STATISTICS IN MEDICINE, VOL. 10, 697–709 (1991)

THE EFFECTS OF TRANSFORMATIONS AND PRELIMINARY TESTS FOR NON-LINEARITY IN REGRESSION

PATRICIA M. GRAMBSCH

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, U.S.A.

AND

PETER C. O'BRIEN

Section of Biostatistics, Department of Health Sciences Research, Mayo Clinic and Mayo Foundation, Rochester, MN 55905, U.S.A.

SUMMARY

Non-linear relationships between two variables are often detected as a result of a preliminary statistical test for linearity. Common approaches to dealing with non-linearity are to (a) make a linearizing transformation in the independent variable or (b) fit a relationship that is non-linear in the independent variable, such as including a quadratic term. With either approach, the resulting test for association between the two variables can have an inflated type I error. We consider testing the significance of the quadratic term in a quadratic model as a preliminary test for non-linearity. Using simulation experiments and asymptotic arguments, we quantify the type I error inflation and suggest simple modifications of standard practice to protect the size of the type I error. In the case of quadratic regression, the type I error will be increased by roughly 50 per cent. The simple strategy of appropriately correcting the α-level is shown to have minimal loss of power if the relationship is truly linear. In the case of a linearizing transformation, the impact on the type I error will depend on the values of the independent variable and on the set of potential linearizing transformations considered. Simulation results suggest that a procedure which adjusts the test statistic according to the results of the preliminary test may offer adequate protection.

INTRODUCTION

In testing for association between two continuous variables, one typically assumes the simple linear regression model with independent and normally distributed errors, $Y = \alpha + \beta X + e$, and performs the usual t test of the null hypothesis: $\beta = 0$. In practice, this approach may be modified if any of the following circumstances is observed or suspected:

1. the distribution of the error term is non-normal;
2. the error terms are not independent;
3. the variance of e varies with X, or
4. the association between Y and X is non-linear.

In this paper, we suppose that the error term is approximately normally and independently distributed with constant unknown variance σ^2. Our concern centres on the situation in which the association between X and Y appears to be non-linear.

0277–6715/91/050697–13$06.50

Received December 1989
Revised August 1990

Many solutions to this problem have been proposed. Two that are commonly used are:

1. Fit a relationship not linear in X, $Y = g(X) + e$. A quadratic polynomial in X is often appropriate, or
2. Make a linearizing transformation on X of the form $X' = f(X)$ where $f(X)$ is any function such that the association between Y and X' is approximately linear and fit the linear model $Y = a + bX' + e$.

In this paper we examine these solutions to the problem of non-linearity. Because they are frequently used in practice, it is important to be aware of their characteristics. It is not our purpose to argue the relative merits of these strategies against others that have been proposed. Our goal is to evaluate the effect of commonly used linearizing transformations and non-linear relationships, particularly quadratic regression, on the overall size of the test for association between X and Y.

Actual practice may rely on a combination of prior information, visual inspection of the data and formal statistical tests to detect non-linearity. However, in order to obtain theoretical results and to perform simulations, we found it necessary to use one simple statistical test for detecting non-linearity. The test for non-linearity to be considered here is the standard F-test of statistical significance for the quadratic term in the second-degree polynomial model

$$Y = \alpha + \beta X + \gamma X^2 + e.$$

There is some arbitrariness in this choice because there is no standard test of non-linearity for regression. We choose this test because of its extensive use in statistical applications, and because it is also recommended as a test for non-linearity in popular textbooks.[1,2] Although the test of the quadratic term is not sensitive to every possible form of non-linearity, it may be expected to be reasonably sensitive to a broad class of non-linear relationships that commonly occur. In quadratic regression the slope of the tangent is linear in X. Thus, testing the quadratic term should prove to be useful in detecting non-linear relationships where the slope of the tangent changes monotonically with X.

We consider the properties of procedures, where our chosen preliminary test for non-linearity is performed and the linear model is abandoned for another when the test exceeds a critical value. We evaluate as an alternative to simple linear regression first the use of quadratic regression, and then the use of some commonly used linearizing transformations in the next two sections. In the fourth section we apply the linearizing transformation and quadratic regression procedures to a real data set. We conclude with some practical suggestions and discuss the relationship of our results to other transformations in regression (the Box–Cox transformation and the ACE algorithm) and extensions of some standard two-sample test procedures.[3]

QUADRATIC REGRESSION

Typically, non-linearity is initially detected by visual inspection of the scatter diagram. One often assesses this visual impression more formally by fitting the quadratic model

$$Y = \alpha + \beta X + \gamma X^2 + e$$

and testing the null hypothesis of linearity (that is, $\gamma = 0$). Let α_P denote the critical level specified for this preliminary test. If the P value is less than α_P, one bases the overall test for association on the 2 degrees of freedom (d.f.) F-test for $\beta = \gamma = 0$. Conversely if the P value exceeds α_P, one uses the 1 d.f. test for association based on simple linear regression. Let α_F denote the critical level for the final test of association between X and Y, whether based on one or two degrees of freedom;

then α_F gives the nominal level for the procedure. One can qualitatively predict the impact of the preliminary test on the true size of the type I error. Testing $\gamma = 0$ provides an indication of association for X and Y, so the entire procedure involves more than one test of the overall relationship. As a result, the size of the type I error is increased, relative to the nominal level α_F.

To quantify the effects of such preliminary testing in large samples, we consider initially the case in which σ^2 is known. Without loss of generality, we can reparameterize the model to:

$$Y_i = a + bT_i + cZ_i + e_i \tag{1}$$

where T and Z are orthogonal linear and quadratic contrasts, respectively. Using this device, we show in the Appendix that

$$Pr\{\text{type I error}\} = (1 - \alpha_P)\alpha_F + \int_{q_P}^{\infty}\left[\int_{q_2 - y}^{\infty} g(u)\,du\right] g(y)\,dy \tag{2}$$

where q_P and q_2 are the critical values for the preliminary test and the final 2 d.f. test, respectively, and $g(\cdot)$ is the density function for a chi-squared distribution with one degree of freedom. We have calculated these probabilities by numerical integration (Table I). The effect of preliminary testing is to increase the size of the test by approximately 50 per cent over the nominal level, and this effect is remarkably constant over the range of values of α_P and α_F considered. Thus, to achieve statistical significance at a specified α level, a nominal P-value of approximately $\alpha/1{\cdot}5$ is required. We will refer to this correction for test size as the '50 per cent'.

To evaluate the adequacy of the '50 per cent rule' in the more realistic setting of σ^2 unknown, we performed a simulation study, consisting of 1000 replicated experiments for each of several values of α_F and α_P. In each experiment, 20(X, Y) pairs were used with X taking on the integer values between 1 and 20 and Y being independent standard normal deviates. The results in Table I indicate that the 50 per cent rule worked quite well throughout the range of values of α_F and α_P considered.

Although the 50 per cent rule will enable the experimenter to control the size of the test, it will result in a loss of power relative to the optimal test when the relationship between X and Y is linear. To assess power, we consider again the orthogonalized model (1), assuming $\sigma^2 = 1$ (known) and compare three procedures:

1. Fit the linear model and perform a 1 d.f. test of $b = 0$.
2. Fit the quadratic model and perform a 2 d.f. test of $b = c = 0$.
3. Perform a preliminary test of non-linearity by fitting the quadratic model and performing a 1 d.f. test of $c = 0$. If significant, perform a 2 d.f. test of $b = c = 0$. If not, fit the linear model and perform a 1 d.f. test of $b = 0$.

For a linear relationship ($Y = a + bT + e$), the power of each of these procedures is a function of the non-centrality parameter $\lambda = b^2$.

The first procedure has power (β_1) equal to the probability that a chi-squared random variable with 1 d.f. and non-centrality parameter λ will exceed q_1, the critical value for the final 1 d.f. test. The power of the second procedure (β_2) equals the probability that a chi-squared random variable with 2 d.f. and non-centrality parameter λ will exceed q_2. If $\hat{b}$ and $\hat{c}$ represent the least squares estimates of b and c, the power for the third procedure is given by

$$\beta_3 = Pr(\hat{b}^2 > q_1 | \hat{c}^2 \leqslant q_P)\,Pr(\hat{c}^2 \leqslant q_P) + Pr(\hat{b}^2 + \hat{c}^2 > q_2 | \hat{c}^2 > q_P)\,Pr(\hat{c}^2 > q_P).$$

Simplifying, the power equals

$$\beta_1(1 - \alpha_P) + Pr(\hat{b}^2 + \hat{c}^2 > q_2 | \hat{c}^2 > q_P)\alpha_P.$$

Table I. The size of the type I error in quadratic regression with a preliminary test for non-linearity

α_P	α_F	Type I error σ^2 known	Type I error σ^2 estimated	'50 per cent rule' $1{\cdot}5 \times \alpha_F$
0·25	0·01	0·016	0·016	0·015
0·10		0·017	0·014	
0·05		0·016	0·015	
0·25	0·025	0·038	0·037	0·0375
0·10		0·039	0·027	
0·05		0·039	0·035	
0·25	0·05	0·073	0·087	0·075
0·10		0·075	0·081	
0·05		0·074	0·069	
0·25	0·10	0·139	0·133	0·15
0·10		0·143	0·126	
0·05		0·137	0·135	
0·25	0·25	0·327	0·341	0·375
0·10		0·324	0·323	
0·05		0·288	0·288	

Table II. Power for the three regression tests at level 0·05 (linear model)

	Method 1	Method 2	Method 3
λ	1 d.f. χ^2 test	2 d.f. χ^2 test	($\alpha_F = 0{\cdot}033$) ($\alpha_P = 0{\cdot}10$)
5	0·6089	0·5038	0·5746
10	0·8855	0·8155	0·8633
15	0·9722	0·9440	0·9623

Thus, the power of test 3 is at least $(1 - \alpha_P)\beta_1$. The calculated power for all three procedures is shown in Table II for $\lambda = 5, 10, 15$. The tests for methods 1 and 2 were at the 0·05 level. In keeping with the '50 per cent rule', method 3 was performed with $\alpha_P = 0{\cdot}10$ and $\alpha_F = 0{\cdot}033$. The power achieved by always fitting a quadratic model is surprisingly similar to the optimum achievable, and the power achieved by using preliminary testing would be indistinguishable from the optimum in practical contexts.

In order to verify these results with σ^2 unknown, 1000 simulation experiments were conducted with X again taking on the integer values between 1 and 20, and using the model described above. The results were consistent with the results in Table II.[4]

LINEARIZING TRANSFORMATIONS

Introduction

An alternative approach to non-linearity is to transform the predictor variable. Following visual inspection of the scatterplot with or without formal statistical tests, the investigator identifies a

transformation that improves linearity. The test for association between X and Y is done by regressing Y on the transformed X and using the standard 1 d.f. F-test for testing a linear slope. Because the data have been used to select the transformation, this procedure will have an inflated type I error. Thus, we also considered a second approach, that of adjusting the linear slope F-test by using two instead of one degree of freedom for the numerator and decreasing the denominator degrees of freedom by 1. This adjustment was motivated by the quadratic regression procedure of the previous section and by a recommendation of Box and Tidwell[5] where, in using a power transformation of $X(X' = X^{\theta})$ and picking θ to minimize the residual sum of squares, they recommended adjusting the F-test as above.

Simulation study: methodology

Simulation studies were performed in which the transformation used was selected from a set of eight commonly-used transformations: e^X, X^3, X^2, X, $X^{1/2}$, $X^{1/3}$, $\ln X$ and $1/X$. This set is similar to, but not identical with, the 'ladder of re-expression', suggested for straightening curves by Mosteller and Tukey.[6] In each case, the transformation that made the relationship between X and Y as linear as possible was used. Let X' denote a generic transformed X. For each transformation, we fitted the model $Y = \alpha + \beta X' + \gamma(X')^2 + e$ and performed the usual 1 d.f. F-test for the quadratic effect, (that is, $\gamma = 0$). The transformation that minimized the F-test statistic was chosen.

We evaluated five different methods for assessing the relationship between X and Y. They differed in whether or not they performed a preliminary test for non-linearity and whether or not they adjusted the F statistic. We describe the methods as follows.

Let F_i be the final F-test statistic for association for method i ($i = 1, 2, \ldots, 5$) and let n be the number of data points.

Method 1: Regress Y against the most linear transformation of X. F_1 is the standard one degree of freedom F-test statistic for a linear slope.

Method 2: Proceed as in method 1 and then adjust the F-test statistic.
$F_2 = (n - 3)F_1/2(n - 2)$.
Refer to the F distribution with 2 and $(n - 3)$ degrees of freedom.

Method 3: Use a preliminary test for non-linearity. Fit a quadratic model in X and assess the significance of the quadratic term by the standard F-test. If significant at level α_P, regress Y against the most linear transformation of X and compute the standard F-test statistic for the linear slope. If the preliminary test is not significant, regress Y against X and use the standard F-test for the linear slope.

Method 4: Proceed as in method 3 and then adjust the F statistic, as in method 2.

Method 5: Proceed as in method 3, but adjust the F statistic only if the preliminary test achieved statistical significance at α_P.

It is likely that the properties of these transformation-selection procedures depend simultaneously on the values of X and the candidate transformations. The variety of the non-linear shapes described by the transformations over the range of X values may be important. Therefore, we performed six simulation studies, varying the X values and the subsets of candidate transformations: they comprised a main experiment and five variants. In the main experiment, $X = \{1, 2, \ldots, 20\}$ and all eight transformations were considered. Variant 1 was designed to assess the impact of skewness, with X consisting of the fifth, tenth, etc. percentiles from a chi-square distribution with 1 d.f., linearly transformed to range from 1 to 20. Thus, $X = \{1{\cdot}00, 1{\cdot}03, 1{\cdot}09, 1{\cdot}18, 1{\cdot}30, 1{\cdot}46, 1{\cdot}66, 1{\cdot}90, 2{\cdot}18, 2{\cdot}52, 2{\cdot}92, 3{\cdot}40, 3{\cdot}97, 4{\cdot}66, 5{\cdot}50, 6{\cdot}56, 7{\cdot}95, 9{\cdot}90, 12{\cdot}99, 20{\cdot}00\}$.

In variant 2 we maintained equal spacing among the X's but dramatically shrank the value of X_{max}/X_{min}: $X = \{9{\cdot}1, 9{\cdot}2, \ldots, 10{\cdot}9, 11\}$. In variant 3 we increased the value of X_{max}/X_{min} and also included skewness; $X_i = i^2/10, 1, 2, \ldots, 20$. In variants 4 and 5 we used a subset of the eight transformations. In variant 4 we used $\{X, e^X, 1/X\}$ and in variant 5 we used $\{X, X^2, \ln X\}$. For both, $X = \{1, 2, \ldots, 20\}$.

All experiments consisted of 20 uncorrelated (X, Y) pairs, with Y generated as independent standard normal deviates. The values of α_P were $\{0{\cdot}25, 0{\cdot}10, 0{\cdot}05\}$ and of α_F were $\{0{\cdot}10, 0{\cdot}05, 0{\cdot}025, 0{\cdot}01\}$. We ran 1000 simulations of each of the six experiments for each of the twelve (α_P, α_F) combinations. In each group of 1000 simulations, we computed the proportion of simulations in which the final F-test exceeded the critical value corresponding to α_F.

Simulation study: results

Varying α_P had little effect on the observed size in any experiment.[4] Therefore, we present the results for $\alpha_P = 0{\cdot}10$ only. The pattern of results is the same for each experiment (Table III). Method 1 (choosing the most linear transformation without benefit of a preliminary test and without correcting the F-test) is anticonservative and produces a larger type I error than the other methods. However, always correcting the degrees of freedom, with (method 4) or without (method 2) a preliminary test is too conservative. Thus, the methods of choice are methods 3 and 5.

When the number of transformations is small and the nature of the transformations sufficiently modest (variant 5) or when X_{max}/X_{min} is small (variance 2), the selected transformation is only modestly non-linear and method 3 performs quite well. Conversely, when the transformations are more extreme, method 3 is noticeably anticonservative. Method 5 is never anticonservative and, in general, has type I error levels close to the nominal α_F. The good behaviour of these two methods is due to the fact that the preliminary test provides an appropriate mixture of an anticonservative (method 1) with a conservative procedure (method 2).

The proportion of simulations in which each transformation was chosen in the five simulation experiments under method 1 is indicated in Table IV.

Alternative approach: the best-fitting non-linear transformation

We also evaluated an alternative approach in which the best-fitting of the eight transformations as measured by the residual sum of squares was chosen, rather than the most linear. The design of the experiment was the same as the main experiment described above: the predictor variable consisted of the integers from 1 to 20, and the dependent variables were standard normal deviates. However, there was a difference in the preliminary test for non-linearity. Instead of testing the significance of the squared term in quadratic regression, we tested whether any of the transformations fit better than the linear one. We used a very simple *ad hoc* procedure based on the residual sums of squares, although the more complicated theory of testing non-tested hypotheses[7] would also have served. Denote the minimum residual sum of squares over the eight transformations by RSS(min) and denote the residual sum of squares for the untransformed data by RSS(iden). Let $F' = (\text{RSS(iden)} - \text{RSS(min)})/\text{RSS(min)}$. We used the transformation minimizing RSS if and only if $(n-3)F'$ exceeded the α_P critical level of the F distribution on 1 and $n-3$ degrees of freedom $[F_{\alpha_P}(1, n-3)]$. Thus, the transformation was used if and only if the usual linear model increased the residual sum of squares by $[F_{\alpha_P}(1, n-3)/(n-3)] \times 100$ per cent, the same per cent by which we judged the best quadratic fit relative to a linear model. The final test was based on the standard F-test of the linear slope. One thousand simulations were run for each

Table III. Size of type I error when choosing the most linear transformation ($\alpha_P = 0{\cdot}10$)

	α_F	F_1	F_2	F_3	F_4	F_5
Main experiment	0·10	0·146	0·054	0·129	0·043	0·111
$X = \{1, 2, \ldots, 20\}$	0·05	0·081	0·023	0·068	0·017	0·061
	0·025	0·047	0·014	0·036	0·011	0·028
	0·010	0·021	0·004	0·016	0·002	0·010
Variant 1						
Skewed X	0·10	0·154	0·058	0·128	0·047	0·114
All 8 transformations	0·05	0·084	0·017	0·059	0·011	0·046
	0·025	0·052	0·018	0·041	0·014	0·033
	0·010	0·021	0·006	0·016	0·005	0·012
Variant 2						
Reduced X_{max}/X_{min}	0·10	0·106	0·043	0·088	0·039	0·084
	0·05	0·062	0·012	0·054	0·011	0·049
All 8 transformations	0·025	0·025	0·004	0·025	0·005	0·020
	0·010	0·010	0·003	0·009	0·003	0·007
Variant 3						
Skewed X	0·10	0·132	0·051	0·119	0·042	0·105
Expanded X_{max}/X_{min}	0·05	0·073	0·026	0·054	0·021	0·045
All 8 transformations	0·025	0·047	0·010	0·041	0·009	0·029
	0·010	0·022	0·005	0·014	0·003	0·011
Variant 4						
$X = \{1, 2, \ldots, 20\}$	0·10	0·125	0·036	0·117	0·034	0·096
3 transformations	0·05	0·053	0·011	0·055	0·013	0·049
X, e^X, $1/X$	0·025	0·030	0·011	0·027	0·011	0·023
	0·010	0·016	0·006	0·016	0·006	0·013
Variant 5						
$X = \{1, 2, \ldots, 20\}$	0·10	0·129	0·043	0·107	0·034	0·097
3 transformations	0·05	0·070	0·025	0·058	0·020	0·049
X, X^2, $\ln(X)$	0·025	0·042	0·011	0·033	0·009	0·025
	0·010	0·009	0·001	0·008	0·000	0·007

of the values of α_P and α_F listed previously. As with the simulations for the most linear transformation, the type I error was approximately constant for all values of α_P considered.

For $\alpha_F = 0{\cdot}05$, the observed type I error rates were 0·141, 0·041, 0·106, 0·033, and 0·067 for methods 1, 2, 3, 4, and 5, respectively. The type I errors for $\alpha_F = 0{\cdot}025$ and 0·01, showed a similar pattern.[4] The nominal level is closest to the true type I error for method 2 which simply takes the transformation that fits best (in terms of smallest residual sum of squares) and corrects the degrees of freedom. This finding differs from the previous approach (where the most linear transformation was used) for which method 5 provided most accurate control over the size of the test and agrees with the recommendations of Box and Tidwell.[5]

Method 1 for the two approaches is clearly anticonservative, but noticeably more so using the best-fitting rather than the most linear transformation.

The percentage of experiments for which each transformation gave the best fitting model were:

e^X: 31·3 per cent; X^3: 15·0 per cent; X^2: 6·9 per cent; X: 6·6 per cent;

$X^{1/2}$: 3·3 per cent; $X^{1/3}$: 3·2 per cent; $\ln X$: 7·5 per cent; $1/X$: 26·3 per cent.

Table IV. Proportion of 12,000 simulations in which each transformation was chosen as most linear

	Transformation							
	e^X	X^3	X^2	X	$X^{1/2}$	$X^{1/3}$	$\ln(X)$	$1/X$
Main experiment $X = \{1, 2, \ldots, 20\}$	16·3%	14·3%	12·2%	12·1%	7·9%	6·7%	13·0%	17·5%
Variant 1 Skewed X	15·9%	12·0%	13·1%	14·1%	8·3%	7·0%	12·1%	17·5%
Variant 2 Restricted range	46·7%	8·9%	2·1%	1·5%	0·7%	0·6%	1·4%	38·1%
Variant 3 Skewed X Expanded range	14·1%	12·6%	11·2%	12·6%	10·1%	9·9%	14·4%	15·1%
Variant 4 $X = \{1, 2, \ldots, 20\}$	33·2%	—	—	33·1%	—	—	—	33·7%
Variant 5 $X\{1, 2, \ldots, 20\}$	—	—	35·7%	23·5%	—	—	40·9%	—

Comparison with the main experiment in Table IV shows that the 'best-fitting' transformation is more likely to be e^X or $1/X$ than is the 'most linear' transformation.

Problems with using the best-fitting transformation

Although one can control the size of the test for association by method 2, nevertheless the approach of picking the best-fitting transformation has two serious drawbacks, one theoretical and one practical. Theoretically, the set of transformations we are considering extends naturally to the power family in which $E(Y) = \alpha + \beta X^{\theta}$. However, under the null hypothesis of no association between X and Y, the parameters are not all estimable.[8] Practically, the best fitting transformation is frequently one of the extreme transformations (reciprocal or exponential) and tends to fit the extremes of the data (smallest or largest X values), thus spuriously inflating goodness-of-fit statistics. Figure 1 shows a typical example from a simulation where the exponential transformation seems to fit markedly better than no transformation because F' is 'significant' at the 0·05 level. However, there is no visible evidence of non-linearity.

EXAMPLE

We illustrate the practical implication of the preceding discussion with an example consisting of a study performed at the Mayo Clinic in which the relationship between sensation and age was of interest. Specifically, the threshold at which a vibratory stimulus could be detected was measured on 51 healthy males aged 50 and over using the methodology described in Dyck *et al.*[9] Preliminary analysis of the data showed strong evidence of non-normality which as ameliorated by squaring the observed threshold. Further analysis used the squared threshold as the dependent variable. The data are shown in Figure 2. A low threshold corresponds to good sensation. Although the linear association is not statistically significant ($P = 0{\cdot}137$), there is visual evidence of a non-linear association. The improvement obtained with the quadratic model is significant at the $P = 0{\cdot}054$ level, and the corresponding overall 2 degrees of freedom test for association using the quadratic model is significant at the $P = 0{\cdot}052$ level.

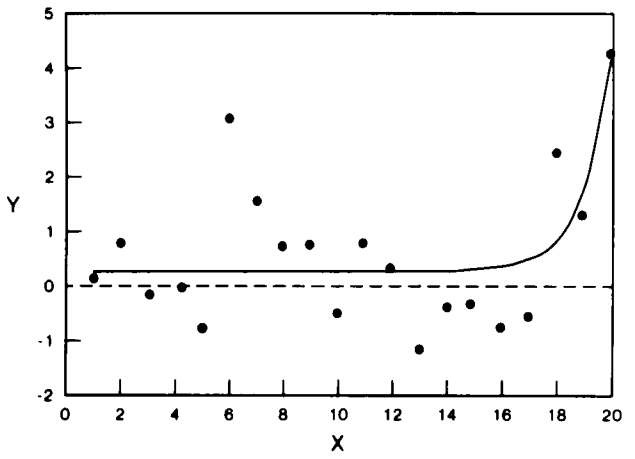

Figure 1. A simulated data set in which the Y variables are independent standard normal deviates. The line on the plot is the least squares regression line of Y on $\exp(X)$. The plot gives an example of how end effects can spuriously inflate the goodness-of-fit of extreme transformations

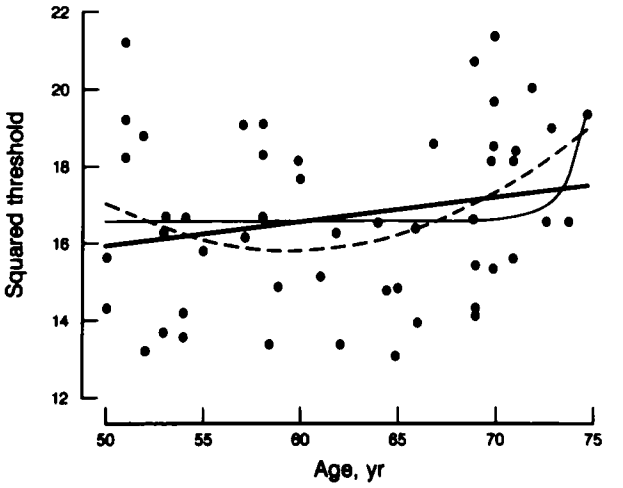

Figure 2. The relationship between the squared vibratory stimulus threshold and age in a simple of 51 men. The thick solid line is the linear regression line; the dashed line shows quadratic regression and the thin solid line is the regression on exp(age)

The results of such a standard analysis raises questions about the interpretation of P values reported in conjunction with the quadratic model, since a departure from simple linear regression was motivated from visual inspection of the data. Of course, one cannot know for certain what changes in sensation occur with ageing. However, the discussion in Sections 2 and 3 leads us to conclude that a statistically significant association has not been demonstrated, using either the linear or quadratic models, since the adjusted P value for the latter analysis (using the 50 per cent rule) is 0·078. Although the possibility that sensation diminishes with age at a possibly

accelerating rate is suggested, the possibility that sensation actually improved during the initial 10 year period (as suggested by the data) seems implausible to us.

Finally, the most linear transformation obtained from the family of 8 transformations studied in Section 3 was $X' = e^X$. The P value obtained using simple linear regression with this transformation was 0·182. Alternatively, using a 2 d.f. F-test as described in method 5 yields $P = 0{\cdot}414$.

DISCUSSION

The discussion of transformations in regression frequently encompasses more than just transformations of the predictors and includes transformations of the dependent variable. A prime example of transformation of the dependent variable only is the Box–Cox transformation and a recent example of simultaneous transformation of independent and dependent variables is the ACE algorithm.[10]

Transformation of the dependent variable alone can have a substantially smaller impact on test statistics than does transformation of the predictor as reported here. The Box–Cox transformation of Y is a power transformation, $h(Y) = (Y^\lambda - 1)/\lambda$ where λ is estimated from the data so as to maximize the normal theory linear model likelihood for the transformed Y. Doksum and Wong[11] have shown by asymptotic argument that the level and power of standard tests in simple linear regression using the transformed data and ignoring the transformation selection procedures are the same as when the appropriate transformation is assumed known. Berry[12] considers a different transformation ($h(Y) = \log(Y + \lambda)$) and optimizes the symmetry and kurtosis of the residual rather than maximizing likelihood. He shows by simulation in a 2-way ANOVA model that the size and power of F-tests are minimally affected by the fact that λ is estimated from the data.

However, as we have shown, when the predictor variable is transformed, the null distribution of the test statistic is affected by the fact that the transformation is estimated from the data. When the transformation is chosen to optimize linearity, the true type I error can be increased by more than 50 per cent relative to the nominal error level (see the results for method 1 in Table III). When the transformation is chosen to minimize the residual sum of squares, the impact on the test size is even greater.

Another current use of transformation in regression is the ACE algorithm.[10] ACE attempts to maximize the proportion of variance explained by linear regression (R^2) by using a bivariate smoother to develop transformations for the dependent variable and each of the independent variables. This procedure can find structure in data that standard techniques would miss but, as pointed out by Fowlkes and Kettenring,[13] it will also find structure where none exists. They showed that, for independent bivariate normal data, ACE found correlations averaging about 0·2 for the transformed variates.

We would argue that this problem is largely because ACE has no way of controlling for the transformation selection procedure and no preliminary assessment of the need for transformation. Fowlkes and Kettenring[13] suggest that 'criteria that focus directly on achieving linearity of regression, rather than maximizing correlation, would be worth exploring'. Our results suggest that this modification will not suffice. One must adjust for the fact that the best transformation, whether 'most linear' or 'best fitting', is estimated from the data; otherwise, transformations can suggest spurious relationships where none exists.

An alternative strategy for dealing with non-linearity, when the association is known to be monotone, is to test for association using rank correlation. Although we have not evaluated this

method, we note that rank correlation may be viewed simply as the test of another transformation for achieving linearity.

The approach based on a preliminary test for non-linearity and fitting a quadratic model as needed, has potential application beyond the usual regression context. O'Brien[3] uses non-linear regression to extend various standard tests for two-sample comparisons (two-sample t, rank sum, and logrank test) to deal with heterogeneity of response between the two samples. In medical experimentation, a specific treatment may not be equally efficacious for all patients. This fact may mean that both the mean and the variance are increased in the treatment group compared with the control group. In that case, the regression of the treatment indicator variable on the response variable will have a non-linear component. On the other hand, when a shift in location model holds and there is no heterogeneity, the regression will be linear. Thus, O'Brien has suggested the following procedure: first, test for non-linearity by testing the significance of the quadratic term; then use the outcome of that test to determine whether to use a linear model with a 1 d.f. test or a quadratic model with a 2 d.f. test. O'Brien has shown by simulation and several real examples that this procedure can detect treatment effects that standard procedures miss. However, his procedure did not explicitly take into account the potential distortion in test size due to the preliminary test for non-linearity.

To investigate the properties of O'Brien's procedure, we performed a small scale simulation study of the null case with 25 subjects in each group and the data in each group obtained from a standard normal distribution. Each experiment was replicated 1000 times. For $\alpha_P = 0{\cdot}10$, we estimated a true type I error rate of 0·01 for $\alpha_F = 0{\cdot}01$, 0·031 for $\alpha_F = 0{\cdot}025$, 0·066 for $\alpha_F = 0{\cdot}05$ and 0·150 for $\alpha_F = 0{\cdot}10$. Results for other values of α_P^4 were consistent. In this application, the '50 per cent rule' is somewhat conservative; the nominal and true type I error rates are closer together.

We conclude with some practical suggestions for the analysis of regression data where the primary concern is the possibility of non-linearity and where the investigator plans to use modifications of the independent variable alone to achieve linearity, assuming that other problems such as heteroscadescity, non-normality and outliers are not of concern or have been dealt with. Using the quadratic regression approach, we recommend modifying usual practice as follows. If the preliminary test fails to reject linearity at conventional significance levels ($0{\cdot}05 \leqslant \alpha_P \leqslant 0{\cdot}25$), regress Y on X and compute the standard F-test for regression, referring it to the central F distribution with 1 and $n - 2$ degrees of freedom. If the test is significant, regress Y on X and X^2 and compute the usual 2 degree of freedom F statistic to assess the overall significance of the regression and refer to the central F distribution on 2 and $n - 3$ degrees of freedom. The resulting P-value must be increased by 50 per cent irrespective of the outcome of the test. Thus, in designing an experiment in which potential non-linearity is of concern, one should set the α-level at 2/3 of that desired. This procedure will have minimal loss of power in case the relationship is truly linear.

We are more cautious about making recommendations in using linearizing transformations. Our simulations cover a limited, albeit varied, set of circumstances, and the properties of any transformation-selection procedure depend on the family of transformations considered. Nevertheless, the results of our various simulation experiments were concordant. It is clear that simply picking the most linear transformation from a candidate set and doing standard linear regression can result in a substantially inflated type I error. Method 5 (using a preliminary test based on quadratic regression and choosing a transformation only if the quadratic term is significant) is easily implemented and appears to provide improved control over the size of the test, although further work is indicated.

To summarize, we have evaluated two commonly used methods for dealing with the problem of non-linearity in regression. Our results indicate that simple modification of standard practice, resulting in only slight loss of power, is needed to control the size of the test for association.

APPENDIX

This appendix provides the proof for equation (2). From (1), we have

$$Y_i = a + bT_i + cZ_i + e_i$$

where

$$T_i = (X_i - \bar{X})/\Sigma(X_i - \bar{X})^2$$

$$Z_i = Z_i'/\Sigma Z_i'^2$$

and

$$Z_i' = T_i^2 - (\Sigma T_i^3/\Sigma T_i^2)T_i - \Sigma T_i^2/n.$$

Without loss of generality, we can assume $\sigma^2 = 1$.

Let $\hat{b}$ and $\hat{c}$ denote the least squares estimates of b and c. With σ^2 known, chi-squared tests replace F-tests. Let $q_P = \chi_1^2(1 - \alpha_P)$ where $\chi_V^2(\mu) =$ the μ percentile of the chi-squared distribution with V d.f. so q_P is the critical value for the preliminary test. Similarly, let $q_1 = \chi_1^2(1 - \alpha_F)$ and $q_2 = \chi_2^2(1 - \alpha_F)$ so that q_1 and q_2 are the critical values for the final 1 d.f. and 2 d.f. tests, respectively. If $\hat{c}^2 \leqslant q_P$, one assumes a linear model. One rejects the null hypothesis, H_0, of no relationship between X and Y if $\hat{b}^2 > q_1$. If $\hat{c}^2 > q_P$, one assumes the quadratic model and rejects H_0 if $\hat{b}^2 + \hat{c}^2 \geqslant q_2$.

Under H_0, $\hat{b}$ and $\hat{c}$ are independent standard normal random variables and therefore the true probability of type 1 error is easily derived.

$$\begin{aligned} Pr\{\text{type I error}\} &= Pr\{\text{rejecting } H_0 \text{ and } \hat{c}^2 \leqslant q_P\} + Pr\{\text{rejecting } H_0 \text{ and } \hat{c}^2 > q_P\} \\ &= (1 - \alpha_P)\alpha_F + \alpha_P P^* \end{aligned}$$

where

$$\begin{aligned} P^* &= Pr\{\hat{b}^2 + \hat{c}^2 > q_2 \mid \hat{c}^2 > q_P\} \\ &= \alpha_P^{-1} \int_{q_P}^{\infty} \left[\int_{q_2 - y}^{\infty} g(u)\mathrm{d}u \right] g(y)\mathrm{d}y \end{aligned}$$

when $g(y)$ is the density for a χ_1^2 random variable.

ACKNOWLEDGEMENT

We are indebted to the referees for helpful suggestions.

REFERENCES

1. Snedecor, G. W. and Cochran, W. G. *Statistical Methods*, 7th edn, Iowa State University Press, Ames, Iowa, 1980.
2. Weisberg, S. *Applied Linear Regression*, 2nd edn, Wiley, New York, 1985.
3. O'Brien, P. C. 'Comparing two samples: Extensions of the *t*, rank sum, and log rank tests', *Journal of the American Statistical Association*, **83**, 52–61 (1988).

4. Grambsch, P. M. and O'Brien, P. C. 'The effects of preliminary tests for nonlinearity in regression', *Mayo Clinic Biostatistics Technical Report*, No. 39, (1988).
5. Box, G. E. P. and Tidwell, P. W. 'Transformation of the independent variables', *Technometrics*, **4**, 531–550 (1962).
6. Mosteller, F. and Tukey, J. W. *Data Analysis and Regression*, Addison-Wesley, Reading, MA, 1977.
7. McAleer, M. and Pesaran, M. H. 'Statistical inference in non-nested econometric models', *Applied Mathematics and Computation*, **20**, 271–311 (1986).
8. Gallant, A. R. 'Nonlinear regression', *American Statistician*, **29**, 73–81 (1975).
9. Dyck, P. J., Karnes, J., O'Brien, P. C. and Zimmerman, I. 'Detection thresholds of cutaneous sensation in man in health and disease', *in* Saunders, W. B., *Peripheral Neuropathy*, 2nd edn, 1984, Chapter 49, pp. 1103–1139.
10. Breiman, L. and Friedman, J. 'Estimating optimal transformations for multiple regression and correlation', *Journal of the American Statistical Association*, **80**, 580–619 (1985).
11. Doksum, K. A. and Wong, C. W. 'Statistical tests based on transformed data', *Journal of the American Statistical Association*, **78**, 411–417 (1983).
12. Berry, D. 'Logarithmic transformations in ANOVA', *Biometrics*, **43**, 439–456 (1987).
13. Fowlkes, E. B. and Kettenring, J. R. 'Comment: the ACE method of optimal transformations', *Journal of the American Statistical Association*, **80**, 607–613 (1985).