Transform or Link?

Harold V. Henderson
Statistics Section, Ruakura Agricultural Centre, Hamilton, New Zealand

and

Charles E. McCulloch
Biometrics Unit, Cornell University, Ithaca, New York 14853, USA

BU-1049-MA June 1990

TRANSFORM OR LINK?

Harold V. Henderson
Statistics Section, Ruakura Agricultural Centre, Hamilton, New Zealand

and

Charles E. McCulloch*
Biometrics Unit, Cornell University, Ithaca, NY 14853, USA

Abstract

We present data, from fitting lactation curves, which dramatically demonstrates the difference between transforming the dependent variable in regression versus directly fitting a nonlinear model. It has long been recognized that in some cases these approaches give different results but actual examples demonstrating and explaining the differences are difficult to find.

Traditional approaches to transformations in regression analysis have focussed on whether the entire nonlinear equation, including the errors, can be transformed to linear with an additive error. A more modern approach using generalized linear models considers a linearizing link function on the mean, and it does this separately from distributional assumptions.

Keywords: generalized linear models, weighting, scale

This paper was partially supported by the Mathematical Sciences Institute at Cornell University, the New Zealand Mathematics Society and Hatch grant #151-406. It was produced while H V Henderson was a Visiting Fellow in the Biometrics Unit at Cornell University, July - September 1989. Thanks to cow 7450 for such interesting data and to Dr. Arnold M. Bryant for permission to use it.

*To whom correspondence should be addressed.

Introduction

We present data which dramatically demonstrates the difference between transforming the dependent variable in regression versus directly fitting a nonlinear model. It has long been recognized that in some cases these approaches give different results but actual examples demonstrating and explaining the differences for class use are difficult to find.

Traditional approaches to transformations in multiple regression analysis have focussed on whether the entire nonlinear equation, including the errors, can be transformed to linear with an additive error. Such models are called "transformably linear" by Bates and Watts (1988, p.34) in preference to the phrase "intrinsically linear" sometimes used, for example, by Draper and Smith (1981, p.459). Transformably linear models are discussed in standard text books, for example, Snedecor and Cochran (1989, p.403), in recent books on nonlinear regression including Bates and Watts (1988, section 2.1.1), Seber and Wild (1989, sections 1.7, 2.8), and in the encylopedia article on transformations by Atkinson and Cox (1988, p.313).

A more modern approach using the generalized linear model (Nelder and McCullagh, 1989), has emphasized the mean and the "link" function to the mean, i.e., the function which makes the mean a linear function of the explanatory variables. The distribution of the "errors" is then specified separately by making distributional assumptions about the dependent variable, perhaps using a different (and not necessarily the identity) transformation from that used to achieve linearity. And so we may consider transforming (the mean) to linearity separately from transforming to normality or to some particular distribution. Generalized linear models may

be easily fitted by maximum likelihood in GLIM and Genstat which use the algorithm of iteratively reweighted least squares.

The applicability of generalized linear models was extended by Wedderburn (1974) with the notion of quasi-likelihood. This drops the distributional assumptions but retains the corresponding assumptions on the variance and generalizes the well-known connection between ordinary least squares estimation and maximum likelihood estimation when assuming normality. Carroll and Ruppert (1988) detail the equivalence of generalized (weighted) least squares to maximum likelihood in the class of generalized linear models.

This paper is largely expository. We find that using the framework of generalized linear models helps clarify the issues and provides a paradigm for thinking about and discussing them. In particular, we gain insight from the weights used in the iteratively reweighted least squares fitting algorithm used to fit generalized linear models. This we do in the example on average daily fat yield from a cow that follows (Table 1).

Table 1.
Average daily fat yield (kg/day) for each of 35 weeks

.31	.39	.50	.58	.59	.64	.68	.66	.67	.70
.72	.68	.65	.64	.57	.48	.46	.45	.31	.33
.36	.30	.26	.34	.29	.31	.29	.20	.15	.18
.11	.07	.06	.01	.01					

Our data set

We consider fitting the nonlinear equation

$$f(t) = a\, t^b\, e^{ct} \qquad (1)$$

to describe lactation data of dairy cows, where f(t) represents the production of milkfat in week t, and a, b, c are the parameters to be estimated. In the dairy science literature (1) is known after its proposer, Wood (1967). The model has been justified from a biometrical viewpoint in Wood (1977) and has been widely applied, see for example Congleton and Everett (1980). For a review of the history and the variety of models proposed for lactation curves see France and Thornley (1984, pp.220-232) and Grossman and Koops (1988).

Transformably linear: Log transform

Historically, for computational simplicity, Wood's lactation curve was (and is even today is often) fitted by linear regression on the log transformed scale, with (implicit, if not explicit) additive error on the log scale.

$$\log y_t = \log f(t) + e_t = \log a + b \log t + c\, t + e_t \qquad (2)$$

where y_t is the observed production at week t. The usual assumptions on the errors, e_t, for regression are that they are independently, identically distributed with a Gaussian (Normal) distribution with mean 0 and variance σ^2, written as $e_t \sim$ iid $N(0,\sigma^2)$.

This implicitly assumes that (1) has multiplicative error:

$$y_t = a\, t^b\, e^{ct}\, e^{e_t} \qquad (3)$$

Distributional assumptions for multiplicative errors will be discussed subsequently.

Additive (Gaussian) error

Instead, we may assume an additive Gaussian error

$$y_t = at^b e^{ct} + \varepsilon_t \qquad (4)$$

with $\varepsilon_t \sim$ iid $N(0,\tau^2)$, and fit the model, at least for now, using nonlinear regression on the natural scale.

We now illustrate the differences between these error assumptions when fitting Wood's curve to lactation data for a (well chosen) cow from a trial at Ruakura. Coby and Le Du (1978) provide a less dramatic example. That the transformably linear (2) procedure can give poor fits on the natural scale is graphically evident in figure 1 when compared with the fit with additive errors. Figure 2 shows the same fitted curves on a log scale.

Generalized linear model: Log link

In the generalized linear model approach, the mean and distribution of y_t are considered separately. If we assume that equation (1) represents the mean of y_t then a log link would be used, i.e., the log of the mean μ_t is linear. Adjoining this with a distributional assumption gives the model, for example:

$$\begin{aligned} &\log \mu_t = \log a + b \log t + c\, t \\ &y_t \sim \text{independent } N(\mu_t, \sigma^2). \end{aligned} \qquad (5)$$

This distributional assumption gives a model completely equivalent to (4), a point that has yet to be exploited in the lactation curve literature. France and Thornley (1984, Chapter 11) discuss finding approximations for parameters in readiness for nonlinear regression, which is unnecessary with generalised

linear models. Note that the transformably linear approach (2) does not model the mean since the mean on the original scale for (2) is not given by equation (1).

Weighted Regression

The similarities and differences between these models and others can be easily understood in the context of weighted or iteratively reweighted least squares. For example, if we believe model (2), then the standard deviation is constant on the log scale, but proportional to the mean on the original scale (Box, Hunter, and Hunter, 1978, p.233). Thus, an alternative approach to analyzing (2) would be to fit a weighted nonlinear regression with weights proportional to μ_t^{-2}. Since μ_t is unknown, we would in practice use weights equal to y_t^{-2} or we would use iteratively reweighted nonlinear least squares, each time using weights equal to the current estimate of μ_t^{-2}.

Likewise, if we believed that the standard deviations were constant on the original scale, but we wished to analyze log y_t , then we would use weights proportional to μ_t^2 or y_t^2. This weighted regression strategy using y_t^2 is actually the first step in the iterations used in GLIM and GENSTAT to fit model (5). Different weights are applied to each observation depending on what assumptions on the distribution or variance are made. Thus we can see that the extent to which the additive and multiplicative fits will differ will depend on the range of the weights in the data. Our cow has a multiplicative range (largest divided by smallest y_t) of 70 and thus has a 5000-fold difference in the weights. No wonder the fits are quite different!

Multiplicative errors

In generalized linear models the distribution of y_t is a member of the exponential family, so offers much more flexibility than does the normality imposed, or assumed, by the regression on transformed data.

This flexibility can be exploited by specifying a gamma distribution, which implies constant coefficient of variation, on y and a log link. This gives multiplicative errors as discussed in McCullagh and Nelder (1989, Chapter 8). Wood's lactation curve is specified in this way as

$$\log \mu_t = \log a + b \log t + c\, t \qquad (6)$$
y_t has independent Gamma distribution

Its fit to the lactation data is shown in Figure 3 and is similar to the fit obtained from regression on log y_t in (2). In general, as is well known, assuming y_t has a Gamma distribution is similar to assuming that it is lognormal as we did in (2) where log y_t is assumed to be Gaussian. The subtle differences are discussed by Firth (1988) and McCullagh and Nelder (1989, p.286). Note that neither model (2) nor a Gaussian model with constant coefficient of variation is a generalized linear model since they do not model the mean (Carroll and Ruppert, 1988, p.21). Discussion of the choice between the gamma model with a log link and the lognormal model is also given in Atkinson (1982, pp.18-22). In this connection, he reports the critical reanalysis by McCullagh (1980) of data on learning patterns of chimpanzees in a two-way layout.

The estimated coefficients for the 5 methods of fitting Woods's curve we have discussed are in displayed in Table 2. Three methods are variants on fitting multiplicative errors and the other two are variants on fitting additive errors.

Table 2.
Coefficients for Wood's curve $f(t) = at^b e^{ct}$ to observed data y_t from different fitting methods

Log link or transform	Distribution on y_t	weights	Maximum Likelihood Estimates log a	b	c
link	Gaussian	$\propto y_t^{-2}$	-2.60	2.24	-0.274
transform	lognormal	$\propto 1$	-1.56	1.39	-0.186
link	gamma	$\propto 1$	-1.39	1.17	-0.157
link	Gaussian	$\propto 1$	-1.41	1.00	-0.128
transform	lognormal	$\propto y_t^2$	-1.30	0.89	-0.114

Choice of scale

The choice of scale for analysis is an important consideration in model selection. In our example the choice has been between y and log y. What scale should data be analysed on? It depends! It depends on the purpose for which the scale is to be used. Cox and Snell (1981), in a very lucid discussion, introduce the term "extensive" for a variable which is physically additive in a useful sense: its mean value has a physical interpretation.

The mean milkfat production clearly has a physical interpretation as the total production over the lactation. But the log of the production does not have this property, nor does any other nonlinear transform. This is an argument

against the log transform of the data since we would not be modelling the mean. Instead it would advocate use of a log link on the mean with a Gamma, (6), or Gaussian distribution, (4), on y_t.

In discussing this McCullagh and Nelder (1989, p.22) quote from the preface of Jeffreys (1961): 'It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude'.

References

Atkinson, A.C. (1982), Regression diagnostics, transformations and constructed variables. *Journal of the Royal Statistical Society,* B, 44, 1-36.

Atkinson, A.C. (1985), *Plots, Transformations and Regression.* Clarendon Press, Oxford.

Atkinson, A.C. and Cox, D.R. (1988), Transformations. *Encylopedia of Statistical Sciences,* Volume 9. (Kotz, S. and Johnson, N.L. editors), John Wiley & Sons, New York.

Baker, R.J. and Nelder, J.A. (1978), *The GLIM System.* Release 3, *Generalized Linear Interactive Modelling.* Numerical Algorithms Group, Oxford.

Batts, D.M. and Watts, D.G. (1988), *Nonlinear Regression Analysis and its Applications.* John Wiley & Sons, New York.

Box, G.E.P. and Cox (1964), An analysis of transformations. *Journal of the Royal Statistical Society,* B, 26, 211-52.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978), *Statistics for Experimenters, An Introduction to Design, Data Analysis, and Model Building.* John Wiley & Sons, New York.

Carroll, R.J. and Ruppert, D. (1988), *Transformation and Weighting in Regression.* Chapman and Hall, New York.

Coby, J.M. and Le Du, Y.L.P. (1978), On fitting curves to lactation data. *Animal Production,* 26, 127-133.

Congleton, W.R., Jr. and Everett, R.W. (1980), Error and bias in using the incomplete gamma function. *Journal of Dairy Science,* 63, 101-108.

Cox. D.R. and Snell, E.J. (1981), *Applied Statistics, Principles and Examples.* Chapman and Hall, London.

Dobson, A.J. (1983), *An Introduction to Statistical Modelling.* Chapman and Hall, London.

Draper, N.R. and Smith, H. (1981), *Applied Regression Analysis,* Second Edition. John Wiley & Sons, New York.

Firth, D. (1988), Multiplicative Errors: Log-normal or Gamma? *Journal of the Royal Statistical Society,* B, 50, 266-268.

France, J. and Thornley, J.H.M. (1984), *Mathematical Models in Agriculture*. Butterworths, London.

Grossman, M. and Koops, W.J.(1988), Multiphasic analysis of lactation curves in dairy cattle. *Journal of Dairy Science,* 71, 1598-1608.

Jeffreys, H. (1961), *The Theory of Probability*. (Third Edition), Clarendon Press, Oxford.

McCullagh, P. (1980), A comparison of transformations of chimpanzee data. *GLIM Newsletter,* 2, 14-18.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*. (Second edition), Chapman and Hall, New York.

Raaijmakers, J.G.W. (1987), Statistical analysis of the Michaelis-Menten equation. *Biometrics*, 43, 793-803.

Seber, G.A.F. and Wild, C.J. (1989), *Nonlinear Regression*. John Wiley & Sons, New York.

Snedecor, G.W. and Cochran, W.G. (1967), *Statistical Methods*. (Sixth Edition), Iowa University Press.

Wedderburn, R.W.M. (1974), Quasi-likelihood functions, generalized linear models, and the Gauss-Newton Method. *Biometrika*, 61, 439-447.

Wood, P.D.P. (1967), Algebraic model of the lactation curve in cattle. *Nature*, 216, 164-165.

Wood, P.D.P. (1977), The biometry of lactation. *Journal of Agricultural Science,* Cambridge, 88, 333-339.

Figure 1: Lactation curve: Natural scale

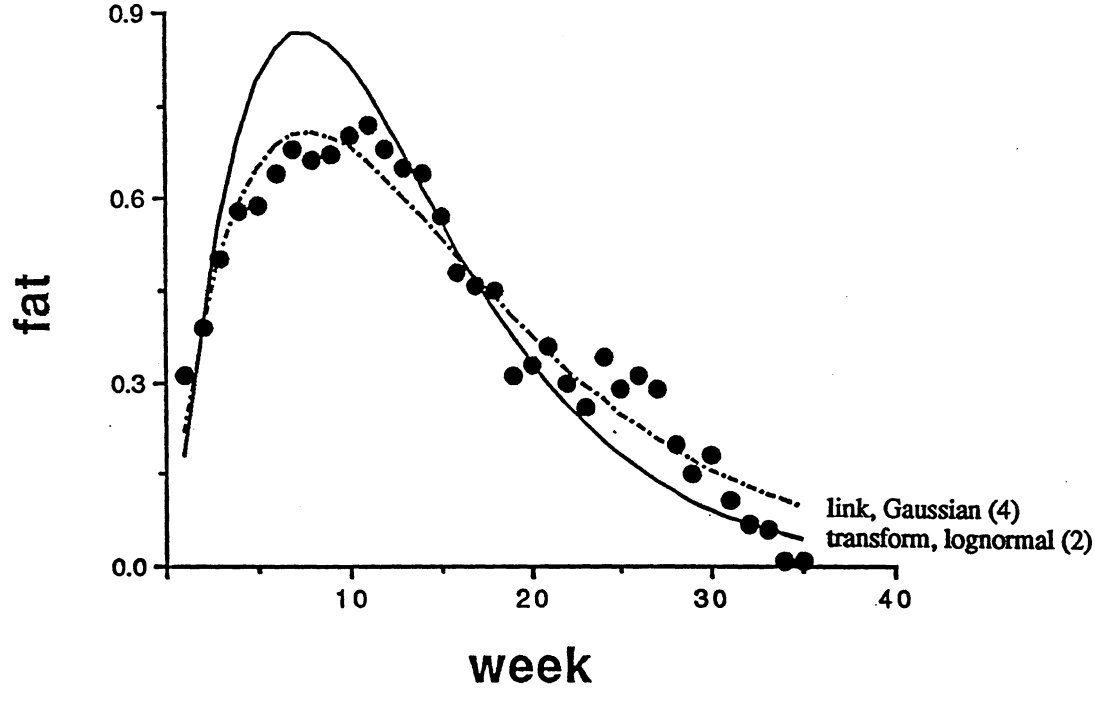

Figure 2: Lactation curve: Log scale

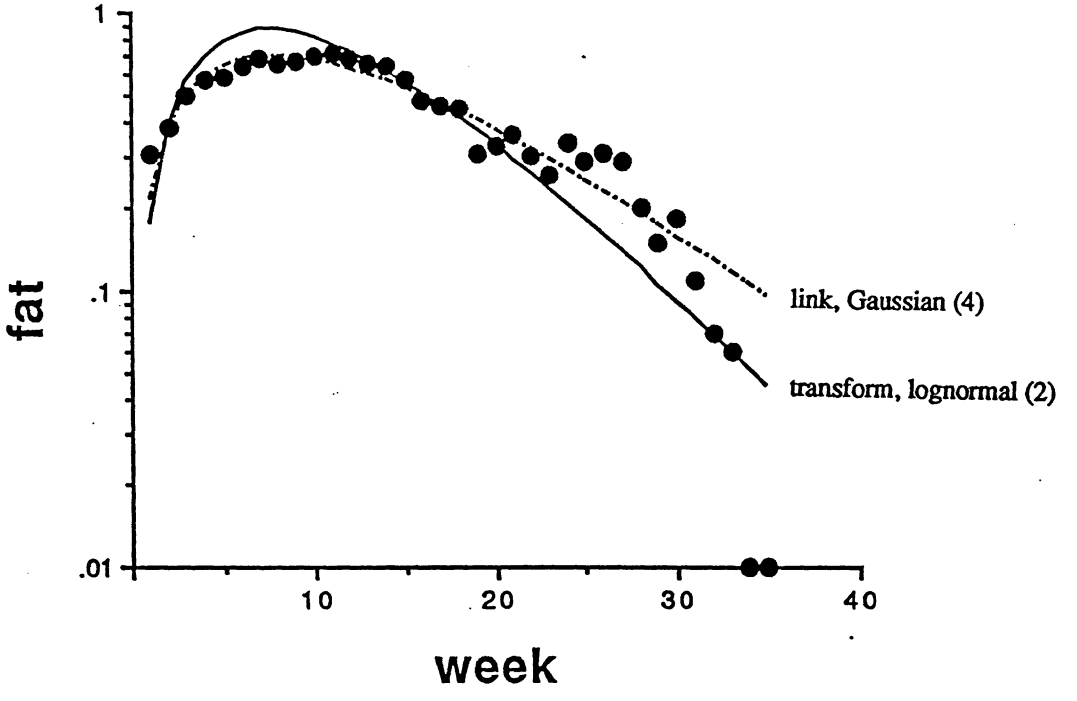

Figure 3: Lactation curve: Natural scale