Z_1 in Figure 8 above, would be examples.

Given the argument of the previous section, the best solution is to add the same very small increment to all the observations of the variable if any one of them is zero and a transformation is desired.

There is one particular situation where a transformation is necessary to obtain the Gauss-Markov conditions. This is the case of heteroscedasticity[3]. Strictly this is getting Weighted Least Squares or Generalised Least Squares. If the data when plotted in a scatter diagram look like Figure 9, then we may argue, following Wonnacott and Wonnacott[4], that

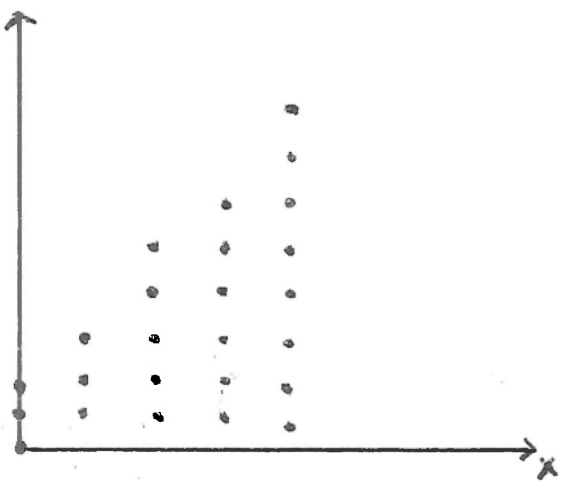
$$\sigma_i = kx_i$$

and that the function to minimise is

$$G = \sum \left(\frac{y}{x} - a\frac{1}{x} - b\right)$$

to obtain $y = \hat{a} + \hat{b}x$

Fig. 9



TREATMENT OF NEGATIVE VALUES

As has probably been noted, Figures 1 (above the b/d asymptote), 3 and 5 (and 4 and 7 if $a > 0$) alone are feasible choices since the k selected must be subtracted from the observed y.

A sufficiently large c value can always be chosen such that, given a k value, all (x+c) will be positive. Since sometimes researchers wish for a curve such as the North-East quadrant of Figure 7 which is ruled

---

3. Error terms are said to be heteroscedastic when they are independently distributed variables but there are differences in the variances of the distributions associated with different values of the independent variables.

4. R.I. Wonnacott and T.H.Wonnacott, Econometrics, New York, John Wiley & Sons, 1970, pp. 132-5.

out, they can obtain it by transforming all their x observations into -x's, then by selecting a suitable c they can calculate the North-West quadrant of Figure 3, i.e. they estimate $y = a(c-x)^b - k$.

By suitable manipulation of c and k in the feasible choices there is no reason to discard any data or to introduce any new biases.

CONCLUSION

This paper has argued that there are a range of curve shapes that are easily tractable for computing which may give better fits to data than the more familiar choices.

It has also shown a satisfactory way of treating strictly non-positive values of variables where logrithmic or reciprocal transformations are desired.

APPENDIX

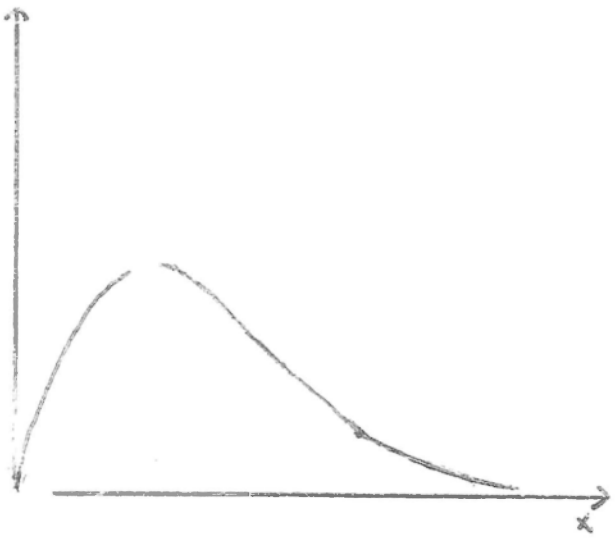
Should plotted data yield a curve such as Figure 10, then rather than fit a quadratic, which imposes symmetry about the peak, it is better to fit

$$y = ax^b e^{cx}$$

which gives the linear transformation

$$\ln y = \ln a + b \ln x + cx.$$

Fig. 10



A NOTE OF WARNING

Since the estimators that are calculated in logrithmic transformations are Maximum Likelihood Estimators, their attractive properties, efficiency and consistency, depend heavily on their being generated from a large sample. They may not be unbiased if the sample size is small for the variance to be zero - the value it approaches asymptotically.