Appl. Statist. (1994)
43, *No.* 2, *pp.* 347–355

Tests of Linearity, Multivariate Normality and the Adequacy of Linear Scores

By D. R. COX†

Nuffield College, Oxford, UK

and NANNY WERMUTH

University of Mainz, Germany

[Received October 1992. Final revision April 1993]

SUMMARY

After some discussion of the purposes of testing multivariate normality, the paper concentrates on two different approaches to testing linearity: on repeated regression tests of non-linearity and on exploiting properties of a dichotomized normal distribution. Regression tests of linearity are used to examine the adequacy of linear scoring systems for explanatory variables, initially recorded on an ordinal scale. Examples from recent psychological and medical research are given in which the methods have led to some insight into subject-matter.

Keywords: Constancy of variance; Linearity of association; Multivariate normality; Normal plots; Ordinal scales

1. Introduction

Much of the 'classical' approach to the multivariate analysis of continuous variables rests to some extent on multivariate normality. Most of the distribution theory and optimality of standard test procedures derive directly from this assumption. More importantly, the methods of data reduction hinge on the calculation of sample mean vectors and covariance matrices, or sometimes of 'robust' versions of these quantities. The exponential family structure of the multivariate normal distribution provides a strong theoretical justification for such data reduction.

There are several reasons for checking multivariate normality. Occasionally the central limit theorem may be thought to have operated when the data were generated. Then this itself may be the hypothesis of interest about data generation. Next we might want to learn about the effect of departures from standard assumptions on the properties of formal tests of significance and interval estimates. Thirdly, the substantive objective might involve use of the multivariate normal form to calculate some derived probabilities, for example concerned with particular extreme regions for future observations. Finally we may be concerned that the reduction of the observations to covariance matrices overlooks important features of the dependences of intrinsic interest.

†*Address for correspondence*: Nuffield College, Oxford, OX1 1NF, UK.

 0035–9254/94/43347

The present paper concentrates on the last objective which we believe to be often the most important in applications. Thus, the procedures discussed here are not direct generalizations of those used for testing univariate normality.

In Section 2 we first review and extend suggestions of Cox and Small (1978) to test normality via repeated standard regression tests of non-linearity; another approach uses properties of a dichotomized normal distribution (McFadden, 1955) to test normality. In Section 3 we outline somewhat similar ideas applied to the analysis of ordinal data.

2. Testing Multivariate Normality

We suppose that observations y_{sl} ($s = 1, \ldots, p$; $l = 1, \ldots, n$) are available for p continuous (response) variables for n independent individuals and concentrate here on two types of systematic departure from multivariate normality: the presence of

(a) curvature in the relationship between a pair of variables and
(b) an 'interactive' effect in which the slope of the linear relationship between two variables depends on the value of a third variable.

Non-linearity in the dependence of y_s on y_t is detected by inserting a squared term y_t^2 in the regression of y_s on y_t. An interaction in the dependence of y_s on y_t and y_u is detected by inserting a cross-product term $y_t y_u$ in the regression of y_s on y_t and y_u. To achieve numerical stability it is helpful to subtract a constant such as its mean from y_t and y_u before squaring and computing cross-products. These are standard procedures (Ezekiel, 1926; Cox and Small, 1978). For testing significance in a single application they lead to Student t-statistics, to be denoted $Q_{s,t}$ and $Q_{s,tu}$ respectively.

In applications some choices are to be made and the following recommendations are based on largely qualitative arguments which may be amended in the light of more extensive practical experience.

(a) All possible $p(p-1)$ Student t-statistics for squared terms should be calculated unless there are prior reasons for concentrating on some particular terms because they are of special intrinsic interest or because experience of previous similar data has shown them to be likely to be non-linear.
(b) In fitting say the term in y^2 in the regression of y_s on y_t and y_t^2, it has to be decided what other terms to include in the fit. Sensible initial strategies seem to be either not to include any other explanatory variable, so that non-linearities in marginal relations are detected, or, if p is small, to include all other important (explanatory) variables, or, if p is large, to include all other (explanatory) variables contributing appreciably to the linear regression.
(c) Similar remarks apply to the $\frac{1}{2}p(p-1)(p-2)$ possible choices of cross-product terms $y_t y_u$ in the linear regression of y_s.

One important general issue is that if the joint distribution of (X, Y) is of interest, then, even if Y is a response variable, it is fruitful to consider both the regression of Y on X, X^2 and of X on Y, Y^2. The reason is that it may happen that the quadratic term in the latter but not in the former regression becomes substantial. A possible interpretation is then (Cox and Wermuth, 1993) that the conditional

variance of the response Y changes systematically with X.

In all, $\frac{1}{2}p^2(p-1)$ Student t-statistics are available and so in any rough assessment of significance allowance for selection is necessary and restrictions on prior grounds on the number of statistics will aid sensitivity. Arranging the t-values in tables permits convenient examination only for few variables. For larger sets of t-values we suggest plotting them against expected values of normal order statistics (Cox and Hinkley, 1974), i.e. we plot the rth largest statistic against

$$\Phi^{-1}\left(\frac{r-\frac{3}{8}}{n+\frac{1}{4}}\right). \tag{1}$$

Although considerable caution is needed in interpretation, we find such probability plots helpful when there are enough points, say 20 or more. Note further that

(a) the degrees of freedom of the t-statistics should be sufficiently large for the standard normal distribution to be a reasonable approximation to the Student t-distribution,

(b) modest correlation between different t-statistics may produce systematic displacement from the unit line, but this effect should become less important as the number of essentially independent points plotted increases, and

(c) outliers in one variable may show in the plot by a number of extreme points associated with that variable and an inspection of scatterplots is necessary to distinguish such outliers from systematic non-linearity. It would be helpful to know more about the effects of correlation on such plots.

Especially when there is a large number of observations an alternative approach to testing consistency with multivariate normality can be based on dichotomizing all variables at their medians. In each of the resulting $2 \times 2 \times 2$ tables frequencies are equal in pairs and given by simple functions of the marginal correlations (McFadden, 1955). For any three of the resulting dichotomized variables A, B and C the corresponding probabilities are

$$\begin{aligned} p_{000}^{ABC} &= \tfrac{1}{8} + \eta_{AB} + \eta_{AC} + \eta_{BC}, \\ p_{100}^{ABC} &= \tfrac{1}{8} - \eta_{\mathrm{AB}} - \eta_{\mathrm{AC}} + \eta_{\mathrm{BC}}, \end{aligned} \tag{2}$$

etc., where for instance

$$\eta_{AB} = (4\pi)^{-1}\sin^{-1}\rho_{12}$$

and ρ_{12} is the correlation between the two normal variables generating A and B. In general the contribution of say η_{AB} in equations (2) is positive if A and B take the same values and negative if they take different values. To see this, note that changing the level of A is equivalent to changing the sign of the underlying normal variable and hence to changing signs of the corresponding correlation coefficients. It follows that the probabilities are equal in pairs:

$$p_{000}^{ABC} = p_{111}^{ABC}, \qquad p_{100}^{ABC} = p_{011}^{ABC}, \qquad p_{010}^{ABC} = p_{101}^{ABC}, \qquad p_{110}^{ABC} = p_{001}^{ABC}.$$

Also the three-factor interaction vanishes, whether calculated directly from the probabilities or from the log-probabilities.

An arbitrary trivariate normal distribution of zero mean and unit variance has

three adjustable parameters whereas an arbitrary three-variable binary distribution with marginal probabilities $\frac{1}{2}$ has four, showing that there is 1 degree of freedom for testing consistency with the special symmetries implied in particular by trivariate normality. Thus, whenever there are enough observations to obtain reliable estimates of medians, a large linear or log-linear three-factor interaction in one of the three-way contingency tables formed by the median dichotomized variables is evidence for non-linearity.

The following examples illustrate the types of interpretation suggested by such analyses. In the first example there are six variables and a small sample size, whereas in the second example there are three variables and a reasonably large sample size.

2.1. *Example* 1 (*Blood Glucose Control*)

From on-going investigations of determinants of blood glucose control (Kohlmann *et al.*, 1993) we have previously analysed observations for 70 diabetic patients, all having had fewer than 10 years of formal schooling (Cox and Wermuth, 1993). We show here that the same two variables which have an interactive effect on metabolic adjustment of all 70 patients show up when checking for non-linearities between the variables in a collective of only 32 patients, all from the second stage of data collection 1 year after the first.

The variables are Y, a particular metabolic parameter (the glycosylated haemoglobin), X, a standardized score for particular knowledge about diabetes, and W, duration of illness in months. Furthermore, three different attitudes of the patients are measured as subscale sum scores of a questionnaire. The attitudes are intended to capture to whom or to what the patient attributes what is happening about his illness: Z, social externality (powerful others are responsible); U, fatalistic externality (mere chance determines what occurs); V, internality (the patient sees himself as mainly responsible). Summary statistics are given in Table 1. There are only two large outlying marginal t-statistics among the 60 product terms; for regressions of Z on $Y, W, Y \times W$ ($Q_{z,yw} = 3.61$) and of Y on $Z, W, Z \times W$

TABLE 1

Observed marginal correlations (lower half), observed partial correlations given all remaining variables (upper half), means and standard deviations for 32 diabetic patients with fewer than 10 years of formal schooling

Variable	Y	X	Z	U	V	W	$Z \times W$
Y: glycosylated haemoglobin	1	0.20	0.26	0.24	0.35	0.02	0.48
X: knowledge	−0.09	1	−0.49	−0.56	−0.45	0.08	−0.11
Z: social externality	0.27	−0.51	1	−0.08	−0.28	0.04	0.03
U: fatalistic externality	0.08	−0.51	0.33	1	−0.56	0.01	−0.22
V: internality	0.15	−0.08	−0.18	−0.41	1	−0.08	−0.24
W: duration	0.02	0.08	0.02	0.02	−0.12	1	0.01
$Z \times W$: $(z_l - \bar{z})(w_l - \bar{w})$	0.47	−0.05	0.20	−0.03	−0.08	0.04	1
Mean	8.63	0.00	28.66	21.66	39.09	108.6	11.99
Standard deviation	2.35	1.00	6.93	7.32	5.38	81.93	512.1

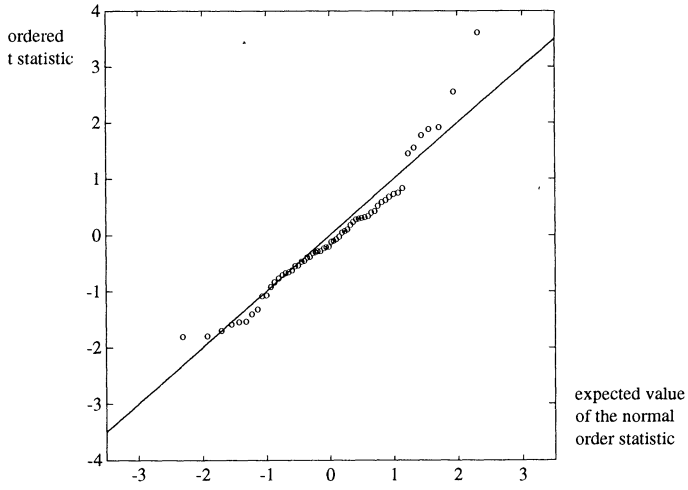

Fig. 1. Normal plot of ordered t-statistics $Q_{s,tu}$.

($Q_{y,zw} = 2.56$). They show up as larger deviations from the diagonal line in the normal plot, in the plot of the marginal $Q_{s,tu}$ terms against function (1) in Fig. 1.

A possible explanation for the non-linear relationships between Y, Z and W is that the dependence of metabolic adjustment Y on social externality Z changes with the duration of the illness, i.e. that

$$E(Y|Z = z,\ W = w) = \alpha + (\beta + \gamma w)z + \delta w,$$

whereas the small correlations among duration of illness and all other variables (Table 1) imply that there are no linear relationships to duration of illness.

2.2. *Example* 2 (*Effects of Parents' Child Rearing Styles*)

From research on effects of child rearing styles on the manifestation of anxiety in the child (Kohlmann *et al.*, 1987), we analyse here observations for 246 children aged 10–14 years. One research hypothesis is that there is an effect of inconsistent behaviour of the mother, X, on the manifestation of anxiety in the child, Y, moderated by supportive behaviour of the father, V. The first part of this statement is reflected in a substantial marginal correlation between Y and X, $r_{yx} = 0.53$; see Table 2. The second part appears to be refuted, at first sight, since the marginal correlations $r_{yv} = -0.12$ and $r_{xv} = -0.06$ point towards linear independence of V and Y, X. However, in the regression of Y on $X, V, X \times V$ the t-statistic for the product term is large, $Q_{Y,XV} = -3.15$. The data summaries in Table 3 show that the effect of V is in the expected direction.

By median dichotomizing on the variable V we compare children receiving little support from their fathers ($n = 126$) with children receiving much support ($n = 120$). The effect of inconsistent behaviour of the mother is considerably stronger in the former group: $r_{yx} = 0.64$ *versus* $r_{yx} = 0.40$ in the latter. An appropriate test statistic corresponding to this observed difference depends on the distributional assumptions for the three variables (Lauritzen and Wermuth, 1989;

TABLE 2
Observed marginal correlations (lower half) and observed partial correlations given all remaining variables (upper half), means and standard deviations for 246 pupils

Variable	*Y*	*X*	*V*	*X* × *V*
Y: anxiety, child	1	0.538	−0.108	−0.198
X: inconsistency, mother	0.527	1	0.012	0.145
V: support, father	−0.117	−0.055	1	−0.034
X × *V*: $(x_l - \bar{x})(v_l - \bar{v})$	−0.141	0.047	−0.015	1
Mean	30.57	23.61	34.67	−3.48
Standard deviation	7.01	6.62	9.53	67.59

TABLE 3
Non-linear dependence of anxiety in the child Y on inconsistent behaviour of the mother X and supportive behaviour of the father V as reflected in the observed correlations between Y and X for two levels of V obtained by median dichotomizing

Results for supportive behaviour of the father			
Low		*High*	
$\bar{y} = 31.21$	$\bar{x} = 23.89$	$\bar{y} = 29.89$	$\bar{x} = 23.32$
$s_y = 7.48$	$s_x = 6.58$	$s_y = 6.45$	$s_x = 6.69$
$r_{yx} = 0.64$	$n = 126$	$r_{yx} = 0.40$	$n = 120$

Cox and Wermuth, 1992). For the purpose here it is enough to note the large t-statistic for the differences in the correlations: 2.65.

The non-linear relationship between X, Y and V is also detected when all three variables are dichotomized at their medians and the resulting counts n_{ijk} in the $2 \times 2 \times 2$ contingency table are inspected; i corresponds to Y, j to X and k to V. If we denote a level with values below the median by 0 and the other by 1, we can list the counts as

$$(n_{000}, n_{100}, n_{010}, n_{110}, n_{001}, n_{101}, n_{011}, n_{111}) = (45, 13, 20, 48, 40, 25, 23, 32).$$

The odds ratio computed for the first four values is, at 8.31, considerably larger than the odds ratio of the last four values, 2.23. Accordingly, the Studentized value of the three-factor interaction term is large, having a value of 2.37 on a log-linear scale and a value of 2.30 on a linear scale. This interaction is in line with the interpretation of the non-linear relationship derived via Table 3.

3. Adequacy of Linear Scores for Ordinal Data

We now sketch a special application of formal tests for non-linearity when we have a continuous response variable and a number of ordinal explanatory variables. The purpose is to decide whether scores on a linear scale or on some modified scale may be used for the ordinal variables. Similar arguments would apply to binary

responses or survival times as responses. The proposed procedure is likely to be most useful when there is a fairly large number of ordinal explanatory variables. In that case it may be cumbersome to apply methods in which effect scores for ordinal characteristics are determined from the data, like those related to Hirschfeld's (1935) work, and it would be inefficient to disregard the ordinal structure by treating the variables as qualitative, i.e. as nominally scaled. Significance testing and interpretation by simple scores for two ordinal variables was suggested by Yates (1948); the relationship to common codings of nominally scaled variables was discussed by Wermuth and Cox (1992).

The following is a readily implemented procedure. Score each ordinal variable on a linear scale, often but not necessarily with equally spaced values, i.e. with scales obtained via the first of the associated standard orthogonal polynomials. For example, a three-point scale could be scored as -1, 0 or 1, a four-point scale as -3, -1, 1 or 3 and so on. Then, for each explanatory variable in turn, we examine non-linearity of regression by including its square, or, equivalently, the second orthogonal polynomial, as an explanatory variable. For an original scale with just three levels such a quadratic scheme is equivalent to treating the variable as nominally scaled. If we find evidence for a particular explanatory variable X that non-linearity is indicated there will be a qualitative interpretation. Thus, for example, if both linear and quadratic terms have the same sign, this implies that the relative scoring of the highest levels of X should be increased. The regression coefficients of the linear and the quadratic term may be used for rescoring of the levels if the interpretation implied is consistent with subject-matter knowledge.

In particular we can conclude that

(a) if a small non-significant quadratic contribution is obtained the data are consistent with a linear scoring scheme in the sense that a 'smooth' quadratic departure in the scores offers no better fit to the response variable and
(b) if an appreciable quadratic component is obtained some modification of linear scoring is indicated.

Our approach is to compute the fitted scores at the points on the original scale to suggest a simple modification.

In the following example this leads to collapsing some of the original three-point scales into two-point scales; in other problems simple expansion or contraction of particular intervals between adjacent points on the scale may be indicated, e.g. by doubling or halving an interval.

3.1. *Example* 3 (*Chronification of Pain*)

In a study of patients treated in a pain clinic (Schmitt, 1990) 10 ordinal scales, each with values -1, 0 or 1, were used to describe different aspects of patients with chronic pain and to construct a new variable, Y, a measure of the chronification of pain. We test here whether this measure is reproduced well by a linear regression on the 10 ordinal variables. These are called duration of pain attacks, X_1, frequency of pain attacks, X_2, changes in intensity of pain, X_3, localization of pain, X_4, treatments of drug addiction, X_5, drug usage, X_6, change of physician, X_7, pain-induced rehabilitative treatment, X_8, pain-induced stationary treatment, X_9 and pain-induced surgery, X_{10}. Observed marginal correlations and partial correlations between Y and each of the X_i are shown in Table 4.

TABLE 4
Observed marginal correlations (upper row) and observed partial correlations given the 12 remaining variables (lower row) with the response Y: stage of chronic pain, for 149 patients

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_1^2	X_6^2	X_9^2
0.53	0.49	0.05	0.71	0.17	0.23	0.14	0.44	0.19	0.28	0.20	−0.09	−0.19
—	0.33	0.21	0.91	0.25	—	0.56	0.51	—	0.28	0.25	−0.55	−0.20

The t-statistics of the quadratic terms of just three variables are large, when included after all linear terms, and also when included in addition to all linear and the two other quadratic terms: the latter are -7.74 for X_6, 2.97 for X_1 and -2.34 for X_9. The regression coefficients are (0.577, -0.542) for (X_6, X_6^2), (0.428, 0.253) for (X_1, X_1^2) and (0.117, -0.178) for (X_9, X_9^2). Each of the remaining seven linear components contributes to predicting Y since the t-statistics are larger than two in the regression of Y on the 10 linear and the three quadratic terms.

We chose scores -1, 0 and 1 for each of the ordinal variables so that the fitted scores computed for X_6 in this scale are obtained as

$$0.5767\begin{pmatrix}-1\\0\\1\end{pmatrix} + (-0.5419)\begin{pmatrix}1\\0\\1\end{pmatrix} = \begin{pmatrix}-1.1186\\0\\0.0348\end{pmatrix}.$$

These fitted scores suggest that the interval between levels 2 and 3 is negligible compared with that between levels 1 and 2. The definition of the levels in the original scale is

(a) -1: irregular intake of pain relieving drugs,
(b) 0: regular intake of at most two pain relieving drugs and
(c) 1: regular intake of more than two pain relieving drugs.

An interval near 0 between levels 2 and 3 implies therefore that only the distinction between irregular *versus* regular drug usage is informative for the response Y, or collapsing the levels 2 and 3 into a single level will not worsen prediction of Y for the patient population involved. In a similar way the fitted values for X_1 imply with values (-0.1745, 0, 0.6812) that levels 1 and 2 (pain attacks of several hours and pain attacks for several days) may be collapsed and just contrasted with the third level, pain attacks lasting more than a week. Finally, the fitted scores for variable X_9 imply with values (-0.2955, 0, -0.0609) that the distinction between levels 2 (two or three pain-induced stationary treatments) and 3 (more than three pain-induced stationary treatments) is rather uninformative for linear prediction of the response.

Acknowledgements

We are grateful to the British German Academic Research Collaboration Programme for supporting our joint work. Computations were carried out with Matlab. We thank C.-W. Kohlmann and N. Schmitt for permission to use their data.

References

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*, p. 470. London: Chapman and Hall.
Cox, D. R. and Small, N. H. J. (1978) Testing multivariate normality. *Biometrika*, **65**, 263–272.
Cox, D. R. and Wermuth, N. (1992) Response models for mixed binary and quantitative variables. *Biometrika*, **79**, 441–461.
——— (1993) Some recent work on methods for the analysis of multivariate observational data in the social sciences. In *Proc. 7th Int. Conf. Multivariate Analysis*. Amsterdam: Elsevier. To be published.
Ezekiel, M. (1926) The determination of curvilinear regression 'surfaces' in the presence of other variables. *J. Am. Statist. Ass.*, **21**, 310–320.
Hirschfeld, H. O. (1935) A connection between correlation and contingency. *Proc. Camb. Phil. Soc.*, **31**, 520–524.
Kohlmann, C.-W., Küstner, E. and Beyer, J. (1993) Kontrollüberzeugungen und Diabeteseinstellung in Abhängigkeit von der Erkrankungsdauer. *Gesundheitspsychologie*, **1**, 32–48.
Kohlmann, C.-W., Schuhmacher, A. and Streit, R. (1987) Parental child rearing behavior and the development of trait anxiety in children: support as a moderator variable? *Anx. Res.*, **1**, 53–64.
Lauritzen, S. L. and Wermuth, N. (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, **17**, 31–75.
McFadden, J. A. (1955) Urn models of correlation. *Ann. Math. Statist.*, **26**, 478–489.
Schmitt, N. (1990) Stadieneinteilung chronischer Schmerzen. *Medical Dissertation*. Universität Mainz, Mainz.
Wermuth, N. and Cox, D. R. (1992) On the relation between interactions obtained with alternative codings of discrete variables. *Methodika*, **6**, 76–85.
Yates, F. (1948) The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, **35**, 176–181.