ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
20, 3 (1990), pp. 377–386

A. BARTKOWIAK, S. ŁUKASIK, K. CHWISTECKI and M. MRUKOWICZ (Wrocław)

INTERDEPENDENCE BETWEEN CHD RISK FACTORS EXAMINED USING LINEAR MODELS WITH AND WITHOUT TRANSFORMATION OF DATA

We try to establish a dependence pattern in large epidemiological data. We do this first in the classical way writing equations for the dependence structure and testing significance of factors introduced into the model.

The executed tests need normality of the considered variables, while in fact these do not follow the normal distribution. To meet this difficulty we applied to the predicted variables the Box–Cox transformation bringing the data nearer to normality. The dependence structure remained the same.

1. Data, the medical problem. We consider epidemiological data collected in the Coronary Heart Disease (CHD) Prevention Study conducted in Wrocław. Systematic investigations on the etiology and epidemiology of CHD are run in many countries (see, e.g., [5]). The results of these investigations are put together and analyzed in the ERICA programme coordinated by the WHO Collaborating Centre in Heidelberg. The data we consider in this paper come from the first screening in Wrocław. They comprise 6651 men working in industrial plants in Wrocław. The scope of the study is a follow-up of a chosen sample of men in working age during successive 10 years. The data we consider in this paper come from the beginning of the study. Our aim is to investigate the relationship between some continuous and categorial variables often considered when indicating risk factors for CHD.

In this paper we consider 7 continuous variables (called in the following also *predicted* or *explained variables*): BP (arterial blood pressure) systolic, BP diastolic, cholesterol, TGL (triglycerides), fraction HDL of cholesterol, glucose, uric acid. We consider 4 categorial variables: A — noise or vibration in the place of work, B — physical effort during the work, C — work in a hurry, D — daily smoked cigarettes.

2. Linear model and its parametrization. We assume the following linear model:

$$y = Xb + e, \tag{1}$$

where y is the $(n \times 1)$-vector comprising observed values of the predicted variable, X is a known $n \times m$ design matrix, the $(m \times 1)$-vector b is the vector of unknown parameters of the model, and e is an $(n \times 1)$-vector comprising errors (inadequacies) of the model. We assume first that the probability distribution of e is multivariate normal $N_n(0, \sigma^2 I)$.

In the following we take as y one after another each of the seven variables listed above.

The matrix X should take into account the considered factors and, additionally, the covariate "age". The elements of the matrix X are the same for all seven predicted variables. The construction of elements of this matrix is given below.

We consider 4 factors which possibly could influence the explained variables. These are:

A – noise or vibration in the place of work. This factor is considered in 3 levels: (1) no noise or vibration, (2) noise, (3) noise and vibration.

B – physical effort during the work. Here we distinguish 3 levels: (1) big effort, (2) medium effort, (3) small effort.

C – work in a hurry. Here we distinguish 2 levels according to the response: (1) no, (2) yes.

D – daily smoked cigarettes. Here we distinguish 3 levels: (1) he never smoked, (2) he smokes less than 10 cigarettes per day and smokes not less than 2 years, (3) he smokes more than 10 cigarettes per day longer than 2 years.

The subdivision of our data into a fourfold contingency table is presented in Table 1. One can see that, despite of having a large sample, the counts of some subclasses are very small.

TABLE 1. Counts of subclasses of individuals subdivided according to four factors: A, B, C, D ($n = 6651$)

		$C1$			$C2$		
		$D1$	$D2$	$D3$	$D1$	$D2$	$D3$
$A1$	$B1$	10	0	26	31	7	100
	$B2$	51	6	123	114	13	250
	$B3$	166	17	171	325	39	485
$A2$	$B1$	25	1	75	107	14	309
	$B2$	119	10	300	288	39	741
	$B3$	44	7	85	130	18	286
$A3$	$B1$	33	4	105	207	21	531
	$B2$	82	11	178	209	28	527
	$B3$	12	4	31	46	5	85

To determine the design matrix X we use the independent variables coding (see, e.g., [3]).

The first column of X is a column of ones. It corresponds to a parameter which is called the *grand mean.*

The next 7 columns correspond to the main effects of the factors A, B, C and D.

Let the factor A occur at a levels. Its counterpart in the matrix is constructed in a block X^A of $a-1$ columns defined rowwise (for subsequent individuals) as follows:

if for the i-th individual ($i = 1, \ldots, n$) the factor A occurs at the level j ($1 \leqslant j \leqslant a-1$), then the element x_{ij}^A is set equal to one, and all other elements in this row are set equal to zero;

if for the i-th individual the factor A occurs at the level a, then all the elements $x_{i1}^A, x_{i2}^A, \ldots, x_{i,a-1}^A$ are set equal to -1.

Next we constructed blocks corresponding to interactions of the considered factors. These were obtained by multiplying appropriate columns corresponding to the main effects.

For our data, introducing besides the main effects also double interactions between factors and taking the variable "age" as covariate, we obtained a matrix X with $m = 27$ columns connected with:

constant term – one column – the 1-st one;
main effects of A – two columns – the 2-nd and 3-rd ones;
main effects of B – two columns – the 4-th and 5-th ones;
main effects of C – one column – the 6-th one;
main effects of D – two columns – the 7-th and 8-th ones;
interactions AB – four columns: no. 9, 10, 11, 12;
interactions AC – two columns: no. 13, 14;
interactions AD – four columns: no. 15, 16, 17, 18;
interactions BC – two columns: no. 19, 20;
interactions BD – four columns: no. 21, 22, 23, 24;
interactions CD – two columns: no. 25, 26;
the covariate "age" – one column – the 27-th one.

Using this method of parametrization we obtained a parsimonious matrix X with rank very likely to be equal to m, the size of the vector b.

The model (1) is formally a linear model although it takes into account also the interactions, which are in fact nonlinearities of the considered factors.

Our goal now is to investigate whether the introduced factors and their interactions have an essential "influence" on the considered variables y. By the *influence* we mean here a statistical influence which results in the possibility of predicting y when x is known.

With the known distribution of e, the error term in (1), we can execute statistical tests verifying whether the observed statistical influence is really statistically significant, i.e., whether the introduced main effects and interaction terms are different from zero.

3. Evaluation of the importance of the considered parameters. The vector of parameters b from (1), when considered for our data, can be subdivided into 7 subgroups:

$$b' = (b_0,\ b_A,\ b_B,\ b_C,\ b_D,\ b_{\mathrm{INT}},\ b_a)'$$

with 1 (b_0), 2 (b_A), 2 (b_B), 1 (b_C), 2 (b_D), 18 (b_{INT}) and 1 (b_a) elements in subsequent groups.

Applying the method of least squares we obtain an estimate of the vector b by solving the normal equations, which are

$$(X'X)\hat{b} = X'y. \tag{2}$$

These can be solved straightforward (due to the independent variables coding, the matrix $X'X$ is likely to be of full rank).

Next, using the estimate $\hat{b}$ and the known values of X we can reconstruct the value of the variable y, obtaining an estimate $\hat{y}$ for each of the considered individuals i ($i = 1, \ldots, n$). The closeness of y_i and $\hat{y}_i$ can be judged by their difference, or generally by the residual sum of squares

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

This in turn can be compared with the total adjusted sum of squares defined as

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

The difference $\mathrm{SS}_{\mathrm{regr}} = \mathrm{SST} - \mathrm{SSE}$ is that part of the sum SST which is explained by the model (in other words, by the regression of the considered predictor variables). If this part is big as compared with SST, then the fit (adequacy) of the assumed model is good.

To obtain SSE, the residual sum of squares, we determined from (2) and from SST the augmented matrix C,

$$C = \left[\begin{array}{c|c} X'X & (X'y)' \\ \hline X'y & \mathrm{SST} \end{array}\right],$$

and applied to this matrix for $k = 1, 2, \ldots, 27$ the sweep operator executing modified Gauss–Jordan transformations (see, e.g., [4] or [1]) which introduce sequentially subsequent factors into the regression set. This done, we got directly the values SSE, and hence, by subtraction, the value $\mathrm{SS}_{\mathrm{regr}}$.

Next we calculated the index z defned as follows:

$$z = 100\,\mathrm{SS}_{\mathrm{regr}}/\mathrm{SST}.$$

One can see that z is the percent of the total variance SST explained by the introduced model.

These calculations were carried out, one after another, for each of the seven variables, i.e., BP systolic (BPs), BP diastolic (BPd), cholesterol (Ch), TGL, fraction HDL, glucose (GL) and uric acid (UA). The values of z obtained for these variables are given in the upper part of Table 2 (the other parts of Table 2 are explained in the next section of this paper).

TABLE 2. Percent of total SST explained by the introduced model with the covariate "age" introduced last. The values of the F-statistic are given in parentheses. $\times$ – significant at the level $\alpha = 0.05$, $\times\times$ – significant at the level $\alpha = 0.01$

y Model	1 BPs	2 BPd	3 Ch	4 TGL	5 HDL	6 GL	7 UA
All 26 explanatory variables	3.4 $\times\times$ (8.9)	2.8 $\times\times$ (7.3)	1.3 $\times\times$ (3.5)	1.0 $\times\times$ (2.6)	2.6 $\times\times$ (6.9)	1.4 $\times\times$ (3.7)	3.4 $\times\times$ (9.1)
A	0.02	0.10 $\times$	0.44 $\times$	0.08	0.12	0.01	0.27 $\times\times$
B	0.13 $\times$	0.04	0.03	0.17 $\times\times$	1.45 $\times\times$	0.06	0.46 $\times\times$
C	0.01	0.01	0.00	0.06 $\times$	0.31 $\times\times$	0.00	0.09 $\times\times$
D	0.26 $\times\times$	0.36 $\times\times$	0.21 $\times\times$	0.06	0.15 $\times\times$	0.22 $\times\times$	2.31 $\times\times$
INT	0.18	0.15	0.54 $\times$	0.62 $\times\times$	0.55 $\times\times$	0.26	0.28
Age	2.8 $\times\times$	2.1 $\times\times$	0.1	0.0	0.0	0.8	0.0
D	0.25 $\times\times$	0.40 $\times\times$	0.15 $\times\times$	0.03	0.25 $\times\times$	0.22 $\times\times$	2.56 $\times\times$
C	0.01	0.01	0.00	0.04	0.26 $\times\times$	0.00	0.07 $\times\times$
B	0.11 $\times$	0.03	0.23 $\times\times$	0.28 $\times\times$	1.44 $\times\times$	0.05	0.46 $\times\times$
A	0.04	0.06	0.30 $\times\times$	0.00	0.08	0.02	0.04
INT			The same figures as above.				
Age			The same figures as above.				

Looking at the values of z shown in Table 2 one can see that the percent of the explained variance oscillates between the value 1.0 (for TGL) and 3.4 (for BPs). So the introduced model with 26 variables explains only a minuscule part of the total variance.

4. Statistical tests in the assumed model. Are the results univocal? Looking at the results presented in Table 2 we could have doubts whether they reflect true effects of the introduced factors. The stated reduction of SST could be purely random. To exclude the last possibility we carry out statistical tests for proving the statistical significance of the results.

First we test the hypothesis

$$H_0\colon (b_A, b_B, b_C, b_D, b_{\mathrm{INT}}, b_a) = 0.$$

To do this we calculate the F-statistic

$$F = \frac{\mathrm{SS}_{\mathrm{regr}}}{m-1} : \frac{\mathrm{SSE}}{n-m-1}, \tag{3}$$

where m is the rank of the matrix X.

If H_0 is true and the errors e are independent and distributed normally, then the F-statistic given by (3) has an F-distribution with $m-1$ and $n-m-1$ degrees of freedom. In our example $m = 27$ and $n = 6651$.

For our data the assumption of normality is generally not satisfied; see, e.g., the coefficients of asymmetry and kurtosis in Table 4. Despite of this we calculate the F-statistic according to (3) and execute the classical tests. The calculated values of the F-statistic are given also in Table 2. One can see that for all seven considered variables the results are highly significant, which means that the introduced model gives statistically significant reductions of the total variance SST.

What is the participation of the considered factors in this statistical significance?

According to the order of introducing the factors by the sweep operator into the regression set we can split the explained variance into some components. Let the order applied be: A, B, C, D, INTERACTIONS (INT), age. In this case the total explained variance SS_{regr} can be splitted into the following parts:

$$(4)\quad SS_{regr} = SS(b_A) + SS(b_B/b_A) + \ldots + SS(b_{INT}/(b_A, b_B, b_C, b_D)) + SS(b_a/(b_A, \ldots, b_{INT})).$$

This done, we can test the following nested hypotheses:

$$H_A\colon\ b_A = 0 \quad \text{(after swept in } b_0),$$

$$H_B\colon\ b_B = 0 \quad \text{(after swept in } b_0, b_A),$$

$$H_C\colon\ b_C = 0 \quad \text{(after swept in } b_0, b_A, b_B),$$

$$H_D\colon\ b_D = 0 \quad \text{(after swept in } b_0, b_A, b_B, b_C),$$

$$H_{INT}\colon\ b_{INT} = 0 \quad \text{(after swept in } b_0, b_A, b_B, b_C, b_D).$$

For example, to test the hypothesis H_A, we have to calculate an F-statistic of the form

$$F_A = \frac{SS(b_A)}{a-1} : \frac{SSE}{n-m}.$$

To test H_B we have to calculate

$$F_B = \frac{SS(b_B/b_A)}{b-1} : \frac{SSE}{n-m},$$

where b is the number of independent columns in the block X^B (this block comprises the columns of the matrix X which correspond to the main effects of the factor B).

Using the subdivision of SS_{regr} given in (4) we calculated also the values of z, denoting the percents of SST explained by the considered factors when introduced into the regression set in a prescribed order. The appropriate values of z are shown in Table 2. The statistical significance is denoted by $\times$ (significance at the level 0.05) or by $\times\times$ (significance at the level 0.01).

Obviously, the decomposition of SS_{regr} into parts accounted to various considered factors depends on the order in which the variables are swept in into the regression set. In Table 2 we show another decomposition of the percents of explained variance. Both decompositions shown in Table 2 were done in such a way that the covariate "age" was introduced into the regression set last. We made two other decompositions introducing the variable "age" first. The results of these decompositions are shown in Table 3.

TABLE 3. Percent of total SST explained by the introduced model with the covariate "age" introduced first

y Model	1 BPs	2 BPd	3 Ch	4 TGL	5 HDL	6 GL	7 UA
Age	2.8××	2.1××	0.1××	0.0	0.0	0.9××	0.0
A	0.01	0.08	0.45××	0.08	0.12×	0.01	0.27××
B	0.11×	0.05	0.03	0.17××	1.46××	0.04	0.46××
C	0.00	0.00	0.00	0.06×	0.32××	0.00	0.09×
D	0.26××	0.37××	0.21××	0.06	0.15××	0.22××	2.31××
INT	0.16	0.15	0.54×	0.62××	0.55××	0.28	0.27
Age	The same figures as above.						
D	0.26××	0.41××	0.15××	0.03	0.25××	0.22××	2.56××
C	0.00	0.00	0.00	0.05	0.27××	0.00	0.07×
B	0.10×	0.05	0.23××	0.28××	1.45××	0.04	0.45××
A	0.02	0.04	0.30××	0.00	0.08	0.02	0.04
INT	The same figures as above.						

Looking at Tables 2 and 3 one can state that the significance pattern is similar in all four decompositions. The interactions are significant for variables 3, 4 and 5. The covariate "age" is statistically significant for the variables BP systolic and BP diastolic, cholesterol and glucose regardless of the order in which the factors appear in the decomposition. Also the significance pattern established for other factors is quite stable and does not depend on the order of these factors in the given decomposition.

Considering generally the values of z accounted by the factors considered in our model, we state that the largest values of z are accounted by the covariate "age" when associated with systolic and diastolic blood pressure. The second most important factor (for which the values of z are the largest) is D (smoking habitudes). It is highly significant for all variables y except TGL.

Now we are going to transform the data and bring them nearer to normality. After this we shall repeat the calculations and see whether we obtain the same results.

To verify the assumption on the normality of errors appearing in the model (1) we calculated the coefficients of asymmetry (γ_1) and kurtosis (k) defined by the formulae

$$\gamma_1 = \left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^3\right]s^{-3},$$

$$k = \left\{\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^4\right]s^{-4}\right\}-3.0$$

with

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

For normal distributions the values γ_1 and k are expected to be equal to zero.

We calculated these coefficients twice: (a) for the direct observed values $y_1, \ldots, y_n$ and (b) for the residuals $\hat{e}_1, \ldots, \hat{e}_n$, where $\hat{e}_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$, with $\hat{y}_i$ being the expected value of y_i assuming the model (1). The coefficients γ_1 and k calculated using methods (a) and (b) are shown in Table 4.

TABLE 4. Coefficients of asymmetry γ_1 and kurtosis k calculated for original and transformed data. λ is the parameter of the Box–Cox transformation

Variable	Original data		Transformed data		λ
	direct y	residuals	direct y	residuals	
Coefficients of asymmetry γ_1					
BPs	0.96	0.89	0.06	0.01	−.8
BPd	0.81	0.75	0.06	0.02	−.3
Ch	0.78	0.76	−0.01	−0.03	ln
TGL	2.50	2.50	−0.35	−0.32	−.6
HDL	1.14	1.09	−0.00	−0.03	−.3
GL	6.07	6.03	−0.16	−0.20	−1.5
UA	0.54	0.54	0.15	0.15	0.6
Coefficients of kurtosis k					
BPs	2.01	1.93	0.54	0.54	
BPd	3.04	2.78	0.57	0.51	
Ch	2.49	2.52	0.57	0.60	
TGL	8.50	8.58	−0.04	−0.03	
HDL	2.33	2.26	0.21	0.22	
GL	61.03	60.86	4.91	4.77	
UA	1.23	1.27	1.22	1.27	

We see that the distributions, especially TGL and glucose, exhibit a considerable asymmetry and are heavy-tailed ($k > 0$).

We see that the difference between the coefficients calculated from the original values and from the residuals is very small. This is due to the fact that the introduced model explains only a minuscule part of SST, the total variability of y.

Next we tried to bring the data nearer to normality transforming them by the use of the formula proposed by Box and Cox [2]:

$$s = \begin{cases} (y^{\lambda}-1)/\lambda & \text{for } \lambda \neq 0, \\ \ln y & \text{for } \lambda = 0. \end{cases}$$

By trial and error we found such values of λ which transformed the values of the variables $y_1, y_2, \ldots, y_7$ so that the coefficient of asymmetry was practically equal to zero and the coefficient of kurtosis was as small as possible. The values of λ satisfying both these criteria are given in the last column of Table 4.

The coefficients γ_1 and k calculated for the transformed data are given again in Table 4. One can see that – except for glucose and perhaps uric acid – the coefficients of kurtosis are now very near to zero.

We repeated the calculations of SSE for the transformed data. The percents $z = 100\ SS_{regr}/SST$ are shown in Table 5. The decomposition of the

TABLE 5. Percent of variance explained by the model (1) with 27 variables for untransformed data. Model with $m = 27$ variables: the grand mean, 4 factors, their interactions and one covariate

Variable	Data untransformed	Data transformed
1	3.36	3.12
2	2.78	2.63
3	1.34	1.24
4	1.00	1.75
5	2.65	2.47
6	1.42	1.58
7	3.44	3.46

percents is shown in Table 6. We see that these percents for transformed data are similar to those calculated for untransformed data. It follows that the nonnormality of errors in the model (1) had virtually no effects on the results of the tests.

TABLE 6. Percent of total variance explained by the introduced model. Calculations on transformed data according to two orders of introducing the factors into the regression set

Variable Factor	1 BPs	2 BPd	3 Ch	4 TGL	5 HDL	6 GL	7 UA
The whole model	3.12	2.63	1.24	1.75	2.47	1.58	3.46
Age	2.52	2.00	0.12	0.05	0.03	1.04	0.05
A	0.03	0.09	0.41 × ×	0.27	0.11	0.03	0.28 × ×
B	0.08	0.05	0.03	0.51 × ×	1.38	0.07	0.49 × ×
C	0.00	0.00	0.00	0.13 × ×	0.32 × ×	0.01	0.09 × ×
D	0.29 × ×	0.36 × ×	0.19 × ×	0.12 ×	0.06	0.29 × ×	2.28 × ×
INT	0.19	0.13	0.49 ×	0.67 × ×	0.56 × ×	0.15	0.27
Age	The same figures as above.						
D	0.28 × ×	0.40 × ×	0.14 × ×	0.07 ×	0.13 ×	0.28 × ×	2.53 × ×
C	0.00	0.00	0.00	0.09 × ×	0.28 × ×	0.01	0.07 × ×
B	0.07	0.05	0.22	0.85 × ×	1.40 × ×	0.04	0.48 ×
A	0.05	0.05	0.27 × ×	0.02 × ×	0.08 ×	0.07	0.05
INT	The same figures as above.						

References

[1] A. Bartkowiak, *SABA, An Algol package for statistical data analysis on the ODRA 1305 computer*, Universitas Wratislaviensis, Wrocław 1984.

[2] G. E. P. Box and D. R. Cox, *An analysis of transformations with discussion*, J. Roy. Statist. Soc. B 26 (1964), pp. 211–252.

[3] R. L. Horton, *Data Analysis in the Social and Behavioral Sciences*, McGraw-Hill, New York 1978.

[4] R. I. Jennrich, *Stepwise regression*, in: K. Enslein, A. Ralston and H. Wilf (Eds.), *Statistical Methods for Digital Computers*, J. Wiley, New York 1977.

[5] A. Keys (Ed.), *Seven Countries, A Multivariate Analysis of Coronary Heart Diseases and Death*, Harvard Univ. Press, Cambridge, Mass., 1980.

ANNA BARTKOWIAK
INSTITUTE OF COMPUTER SCIENCE
UNIVERSITY OF WROCŁAW
UL. PRZESMYCKIEGO 20
51-151 WROCŁAW, POLAND

SEWERYN ŁUKASIK, KRZYSZTOF CHWISTECKI, MIROSŁAW MRUKOWICZ
DEPARTMENT OF CARDIOLOGY
MEDICAL ACADEMY WROCŁAW
UL. PASTEURA 4
50-367 WROCŁAW, POLAND

Received on 1987.11.24