Biometrika (1980), 67, 3, *pp.* 669–74
Printed in Great Britain

Monotonic transformations to additivity using splines

BY S. WINSBERG

Faculté des sciences de l'éducation, Université de Montréal

AND J. O. RAMSAY

Department of Psychology, McGill University, Montreal

SUMMARY

A class of monotonic integral transformations derived from B-splines is fitted to the independent and dependent variables in multiple regression so that the resulting additive relationship is optimized. The fit is achieved by maximizing a log likelihood criterion with inequality constraints on the parameters. Some examples of the analysis of artificial and real data are offered. An algorithm which works reliably is outlined.

Some key words: Additivity; B-spline; Monotone transformation; Multiple regression.

1. INTRODUCTION

In a typical multiple regression analysis one assumes that there is a linear relationship between dependent and independent variables. Sometimes there is reason to doubt this assumption when the relationship is stated in terms of the original observations, and a non-linear transformation of the variables may help to produce a linear relation among the transformed variables. Moreover, it is often the case, particularly in the behavioural sciences, that some of the variables in the analysis are not measured on an interval scale and, consequently, there is no reason to exclude possible nonlinear transformations.

Transformation of only the dependent variable has received much attention. Anscombe (1961) and Anscombe & Tukey (1963) discussed the analysis of residuals to detect useful transformations. Box & Cox (1964) and Schlesselman (1973) treated one- and two-parameter power transformations. Kruskal (1965) used only rank order information in the dependent variable to transform to additivity in two-way factorial designs. Ramsay (1977) fitted generalized power transformations to both the independent and dependent variables in multiple regression to optimize the resulting additive relationship. In this paper we shall extend Ramsay's (1977) work by considering a class of smooth monotonic transformations which have a more flexible shape than is possible with generalized power transformations.

The data consist of N observations x_{ij} ($i = 1, \ldots, N$; $j = 0, \ldots, L$) on a dependent variable and on each of L independent variables. A dependent variable observation is taken to be x_{i0}. The problem is to find a set of transformations f_j and possibly a constant term such that an objective function Q is optimized with respect to $f_j(x_{ij})$, subject to monotonicity constraints on these transformations. Other constraints may also be required. For example, a theoretically inspired fixed transformation or no transformation at all may be applied to a subset of the variables, or a subset of the transformations may be constrained to be identical.

Since a general nonlinear transformation is not necessarily invariant with respect to changes

of scale and location, one might specify that all variables be normalized in some way before analysis begins. We shall assume that each variable enters the regression equation with a positive weight so that only monotonic increasing transformations need be considered.

2. INTEGRATED B-SPLINES

For a transformation of the form

$$f(x) = \int_{x_0}^{x} v(t)\,dt,$$

for $v(t) \geqslant 0$, the problem of choosing a monotone function reduces to the problem of finding a suitable nonnegative kernel $v(t)$. A desirable feature of the integral transformation is that the monotone function will be smooth even when the corresponding kernel is not. We are looking for a simple family of nonnegative functions which depend on a relatively small number of parameters. The parameters can then be chosen to optimize some fitting criterion.

Linear combinations of B-splines, defined as the divided differences of truncated power functions, make an attractive choice for defining $v(t)$ (de Boor, 1978, p. 108; Smith, 1979). The resulting curve has a very flexible shape, and does not depend on an excessive number of parameters. A spline of the order k is a piecewise polynomial of degree less than k; that is, it is a set of polynomials of degree less than k joined at values of their arguments called knots. The choice of the knot sequence determines the smoothness at a knot. The location of the interior knots is relatively unimportant if the curve to be fitted is reasonably smooth, although it is helpful to have more knots in regions where the nonlinearity is most severe. This insensitivity to knot choice implies that in practice the knots may be chosen *a priori* and held fixed during the analysis. Afterwards one may use some fairly crude technique to improve the choice of knots.

The sequence $B_1, \ldots, B_m$ of B-splines of order k for knot sequence $t_1, \ldots, t_{m+1}$ is a basis for the piecewise polynomials considered as functions on (t_k, t_{m+1}). In order to evaluate the kernel $v(t)$ at a point $\hat{t} \in (t_p, t_{p+1})$, where $k \leqslant p \leqslant m$, one must calculate the k numbers $B_{nk}(\hat{t})$ $(n = p-k+1, \ldots, p)$ by a recursion formula. It should be noted that $B_{nk}(\hat{t})$ is positive for $t_n < t < t_{n+k}$ and zero otherwise. It follows that only k B-splines have any interval (t_p, t_{p+1}) in their support.

Thus, for $t_p \leqslant \hat{t} \leqslant t_{p+1}$,

$$v(\hat{t}) = \sum_{n=p-k+1}^{p} a_n B_{nk}.$$

From the positivity of B-splines and their limited support it follows that they are a well-conditioned basis for approximation. They are also invariant with respect to scale transformations, provided that the knot sequence is similarly transformed. Of particular interest in this application is the ease with which linear combinations of B-splines can be integrated and differentiated.

3. ESTIMATION

In fitting integrated B-splines to the variables in a regression problem, it will be assumed that each variable has a knot sequence $t_{j1}, \ldots, t_{j,m_j+k}$. The transformed dependent variable observations $f_0(x_{i0})$ are assumed to be independently distributed with probability density

function $p\{f_0(x_{i0}) \mid \Sigma_j f_j(x_{ij}) + c\}$. This implies a log likelihood of the form

$$Q = \Sigma_i \log p\{f_0(x_{i0}) | \Sigma_j f_j(x_{ij}) + c\} + \Sigma_i \log D\{f_0(x_{i0})\}.$$

It is necessary to fix the scale of the transformations by requiring an equality condition such as $f_0(1) = 1$. The maximization of Q is then a nonlinear programming problem with $m_0 + \ldots + m_L + 1$ parameters, subject to one linear equality constraint and $M - 1$ linear inequality constraints of the form $a_{jm} \geqslant 0$.

When comparing the log likelihood for this model against that achieved by some specialization such as the conventional linear regression model, it is necessary to settle how many parameters are actually being fitted. At most, this would be $M - 1$. As a rule, some of the parameters will be at the boundary, and they can be thought of as at least partially estimated because the data determined whether or not they were at the boundary. A conservative decision would be to set the number of estimated parameters to $M - 1$ when computing the chi-squared statistic from the two log likelihoods. The asymptotic variance–covariance matrix for the parameter estimates can be computed by the use of the Moore–Penrose inverse as described by Ramsay (1978). For a fixed value of the argument, the asymptotic variance for the transform is given by

$$\text{var}\{f(x) \mid x\} = \left\{\frac{\partial f(x)}{\partial x}\right\}^{\mathrm{T}} \Sigma \left\{\frac{\partial f(x)}{\partial x}\right\},$$

where Σ is the variance–covariance matrix of the parameter estimates. This relation can be extended to yield a variance–covariance matrix for a set of function values as well. The resulting p-values would be rough because of failure of the usual regularity conditions.

One way to solve the nonlinear programming problems with constraints is the penalty function approach described by Fiacco & McCormick (1968, p. 40), in which one solves a sequence of unconstrained problems. The penalty function we chose was

$$P(a_{01}, \ldots, a_{Lm_L}, r) = -2^{-r} \Sigma_j \Sigma_q \log a_{jq} \quad (a_{jq} \geqslant 0).$$

Generally after a few steps in r, changes in the parameter estimates become negligible. An obvious procedure for improving the values of the knots t_{jq} given a solution for the parameters a_{jq} is to carry out a few steepest descent iterations with respect to the knot values, computing the optimal values of the a_{jq}'s after each iteration. Only the interior knots need be considered.

4. Examples

The first example indicates the algorithm's ability to recover a set of monotone continuous functions observed with error. The test problem is

$$\log \frac{x_0}{1 - x_0} = 3 \sin\{\pi(x_1 - \tfrac{1}{2})\} + 6x_2^2 - 3 + \varepsilon_i.$$

The data consist of 50 independently generated random values of x_1 and x_2 sampled uniformly between 0 and 1. The values of x_0 were then generated by applying the appropriate inverse transformation of the sum of the transformed values of x_1 and x_2 and a random Gaussian deviate ε_i was added, having a standard deviation of 20% of the standard deviation of the unperturbed sum of the two transformations. Only a single run was carried out for these data and it was assumed that residuals were $N(0, \sigma^2)$. We chose to use order two B-splines for this analysis because this choice usually produces a good fit while requiring a minimal number of parameters for a given number of interior knots. The knots used for all three variables were

the values 0·0, 0·0, 0·2, 0·5, 0·8, 1·0 and 1·0. The penalty parameter was given values 2, 4, 6 and 8; at which point the change in the log likelihood was less than 0·05.

The estimated transformations are compared to the true transformations in Fig. 1, and conditional asymptotic confidence intervals for each transformation are also shown. The squared multiple correlation coefficient was 0·9952. The log likelihood was 83·1 as compared to a log likelihood of 39·1 when the data were analysed by means of multiple regression. The chi-squared statistic was 88·0 with 12 degrees of freedom, and this is significant at $p < 0·001$. The value of all of the transformations at the origin is precise by definition. In each case the true transformation lies within the confidence interval. The confidence intervals are fairly tight, apart from some ballooning near the first interior knot for the independent variables.

The second example deals with data presented by Durbin & Watson (1951) in which the dependent variable was the log consumption of spirits per capita from 1870 to 1938, and the independent variables were log real income per capita and log relative price of spirits. Both of these variables entered the regression equation with negative coefficients, and as a consequence we changed the sign for each of these. In this example the origin of the scale for each variable was arbitrary, so the minimum and maximum values for each variable were set to zero and one, respectively. Order two splines were used with the knots shown in Table 1.

Table 1. *Knots used with order two splines, for example in Fig.* 2

Dependent variable	0·00	0·00	0·40	0·60	0·75	1·00	1·00
1st independent variable	0·00	0·00	0·25	0·35	0·50	1·00	1·00
2nd independent variable	0·00	0·00	0·25	0·45	0·60	1·00	1·00

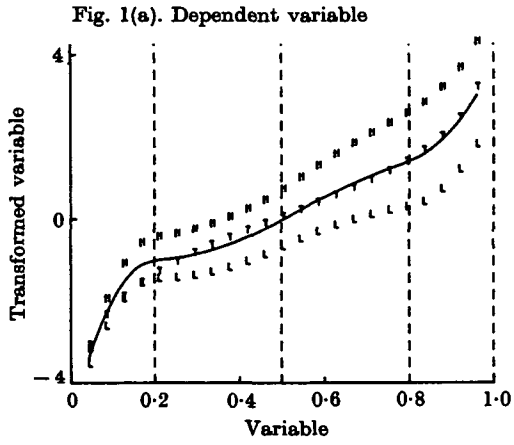
The squared correlation coefficient was 0·9838 compared with a squared multiple correlation coefficient of 0·9558 for the case of no variables transformed. The log likelihood was 119·9 compared with a log likelihood of 86·5 obtained for no transformation. The appropriate chi-squared at 12 degrees of freedom was 66·8 which is significant at $p < 0·001$. The transformations for the dependent and independent variables are shown in Fig. 2 along with conditional asymptotic confidence intervals for each transformation.

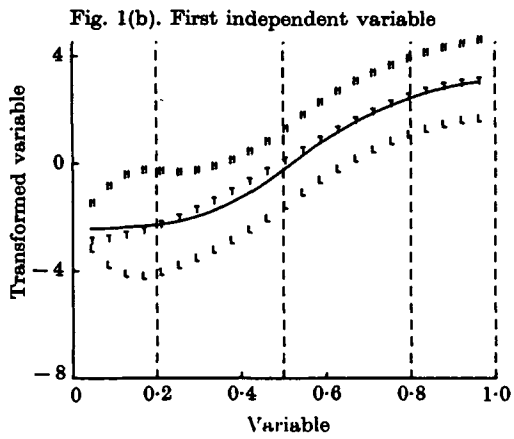
The plateau in the transformation for the log consumption of spirits occurs at values corresponding to the period just before and during the First World War. An examination of the transformations for the independent variables reveals a nonlinear transformation for log real income and some need to transform log price of spirits. These transformations indicate that decreasing income by a fixed factor will have more of an effect when the log real income per capita is high than when it is low. They also indicate that the effect of decreasing relative price of spirits is greater when the relative price is high. The confidence intervals are fairly tightly defined throughout the domain of each variable. These data have been analysed extensively in terms of their linear model residuals without noting the need to transform the variables.

5. DISCUSSION

In general this algorithm has proceeded to solutions reliably, and improvements in the log likelihood were minimal after only a few steps. Although the performance can depend on the choice of knots, it is usually not difficult to make a reasonable initial selection of knots. Moreover, the shape of the solution is usually only slightly affected by the choice of knots.

The need to transform variables to additivity in linear models has been recognized for some time, and practical algorithms have been developed for the dependent variable. Although





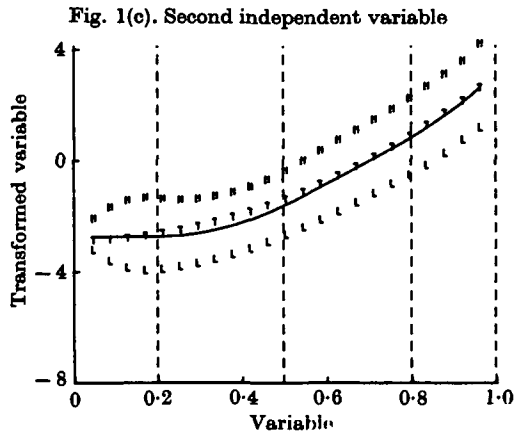
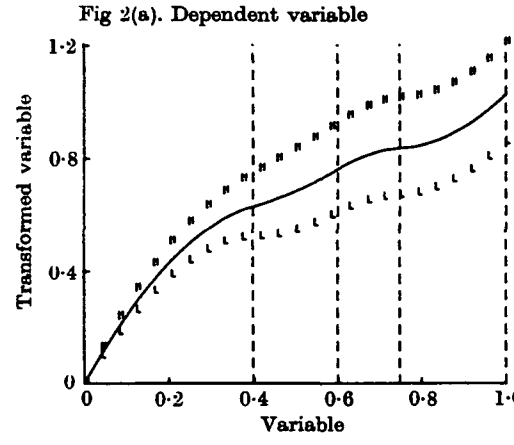
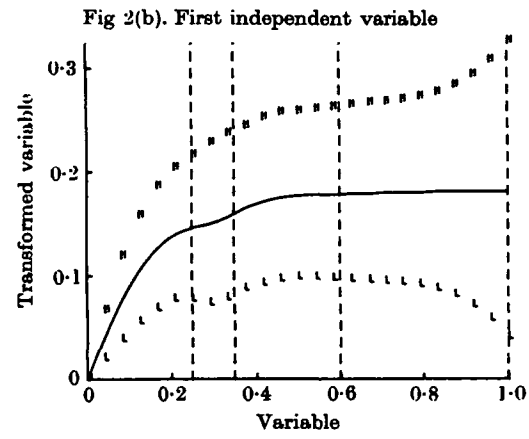


Fig. 1 (left). Fitted transformations of variables for the analysis of artificial data with error; $N = 50$.





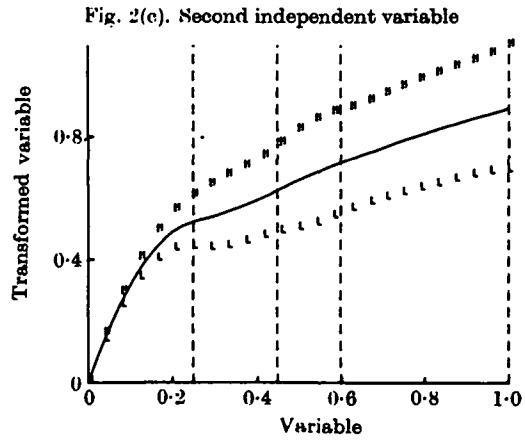


Fig. 2 (right). Transformations of variables for spirits data.
T's, true transformation; H's, upper bound of 95% conditional asymptotic confidence interval; L's, lower bound. Vertical dashed lines, position of knots.

Ramsay (1977) developed an algorithm for monotonic transformation for some or all of the variables to a linear relation, its use introduces bias when the dependent variable is transformed, and it cannot accommodate some types of transformations. The present paper describes a procedure which can be used with confidence when some or all of the variables are transformed, and can accommodate a much wider range of monotone transformations.

REFERENCES

ANSCOMBE, F. J. (1961). Examination of residuals. *Proc. 4th Berkeley Symp.* **1**, 1–36.

ANSCOMBE, F. J. & TUKEY, J. W. (1963). The examination and analysis of residuals. *Technometrics* **5**, 141–60.

BOX, G. E. P. & COX, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.* B **26**, 211–52.

DE BOOR, Ç. (1978). *A Practical Guide to Splines.* New York: Springer-Verlag.

DURBIN, J. & WATSON, G. S. (1951). Testing for serial correlation in least square regression, II. *Biometrika* **38**, 159–78.

FIACCO, A. V. & MCCORMICK, G. P. (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* New York: Wiley.

KRUSKAL, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *J. R. Statist. Soc.* B **27**, 251–63.

RAMSAY, J. O. (1977). Monotonic weighted power transformations to additivity. *Psychometrika* **42**, 83–109.

RAMSAY, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika* **43**, 145–60.

SCHLESSELMAN, J. J. (1973). Data transformation in two-way analysis of variance. *J. Am. Statist. Assoc.* **68**, 369–78.

SMITH, P. L. (1979). Splines as a useful and convenient statistical tool. *Am. Statistician* **33**, 57–62.

[*Received July* 1979. *Revised February* 1980]