Behavioural Processes 81 (2009) 216–222

Contents lists available at ScienceDirect

Behavioural Processes

journal homepage: www.elsevier.com/locate/behavproc

Variance matters: The shape of a datum

Michael Davison*, Douglas Elliffe

The University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 16 August 2008
Received in revised form 27 December 2008
Accepted 21 January 2009

Keywords:
Linear regression
Generalized matching
Curve fitting
y variance
x variance

ABSTRACT

In the quantitative analysis of behaviour, choice data are most often plotted and analyzed as logarithmic transforms of ratios of responses and of ratios of reinforcers according to the generalized-matching relation, or its derivatives such as conditional-discrimination models. The relation between log choice ratios and log reinforcer ratios has normally been found using ordinary linear regression, which minimizes the sums of the squares of the *y* deviations from the fitted line. However, linear regression of this type requires that the log choice data be normally distributed, of equal variance for each log reinforcer ratio, and that the *x* (log reinforcer ratio) measures be fixed with no variance. We argue that, while log transformed choice data *may* be normally distributed, log reinforcer ratios do have variance, and because these measures derive from a binomial process, log reinforcer ratio distributions will be non-normal and skewed to more extreme values. These effects result in ordinary linear regression systematically underestimating generalized-matching sensitivity values, and in faulty parameter estimates from non-linear regression to assume hyperbolic and exponential decay processes. They also lead to model comparisons, which assume equal normally distributed error around every data point, being incorrect. We describe an alternative approach that can be used if the variance in choice is measured.

1. Introduction

While we are all aware that every measurement we make has an associated error variance, we generally pay only lip service to this knowledge. We tend to assume, when conducting quantitative analyses of behaviour, that each datum we obtain is a good estimate of a true point that lies on some theoretical function. We know each datum is likely to miss the "real" function, but we assume that our datum is a sample from a normal distribution on the *y* axis around a point on the "real" function and, furthermore, that the variance of that distribution is the same at every point on the *x* axis (the assumption of homoscedasticity). These are assumptions: we do not *know* either of these assumptions to be the case from empirical research, though, for example, Tustin and Davison (1978) attempted to assess the assumption of homoscedasticity of log behaviour ratios across changing log reinforcer ratios, and found no evidence against the assumption. The two assumptions of normality and homoscedasticity are the standard assumptions required for least-squares linear and non-linear regression, and have received a lot of attention in relation to both inferential statistics and regression (see the recent review by Erceg-Hurn and Mirosevich, 2008).

A third assumption made by least-squares regression is that the independent variable (*x*) value for each data point is fixed and has no variance (Davison and McCarthy, 1981). Both of the first two assumptions are likely to be incorrect and, in any case, neither can be sustained without empirical research or theoretical analysis. As we will show, least-squares fitting procedures are not generally robust against violations of these assumptions. The third assumption is simply not met by many, perhaps all, of our regression analyses. Thus, some or all of the assumptions for linear regression are likely to be violated, and the resulting regression parameter values will therefore be inaccurate, or even systematically biased, by such violations.

Fitting a straight line to data in which the *y* axis is a proportional (relative) measure must underestimate the slope of the relation simply because the distributions of data around extreme proportions (close to 0 and 1) are necessarily truncated. This problem is often addressed by transforming the proportional measure to one that is not truncated at the extremes, and for this we often use the logistic transform, $\log(p/(1-p))$, as in generalized-matching analyses. However, such a transform, as we show below, will likely skew data-sampling distributions differently for each value of the dependent variable, and will not result in a true estimate of the slope of the relation. For example, if data are binomially distributed, the assumption of homoscedasticity is not met by logistically transformed proportional data. Thus, with 100 data, 1 standard deviation around a probability of .5 is ±5 responses (45–55), a logistic range from −0.087 to +0.087. With a true probability of .95 (logistic 1.28),

* Corresponding author at: Psychology Department, The University of Auckland City Campus, Private Bag 92019, Auckland 1142, New Zealand.
Tel.: +64 9 373 7599x88540; fax: +64 9 373 7450.
E-mail address: m.davison@auckland.ac.nz (M. Davison).

0376-6357/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.beproc.2009.01.004

1 standard deviation is ±2.18 responses, a logistic range from 1.11 to 1.54. Thus, logistic variances are not homoscadastic with probability changes if the underlying distribution is binomial, and increase as the probability deviates more from .5. Non-homoscedasticity that is unrelated to probability value does not systematically bias regression slope estimates, and simply increases the variance of the slope estimate. But non-homoscedasticity that is systematically and linearly related to the probability value will systematically bias estimates of slope. The latter will occur with binomially distributed data.

The third assumption – that each value of the independent variable is fixed with no variance – *may* be true if we use the arranged value of *x*, but will be untrue of we use the obtained (a sampled) value of *x*. We might suppose that if arranged *x* is a probability, and we expose an animal to each arranged *x* value for very many trials until the mean obtained *x* value equals the arranged *x* value, it will not matter whether the obtained or the arranged *x* was used in regression—both could be taken as fixed with no variance. But this situation, which we discuss further below, is not the end of the story: first, the total *N* trials is likely to be made from a series of sessional samples, containing many fewer trials. Second, if the animal had anything less than perfect memory for all previous events that had occurred at this particular *x*, and no memory for what had occurred for prior, different, values of *x*, the effective *N* may be much smaller than a simple count of the trials. Thus, even when many trials are conducted, variance in obtained, or effective, *x* remains. In a system that learns and changes behaviour as a function of environmental change, such an arranged proportional *x* can never be fixed and have no variance. The wider question is whether an *x* value can be constant with no variance in *any* system that learns.

The above argument implies that simply aggregating *x* over very large numbers of trials, or sessions, to reduce its variance is not a satisfactory solution to our problem. That practice might indeed make ordinary least-squares regression more statistically defensible, but it distances the measure of *x* out of all recognition from the independent variable *x* that we understand to control behaviour. That is, taking few but very large samples may make statistical sense, but is behavioural folly.

These considerations suggest that, if we are trying to discover the quantitative laws underlying behaviour, then we may be going about fitting quantitative models in the wrong way. As a corollary, we may not be obtaining correct information when we statistically select one model against competing models using residual analysis or a model-comparison approach, such as the Akaike information criterion and related criteria (Burnham and Anderson, 2002). The purpose of the present paper is to highlight the problems with current practice, and to offer a new way of fitting data and assessing quantitative models.

1.1. An example

We start with a concrete example that concerns an experiment that arranges a series of different concurrent variable-interval (VI) VI schedules with different reinforcer ratios for two response alternatives. The *x* variables are the logarithms of the *obtained* reinforcer ratios (under the reasonable assumption that behaviour can only be a function of what animals receive, not of what experimenters arrange); the *y* variables are log response ratios, in each experimental condition. Both measures are usually averaged over a number of sessions, often 5, conducted after choice has stabilized. We want to fit the generalized-matching relation (Baum, 1974), which is

$$\log \frac{B_1}{B_2} = a \log \frac{R_1}{R_2} + \log c, \quad (1)$$

where B_1 and B_2 are the response numbers, and R_1 and R_2 are the obtained reinforcer numbers, on the alternatives subscripted 1 and 2. Since this is a linear relation, we use least-squares linear regression—which means that the parameters *a* and log *c* are given by the equation that minimizes the sums of the squares of the deviations of the data points from the straight line (given as $\sum (yi - ypi)^2$, where *i* is an index identifying each data point, y_i is the measured log choice, and y_{pi} is the *y* value for each *x* value predicted by the fitted line). That is, the sum of the squared deviations on the *y* axis is minimized. Does this provide a good estimate of the system that produced the data? No, it does not, because some of the assumptions of linear regression have not been met: the variances in the *y* measures may not be homoscedastic (but see Tustin and Davison, 1978), they may not be normal, and *x* variance has not been taken into account (Davison and McCarthy, 1981). We have fitted a line according to a set of assumptions without demonstrating that we have met those assumptions, so our reported values of sensitivity and bias may be wrong. Moreover, they may be systematically and predictably wrong.

In this particular example, we might be able to estimate some of the things we do not empirically know, and which we have ignored in the regression. First, we can recognize that the choice situation is a binary situation, and that if we were to assume for the purposes of exposition that the binomial distribution applies (this may be an oversimplification because responses usually occur in runs, rather than as individual instances with a particular probability [e.g., Nevin and Baum, 1980; Pear and Rector, 1979]), we could estimate the *y* variance around each datum from $\sqrt{Np(1-p)}$, the standard deviation of the number of responses on one alternative given *N* total responses and an underlying probability *p* of responses on that alternative. In probabilistic terms, this estimated variance is symmetrical about the mean *Np*. But, in order to fit the generalized-matching relation, we have logistically transformed the *p* values as $\log [p/(1-p)]$, which makes the distribution around the $\log (B_1/B_2)$ data point asymmetrical, with a longer tail toward more extreme values than toward less extreme values. This means that, while an equal *number* of sample estimates will come from above and below the "true" $\log (B_1/B_2)$, estimates that are more extreme than the "true" value are likely to deviate from the "true" value by larger amounts. Furthermore, the degree of asymmetry increases as $\log (B_1/B_2)$ deviates further from 0. Fig. 1 shows *y* axis standard deviations around a set of true data on a simple straight line. Because of the logistic transformation of the probability of a B_1 response, and because the datum indicates a higher probability of B_1 relative to B_2, the standard deviation above the datum is greater than the standard deviation below the datum when $y > 0$, and vice versa when $y < 0$.

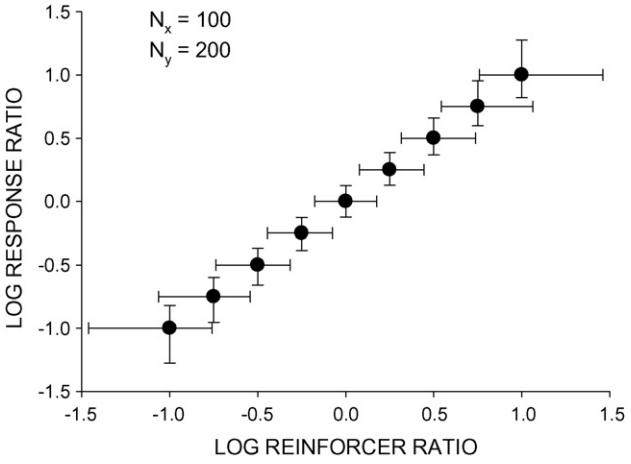

Fig. 1. Two standard deviations around a set of logistically transformed data points derived from homoscedastic binomial distributions around each datum. Log response ratios strictly match log obtained reinforcer ratios.