Listy Biometryczne – Biometrical Letters
Vol. 32 (1995), No. 2, 61-80

Regression modeling of the Polish mortality data 1989-1991: mortality of men

Anna Bartkowiak and Witold Kupść

Institute of Computer Science, University of Wrocław, Przesmyckiego 20, 51-151 Wrocław
Department of Epidemiology and Prevention of CVD, Institute of Cardiology, Niemodlińska 33, 04-635 Warszawa

Summary

We consider the total mortality rates of male population as observed in 49 voivodeships of Poland. We look for a linear regression explaining the mortality in men (Y) as a linear function of nine variates denoted in the following as $X_1,\ldots,X_9$ and reported for each voivodeship. The squared multiple correlation coefficient for the established regression is about 0.75. However, constructing a full regression of all the potential predictors and looking at the t statistics indicating the importance of the considered explanatory variables, we do not obtain an unique and univocal indication of which variables are good predictors for the considered Y-variable, and which are not.

It should be stressed, that the importance of variables – as indicated by the t statistics – is valid only in the context of the established regression; variables appearing in the established "full" regression as "nonsignificant" (in the meaning of the t statistics) can have big predictive power not indicated by the full regression.

We illustrate this by finding alternative subsets of variables. This is done by some detailed considerations based on exploratory data analysis (EDA) techniques and also by detecting near collinearities amongst the variables using a method proposed by Hawkins (1973), which allows to deduce from the established relations which variables are exchangeable.

1. Introduction, the data

The mortality rates in Poland exhibit considerable differentiation among the administrative regions (voivodeships). It is supposed that the observed variability

Key words: LSE (Least Squares Error) regression, subsets of predictors, alternative subsets, exploratory data analysis, mortality of men.

might be related to several environmental variables which differ among voivodeships.

To throw some light on this problem epidemiologists try to find a model for this phenomenon by using some multivariate statistical techniques. One of such techniques is provided by multiple regression, in which the mortality rate is assumed as the response (predicted) variable, and the environmental variables as the explanatory ones.

For our considerations we took data from the data base created and maintained in the Department of Epidemiology and Prevention of CVD, Institute of Cardiology, Warsaw (Kupść and Jasiński, 1991). From these data we took 9 variables as predictor variables and the mortality rate of men as the predicted variable. The 9 predictors were selected from a larger set of potential risk factors.

The predictor variables, denoted in the following as $X_1,\ldots,X_9$, are (the terms in quotation marks show the variable labels)

X_1: "artf" – artificial fertilizers (kg/ha),
X_2: "u18y" – % of population under age of 18 years,
X_3: "divr" – divorce rate (per 1000 pop.),
X_4: "nati" – natural increase (per 1000 pop.)
X_5: "emin" – employment in industry (% per 1000 men),
X_6: "nmed" – number of medical doctors (per 10000 pop.),
X_7: "mmar" – % of married men,
X_8: "wmar" – % of married women,
X_9: "sece" – % of persons with at least secondary education level.

The predicted variable Y labeled "ymog" denotes the mean total standardized mortality rate of men for the years 1989–1991.

The values of these variables were reported in 49 voivodeships of Poland. Data taken for further analysis are in the form:

$\mathbf{X}_{n\times p}$ – the data table comprising $n = 49$ rows (voivodeships) and $p = 9$ columns corresponding to the considered explanatory variables,

$\mathbf{y}_{n\times 1}$ – the column vector comprising values of the variable Y as reported in the $n = 49$ voivodeships of Poland.

In Table 1 we show the data for three voivodeships: Warszawa, Wrocław and Łódź. In that table Y_w and Y_m denote mortality rates for women and men age-adjusted according to the WHO world standard. In this paper we will analyze only the mortality rates for men – the respective rates for women are shown in Table 1 for comparative purpose only. To make the comparisons between these voivodeships easier, we show in the same table also the respective *normalized* data values ("normalized" means that from each data value the appropriate mean was subtracted, and the obtained difference was divided by the standard deviation of the variable).

Table 1
Data vectors for 3 selected voivodeships: no. 1 (Warsaw), no. 24 (Łódź) and no. 47 (Wrocław)

No.	Voivodeship	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y_w	Y_m
a) Original values												
1	Warsaw	139.0	24.1	1.3	–1.1	29.6	41.5	68.4	58.9	53.6	529.52	972.89
24	Łódź	140.0	23.3	2.1	–3.8	41.3	36.9	70.4	59.2	39.5	599.00	1131.92
47	Wrocław	250.0	27.6	1.5	3.1	26.9	33.7	67.0	61.4	40.2	526.11	999.31
b) Normalized values												
1	Warsaw	–0.46	–2.90	0.73	–2.41	0.80	3.27	0.63	–2.39	4.04	0.08	–0.64
24	Łódź	–0.44	–3.26	2.79	–3.48	2.37	2.62	2.11	–2.22	1.79	2.12	2.07
47	Wrocław	1.64	–1.31	1.25	–0.75	0.44	2.17	–0.40	–1.00	1.90	–0.02	–0.19

In the following we will consider two groups of data, called Group I and Group II. The first group comprises the full set of data containing n=49 voivodeships. From a preliminary analysis it appeared that the voivodeship Łódź is atypical. Some regression diagnostics (not shown here) indicated that this voivodeship might be very influential in the calculated regression. To throw more light on the impact of this voivodeship in the evaluated regression we have removed the data vector describing Łódź from the data – and so we got the second group, called Group II, with n=48, containing all the voivodeships but the vector *no.* 24 identified with Łódź.

2. Preliminary investigation of the structure of the data by inspecting scatterplot matrices and biplots

To see whether the data exhibit some unusual pattern, we performed firstly a kind of exploratory data analysis (EDA). We have chosen for this purpose (a) – scatterplot matrices, and (b) – biplots. A scatterplot matrix permits identifying outliers in pairs of variables by looking at two-dimensional scatterplots exhibiting the values for fixed pairs of the variables. A biplot permits exhibiting simultaneously the relations (correlations) between variables, between individuals, and both between variables and individuals. Also, often it permits identifying gross multivariate outliers.

In the following we present the results of our analysis, when using these two EDA techniques.

2.1. Inspecting the data using scatterplot matrices

To draw our scatterplot matrices we have used the system XLispStat (Tierney, 1990). In Fig. 1 we show the scatterplot matrix for all pairs of variables $X_1,\ldots,X_9, Y$.

We are mostly concerned with the first row of the plots, i.e. in the scatterplots representing scatterdiagrams of the pairs (X_1, Y), (X_2, Y),..., (X_9, Y). Looking at these scatterdiagrams one can see, that there is a positive correlation between the variables (X_1, Y), (X_3, Y), (X_5, Y), (X_7, Y), while the correlation between (X_2, Y)

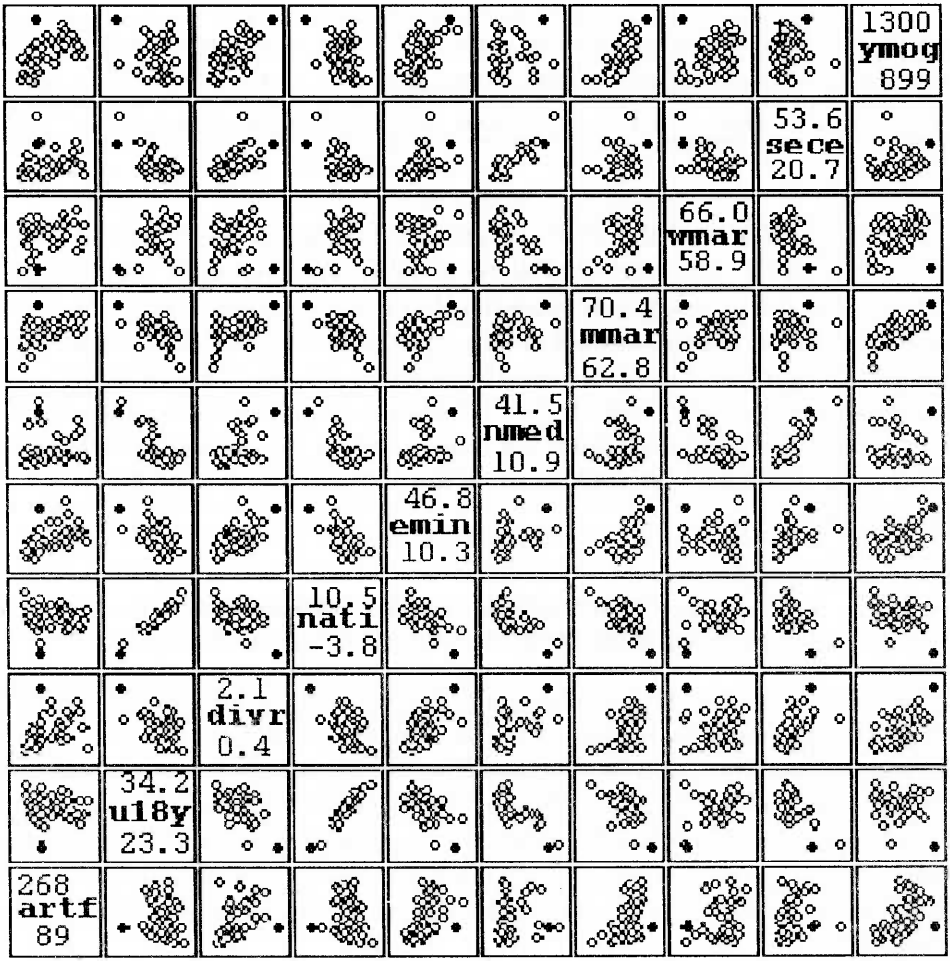

Figure 1. Scatterplot matrix from the variables $X_1,\ldots,X_9, Y$. Group I of data.

Table 2.
Correlation matrix evaluated for 11 variables

1	artf	1.00										
2	u18y	–0.06	1.00									
3	divr	0.47	–0.46	1.00								
4	nati	–0.08	0.96	–0.45	1.00							
5	emin	0.47	–0.54	0.56	–0.55	1.00						
6	nmed	0.08	–0.75	0.48	–0.65	0.37	1.00					
7	mmar	0.45	–0.43	0.50	–0.60	0.54	0.15	1.00				
8	wmar	0.17	0.42	–0.01	0.28	–0.14	–0.56	0.33	1.00			
9	sece	0.27	–0.62	0.62	–0.50	0.45	0.88	0.19	–0.52	1.00		
10	ywog	0.72	–0.31	0.63	–0.39	0.72	0.16	0.66	0.09	0.33	1.00	
11	ymog	0.62	–0.20	0.65	–0.32	0.54	–0.01	0.68	0.31	0.14	0.88	1.00
		1	2	3	4	5	6	7	8	9	10	11

and (X_4, Y) exhibits a negative trend, very blurred though. The mostly correlated pair is (X_2, X_4).

The full matrix of correlations between the considered variables (with the variable "ywog" denoting the standardized mortality rate for women) is shown in Table 2.

Looking at the scatterplots exhibited in Fig. 1 we can see distinctly one outlier, which is identified as the point (data row) *no.* 24 corresponding to the voivodeship Łódź. This voivodeship has really atypical demographic indices and should be investigated separately for its influence exerted in the to be carried out regression analysis. To find out the real influence of this data vector we have established for our analysis two groups of data as explained in Section 1 and have investigated them in parallel.

2.2. Inspecting mutual correlations exhibited in biplots

A biplot is a method of presenting in the same plot both variables and individuals in such a way that their mutual relationships can be revealed. The term "biplot" means that this is a dual representation, both of variables and of individuals, put together in the same plot. Usually the individuals are marked as points, and the variables as vectors.

The principles of constructing a biplot can be found in the papers of Gabriel (1982, 1990), also in the books by Jolliffe (1986) or Krzanowski (1988). For drawing the biplots shown below we have used the program BIPLOT from the package SFAX (Bartkowiak, 1995). The algorithm used in this program is described by Bartkowiak and Szustalewicz (1995).

In Fig. 2 and Fig. 3 we show two biplots constructed from the variables $X_1,\ldots,X_9, Y$. The biplot shown in Fig. 2 was constructed using Group I comprising all voivodeships; the biplot shown in Fig. 3 is based on Group II. Both biplots were constructed from correlation matrices.

The dots in the biplots numbered 1–49 in Fig 2 and 1–23, 25–49 in Fig 3 represent the voivodeships. The vectors labeled $X_1,\ldots,X_9, Y$ represent the considered variables. Generally the points and the vectors in the biplot plane represent projections from the multivariate space onto the plane of the first two principal components.

Let us remind the decomposition formula established in the principal components theory (described e.g. by Morrison, 1967, or Jolliffe, 1986):

$$\mathbf{R} = \sum_{i=1}^{m} \lambda_i \mathbf{a}_i \mathbf{a}_i^T \ .$$

In this formula $\mathbf{R}$ denotes the correlation matrix of size $m \times m$, $m = p+1$; λ_i and $\mathbf{a}_i$ are the eigenvalues and the eigenvectors obtained for $\mathbf{R}$. From the formula we can deduce how much of the diagonal of $\mathbf{R}$ (i.e. of the *trace* of $\mathbf{R}$) is reproduced by consecutive principal components.

Returning to our biplots shown in Fig 2 and Fig 3: the goodness (adequacy) of the representation in the exhibited plots can be measured by the percentage of exhaustion of the trace – when taking into account the first two principal axes, i.e. the first two components in the decomposition formula shown above. For the biplots shown in Fig. 2 and 3 the goodness of representation is about 70 %; the detailed numbers are shown in the table below.

Figure:	Fig.2	Fig.3
1st axis:	47.46	42.83
2nd axis:	25.05	27.45
both axes:	72.51	70.28
out of Total	100.00	100.00

Our computational program has used such algorithm (see Bartkowiak and Szustalewicz, 1995) that the vectors-variables, when computed from correlation matrices, are of unit length.

When working with biplots it is important to know which variables are represented in the constructed plots fairly, and which are not. This can be seen when looking at the reproduction of $\mathbf{R}$ by the diagonals of the rank one matrices $\lambda_1 \mathbf{a}_1 \mathbf{a}_1^T$ and $\lambda_2 \mathbf{a}_2 \mathbf{a}_2^T$ and considering separately each of the elements of the respective diagonals. We could also ask, how much would be gained by adding a third

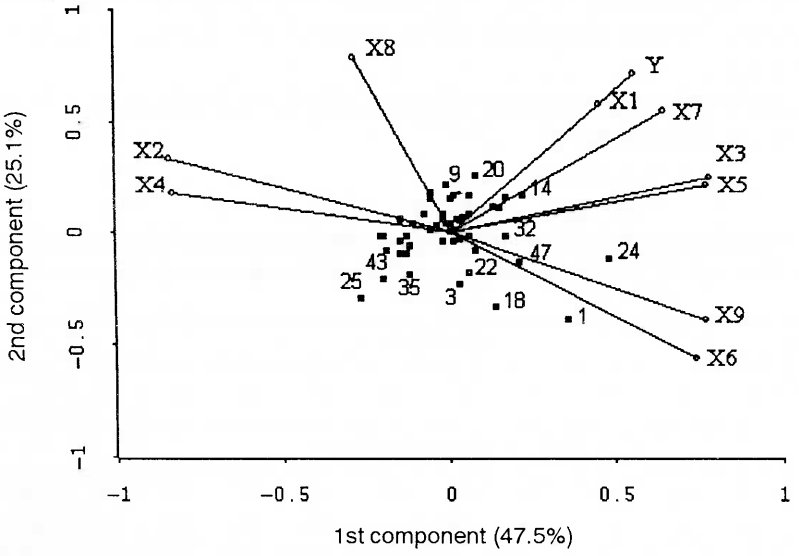

Figure 2. Biplot from the variables $X_1,\ldots,X_9,Y$. Group I of data. Points-dots numbered 1-49 represent the 49 voivodeships of Poland.

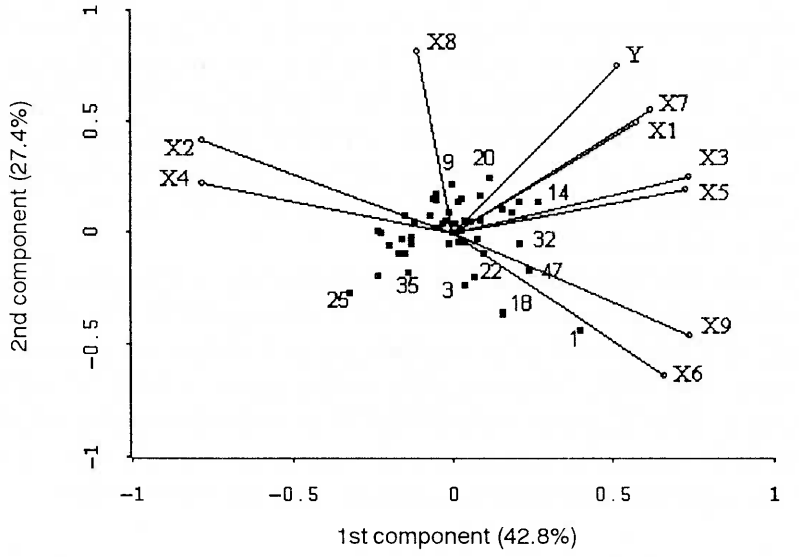

Figure 3. Biplot from the variables $X_1,\ldots,X_9,Y$. Group II of data in which the voivodeship no. 24 Łódź has been omitted.

dimension to the biplot, i.e. by constructing a three-dimensional biplot (this is possible when using appropriate software). The gain of adding the third dimension can be gauged by the magnitude of the values of the diagonal elements of the matrix $\lambda_3 \mathbf{a}_3 \mathbf{a}_3^T$.

The cumulative reproduction of the diagonal of **R** by the first three principal components is shown in Table 3.

Table 3

Cumulative reproduction of the diagonal of the correlation matrix evaluated for Data Set I and Data Set II – by use of first two and first three principal components

Principal components	Variables										% Total trace
	1	2	3	4	5	6	7	8	9	11	
Plot 2 (n=49, with Łódź)											
First two	.54	.82	.67	.73	.63	.87	.72	.70	.75	.83	72.5
First three	.78	.95	.74	.97	.63	.88	.86	.75	.90	.83	83.0
Plot 2 (n=48, without Łódź)											
First two	.59	.78	.61	.65	.58	.85	.70	.68	.76	.83	70.3
First three	.76	.95	.74	.97	.58	.87	.84	.75	.90	.84	82.1

Looking at the values shown in Table 3 one can state that the representation shown in the biplots is not extremely good. The variables X_1, X_3 and X_5 have the worst representation. This can be seen when looking at the length of the respective vectors shown in the plots in Fig. 2 and Fig. 3. Adding the third dimension (i.e. constructing a 3-dimensional biplot) would be helpful, although also this would not yield a very good representation – this follows from the fact that a 3-dimensional biplot would reproduce 83.0 % and 82.1 % of total trace of the respective correlation matrices.

Having in mind the regression of Y on the variables $X_1,\ldots,X_9$ we see in both biplots the same pattern: the most closest to Y are X_1, X_7, X_3 and $X5$. The variables X_2 and X_4 exhibit a week negative correlation and in both plots are located as opposite to the bunch of the other variables. Other characteristics of the two biplots are very similar. Thus we feel justified to conclude that, generally, the voivodeship Łódź, in spite of being an outlier, does not seem to exert any essential influence on the structure (relationship) amongst the considered variables.

3. Calculations of the regression using the method of least squares error (LSE)

The classical linear regression model is given by the equation

$$Y = b_0 + b_1 X_1 + \ldots + b_p X_p + e \ , \tag{1}$$

with $e \sim N(0, \sigma^2)$.

Proceeding in the classical way, i.e. using the least squares error (LSE) method we obtain (see, e.g., Chatterjee and Hadi, 1988)

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \ , \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \ , \tag{3}$$

$$\mathrm{var}(\hat{\mathbf{b}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \ . \tag{4}$$

In the equations (2–4) above $\mathbf{X}$ denotes the augmented data matrix $[\mathbf{1}, \mathbf{X}]$, with the unit vector $\mathbf{1}_n$ corresponding to the intercept b_0 appearing in (1), and $\mathbf{b}$ denotes the vector $(b_0, b_1, \ldots, b_p)^T$.

Let the symbols $s^{(00)}, s^{(11)}, \ldots, s^{(pp)}$ denote the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$. The Studentized estimates of the regression coefficients $b_0, b_1, \ldots, b_p$ appearing in (1) are

$$t_i = \frac{\hat{b}_i}{\hat{\sigma}\sqrt{s^{(ii)}}} \tag{5}$$

We have used these formulae to obtain the estimates of the regression of Y (="ymog") on the variables $X_1, \ldots, X_p$ for the data presented above in Section 1 as Group I and Group II. The results of calculations are shown in Table 4.

The calculations were carried out in parallel for the groups I and II of the data. We have calculated the ordinary LSE regression and two robust regressions: the α-trimmed regression and a robust regression using Huber's weights. All the 3 regressions were calculated using the programs NNREG, RK and STEFF from the package SFAX (Bartkowiak, 1995).

The program RK is based essentially on the algorithm described by Antoch and Bartkowiak (1988). For our calculations we have used the trimming constant $\alpha = 0.08$ (roughly speaking, the constant α denotes fraction of discarded observations, when establishing the final regression estimate).

The program STEFF uses the iteratively reweighted Beaton-Tukey algorithm with Huber's weights as described by Li (1985) or Bartkowiak (1992). In our calculations we have used the accuracy coefficient $\varepsilon = 0.05$ and the tuning constant $\kappa = 1.8$ (A short explanation of ε and κ: The iterative process stops, when

Table 4

Studentized regression coefficients when considering two sets of explanatory variables ($X_1,\ldots,X_9$ and $X_1,X_3,X_5,\ldots,X_9$) and using the least squares error (LSE), α-trimmed (α-tr) and iteratively reweighted with Huber weights (Hub) methods of estimation of the regression coefficients. *RR* denotes multiple correlation coefficient.

Variable	Full set $X_1,\ldots,X_9$			Only $X_1,X_3,X_5,\ldots,X_9$		
	LSE	α–tr	Hub	LSE	α–tr	Hub
Group I, n = 49						
1 – artf	2.76	2.76	3.87	2.39	2.30	2.99
2 – u18y	1.90	1.90	2.63	–	–	–
3 – divr	4.26	4.26	6.04	4.29	4.75	5.48
4 – nati	–2.13	–2.13	–2.88	–	–	–
5 – emin	1.09	1.09	1.64	1.07	1.37	1.72
6 – nmed	–1.74	–1.74	–2.32	–1.45	–1.43	–1.80
7 – mmar	0.19	0.19	0.44	**2.88**	**3.06**	**3.54**
8 – wmar	–0.00	–0.00	–0.23	–0.48	–0.68	–0.68
9 – sece	0.11	0.11	0.09	–0.84	–0.99	–1.10
RR	0.77	0.77	–	0.75	0.78	–
Group II, n = 48						
1 – artf	3.16	3.24	4.29	2.78	2.74	3.81
2 – u18y	1.79	1.97	2.84	–	–	–
3 – divr	3.00	3.10	3.95	3.05	3.01	3.71
4 – nati	–2.10	–2.26	–2.87	–	–	–
5 – emin	0.96	1.06	1.63	0.99	1.15	1.91
6 – nmed	–2.15	–2.18	–3.14	–1.85	–1.61	–2.64
7 – mmar	–0.19	–0.30	–0.40	**2.18**	**2.24**	**2.66**
8 – wmar	0.76	0.93	1.23	0.33	0.27	0.53
9 – sece	0.73	0.85	1.28	–0.12	–0.29	0.08
RR	0.77	0.79	–	0.73	0.74	–

the estimates of **b** obtained in two subsequent iterations differ less then ε. Data vectors, for which the absolute standardized residuals from the regression established in subsequent iteration exceed the constant κ, obtain lesser weight in next iteration).

In Table 4 we show the results of calculations for the two groups of the data. These are: the Studentized regression coefficients – computed according to formula (5) – and RR, the squared multiple correlation coefficient measuring the goodness of fit. This is done taking into account the full set $(X_1,\ldots,X_9)$ and a reduced set $(X_1, X_3, X_5,\ldots,X_9)$ of the predictors.

Variables for which the Studentized coefficients are in absolute value greater than 2.0 can be judged as *significant* in the evaluated regression.

The results shown in Table 4 can be analyzed under several aspects. We may be interested in comparing generally the results obtained for group I and group II of the data. Alternatively, we may want to compare the (Studentized) regression coefficients established by the LSE and robust methods. Also, we may be interested in comparing the goodness of fit and importance of the predictors when introducing into the regression equation all the 9 predictors and a reduced set of 7 predictors only. All these aspects will be considered in details in next 3 subsections.

3.1. Comparing results for groups I and II of the data

The results obtained for the two sets look much alike. This concerns the Studentized regression coefficients and the squared multiple correlation coefficient RR. Therefore it seems that the voivodeship Łódź, in spite of being an outlier, does not have an essential influence on the calculated regression – we obtained very similar (Studentized) regression coefficients.

3.2. Comparing results obtained by the LSE and the robust methods

The α-trimmed regression yields either the same results, or emphasizes slightly more the importance of the variables in the evaluated regression.

The Huberized regression puts generally more importance on the considered predictors, what is deduced from the fact that the Studentized regression coefficients are larger.

3.3. Comparing results obtained for the sets $X_1,\ldots,X_9$ and $X_1, X_3, X_5,\ldots,X_9$

When taking into account all the predictor variables $X_1,\ldots,X_9$ we conclude that the significant variables are: X_1, X_2, X_3, X_4. The variables X_5, X_7, X_8, X_9 exhibit decidedly small Studentized regression coefficients and thus can be judged as "unimportant" in the considered context. The variable X_6 (="nmed") is not always formally established as "significant" although its Studentized regression coefficients are relatively high.

The role of the variables X_2 and X_4 is somehow quizzical: both are important, however looking at the signs of the coefficients, we see that the variable X_2 has a positive sign (hence positive impact on Y), the respective Studentized regression coefficients being equal to 1.90, 1.90, 2.63 in Group I and 1.79, 1.97, 2.84 in Group

II. This seems to be in contradiction with the results obtained from the former exploratory analysis where we have seen both in the scatterplot matrices (Fig. 1) and in the biplots (Fig. 2 and 3), that the variable X_2 is negatively correlated with Y. One of the referees has pointed out that such unexpected result could be caused by the fact of introducing two highly correlated variables into the regression model: then only the unexplained variability may be accounted for by each explanatory variable.

Having X_2 and X_4 in the regression model was not so much interesting for us: it seems quite obvious that when the population is younger, then the mortality rate should be lower. Our aim was to bring into light the role of the other environmental variables in predicting the mortality rate. Therefore we have simply dropped the variables X_2 and X_4 from the set of predictors and decided to consider in parallel to the full set of predictors also the reduced set $X_1, X_3, X_5, \ldots, X_9$.

When considering the reduced subset $X_1, X_3, X_5, \ldots, X_9$ we obtain a change of importance of the variable X_7 (i.e. "mmar"): it becomes now very important in the regression. Looking at the correlation coefficients we find that $r(X_7, X_2) = -0.4318$, $r(X_7, X_4) = -0.5955$ and $r(X_7, Y) = 0.6764$. So it appears, that the variable X_7 has quite a high correlation with the variable Y, however its importance in the regression $Y = b_0 + b_1X_1 + \ldots + b_9X_9$ was suppressed by a linear combination of the variables X_2 and X_4, none of which, when considered alone, is so highly correlated with Y as the variable X_7 is.

We will return to the problem of substituting some variables by others when speaking on finding alternative subsets by inspecting near collinearities found when rotating eigenvectors derived by the method of principal components. This will be considered in detail in Section 4. Before doing that we will apply to our data some classical search methods for optimal or quasi-optimal subsets of predictors.

3.4. Finding relevant subsets by stepwise and all subset search methods

To find the subsets of variables that matter in prediction of the variable Y we have applied the stepwise and the optimal subset search methods to find some quasi-optimal or the optimal subset from the considered predictors. This was done using the jerking algorithm or the optimal subset search algorithm implemented in the program NNREG from the package SFAX (Bartkowiak, 1995). Some results of the search are shown in Table 5.

Table 5

Goodness of fit of alternative regressions. Variables having a significant t statistics are in bold. Evaluations carried out in Group I of data with n = 49 voivodeships.

Retained explanatory variables									RR	Method
1	2	**3**	**4**	5	**6**	7	8	9	0.7742	full regression
1	**2**	**3**	**4**	5	**6**				0.7657	stepwise search from 1...9
1	**3**	**6**	**7**						0.7336	4 optimal from 1...9
1	**2**	**3**	**4**	**6**					0.7657	5 optimal from 1...9
1	**3**	**6**	**7**						0.7336	4 optimal from 1,3,5...,9
1	**3**	5	**6**	**7**					0.7425	5 optimal from 1,3,5...9
1	**3**	**6**							0.6439	suggested by collinearities identified in the matrix shown in Table 6
1	**2**	**3**	**6**						0.7336	
1	**3**	**4**	**6**						0.7325	
1	2	**3**	**9**						0.6697	
1	**2**	**3**	**4**	**9**					0.7405	
1	**3**	**6**	**7**						0.7336	

4. Investigation of interrelations between variables by inspecting the loadings of rotated principal axes

Hawkins (1973) has proposed a method of finding alternative subsets in regression by considering the loadings of principal axes derived by principal component analysis carried out when considering the augmented cross-product matrix $\mathbf{Z}^T\mathbf{Z}$, with $\mathbf{Z} = [\mathbf{y}, \overset{\cup}{\mathbf{X}}]$.

It is the usual practice in principal component analysis to center the columns of the matrix $\mathbf{Z}$ to zero mean. Very often the matrix $\mathbf{Z}$ is normalized (by dividing its values by appropriate constants) in such a way, that the computed matrix $\mathbf{Z}^T\mathbf{Z}$ is in fact the correlation matrix between the considered variables.

Let $\lambda_1 \geq \ldots \geq \lambda_m$ and $\mathbf{a}_1,\ldots,\mathbf{a}_m$ denote the eigenvalues and the eigenvectors of the matrix $\mathbf{Z}^T\mathbf{Z}$. Let us further assume that all the eigenvalues are positive.

Consider a hyperplane in the m-space ($m = p + 1$) given by the equation:

$$Y - \sum_{i=1}^{p} \beta_i X_i = 0 \tag{6}$$

For a given point $P = (y, x_1,\ldots,x_p)$ we can measure its distance from the hyperplane (6) in two ways:

– by measuring its distance along the y-axis as

$$y - \sum_{i=1}^{p} \beta_i x_i \tag{7}$$

(this is the way of proceeding when computing the ordinary LSE regression),

– by measuring its distance along the normal to the hyperplane as

$$\frac{y - \sum_{i=1}^{p} \beta_i x_i}{\sqrt{1 + \beta_1^2 + \ldots + \beta_p^2}} \tag{8}$$

(this is the way of proceeding when carrying out the principal component analysis).

The first distance is called by Hawkins the *y-norm distance*, and the second – the *vertical distance*.

Let s^2 denote the squared distance along the y-axis (i.e. the squared y-norm distance), and λ – the squared distance along the normal to the hyperplane (6). Obviously,

$$s^2 = l^2\lambda, \quad \text{with} \quad l^2 = 1 + \beta_1^2 + \ldots + \beta_p^2 \ . \tag{9}$$

Looking at the coefficient vector $(1, -\beta_1, \ldots, -\beta_p)$ appearing in (6) one can state that l^2 in the equation (9) above denotes just the squared length of the coefficient vector.

Let us return to equation (6). We can also write down this equation as

$$a_0 y + a_1 x_1 + \ldots + a_p x_p = 0 \ , \tag{10}$$

with directional cosines

$$a_0 = 1/l, \quad a_i = -\beta_i / l, \quad i = 1, \ldots, p \ .$$

Substituting $l^2 = 1/a_0^2$ into (9) and solving for λ we obtain that

$$\lambda = a_0^2 s^2 < s^2, \quad \text{or alternatively} \quad s^2 = \frac{1}{a_0^2}\lambda \ . \tag{11}$$

Hawkins (1973) has made the following remark: it is true that s^2 and λ both measure the fit of the hyperplane (6). Low s^2 implies low λ, since $0 < a_0^2 < 1$. However, given λ, the y-norm distance s^2 may be either arbitrarily large if a_0 is arbitrarily small, or very near to λ, if a_0^2 is relatively near to 1.

Let us note that a low λ with a low a_0^2 corresponds to a near multicollinearity amongst the predictors. Thus a linear combination of the x_i's which has a low λ identifies a low s^2 predictor of Y if a_0^2 is large, or a near multicollinearity amongst the predictors if a_0^2 is small.

It is possible to find in the considered multivariate m-space some bases of hyperplanes such that all possible hyperplanes in this space could be derived by forming linear combinations of the basic hyperplanes. One such basis is provided by the method of principal components, where each eigenvector $\mathbf{a}$ (computed from the correlation matrix or the cross-product matrix of the points located in the considered space) provides one hyperplane with the vertical norm λ equal to the eigenvalue λ associated with the vector $\mathbf{a}$.

Hawkins proposed to rescale the eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ by square roots of their eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ – to obtain the vectors $\mathbf{d}_j = \mathbf{a}_j / \sqrt{\lambda_j}$ $(j = 1, \ldots, m)$, which in turn, when put together, define the matrix $\mathbf{D}$:

$$\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_m) \ . \tag{12}$$

The components or coefficients of the vectors constituting the matrix $\mathbf{D}$ will be called in the following *the loadings*.

The matrix $\mathbf{D}$ can be rotated (e.g. using the *varimax* principle proposed by Kaiser, 1958; see also Morrison, 1967, or Krzanowski, 1988) to obtain a simpler structure in the loadings. Each column of $\mathbf{D}$ defines a basic hyperplane in the m-space.

Suppose that the variable Y was defined as the first column of the matrix $\mathbf{Z}$. Let $s_{(k)}^2$ denote the sum of the squared y-norm residuals (called also the *residual sum of squares*) obtained from the regression established by the k-th hyperplane. It is proved that $s_{(k)}^2$ can be obtained from the value d_{1k}^2 as its inverse:

$$s_{(k)}^2 = \frac{1}{d_{1k}^2} \ .$$

The procedure proposed by Hawkins proceeds according to the following steps:

1. Calculate the eigenvalues and the eigenvectors of $\mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z} = [\mathbf{y}, \mathbf{X}]$ is a matrix whose columns are centered to zero, i.e. $\mathbf{Z}^T\mathbf{1}_n = \mathbf{0}_{m \times 1}$.
2. Rescale the eigenvectors obtained in step 1 to the matrix $\mathbf{D}$ defined by formula (12) and rotate that matrix using the criterion "varimax" – to obtain a simpler structure of the loadings.
3. Look at the coefficients d_{1j}^2, $j = 1, \ldots, m$ of the rotated matrix $\mathbf{D}$, i.e. at the row of $\mathbf{D}$ corresponding to the predicted variable Y. Find those elements that are relatively large (> 1). The k-th column identified with the largest d_{1k}^2 indicates a parsimonious regression equation – we construct then the regression equa-

tion taking as predictors those variables that have "large" loadings in the k-th rotated column of **D**.

4. To find collinearities amongst the predictors look at the last columns associated with rather small values of λ_i and also with small loadings d^2_{1j}. Variables with large loadings in such columns are nearly collinear and can be mutually exchanged.

We have carried out the steps indicated above using the program HAWK from the package SFAX. The calculations were carried out using the correlation matrix of the variables $Y, X_1, \ldots, X_9$. The results of the calculations, i.e. the rotated matrix **D** and the eigenvalues are shown in Table 6.

For completeness of the presented results we show in the last (11-th) column of Table 6 the sums of squares (SS) calculated from squared loadings of each row – although we will not discuss further the meaning of the SS statistics.

We find in Table 6 that the relevant plane yielding a relatively small y-norm fit is constituted by the 8-th column. It yields the residual sum of squares

$$s^2_{(8)} = \frac{1}{1.75^2} = 0.33 \ .$$

This regression is mainly loaded by the variables X_1, X_2, X_3, X_4, X_6 and X_7.

Table 6

Rotated matrix of loadings together with λ_i's , the eigenvalues of the correlation matrix of the considered variables. Last column (SS) contains sum of squares of all loadings in given row

No.		1	2	3	4	5	6	7	8	9	10	11
Var	$\lambda_i \rightarrow$	4.746	2.505	1.048	0.570	0.409	0.310	0.202	0.130	0.067	0.012	SS
Y	ymog	–0.00	–0.01	–0.81	–0.05	–0.56	–0.30	–0.25	**1.75**	0.30	0.39	4.43
1	artf	0.00	0.01	0.05	–0.01	1.45	–0.17	–0.15	**–0.25**	0.04	–0.19	2.25
2	u18y	–0.45	–0.43	0.37	–0.04	0.24	0.63	–0.38	**–0.48**	–0.12	–6.05	37.92
3	divr	0.00	0.03	1.85	–0.27	0.05	–0.09	–0.05	**–0.46**	0.11	–0.17	3.77
4	nati	–0.58	–0.36	–0.35	–0.53	–0.85	–0.13	1.44	**0.72**	1.20	6.18	43.80
5	emin	0.00	0.00	–0.05	0.14	–0.16	1.49	–0.08	–0.13	0.11	–0.15	2.31
6	nmed	–0.02	–0.46	–0.29	0.10	–0.14	0.14	0.25	**0.34**	2.86	0.05	8.73
7	mmar	0.00	0.00	–0.07	–0.78	–0.42	–0.25	1.82	**–0.37**	0.35	1.11	5.64
8	wmar	0.00	–0.02	–0.22	1.60	–0.02	0.15	–0.26	–0.03	–0.13	–0.08	2.72
9	sece	–0.01	–0.90	–0.78	0.70	–0.33	–0.17	–0.05	0.04	–2.44	–0.76	8.57

Furthermore, we find in Table 6 three collinearities provided by the columns no. 7, no. 9 and no. 10.

Column 7 provides a collinearity between X_4 (="nati") and X_7 (="mmar"). Thus, each of these variables can be replaced by the other.

Column 9 provides a collinearity amongst the variables X_4 (="nati"), X_6 (="nmed") and X_9 (="sece"). Thus each of these variables can be replaced by the remaining two.

Column 10 provides a collinearity between the variables X_2 (="u18y"), X_4 (="nati"), X_7 (="mmar") and eventually X_9 (="sece"). Thus each of these variables can be replaced by the remaining ones.

Comparing these results with the former ones, obtained by the direct regressional methods, we find them concordant. The regression identified by the 8-th eigenvector shown in Table 6 is much similar – also in signs – to those obtained by the LSE method and shown in Table 4 (note that to obtain similar results as in Table 4, the signs in Table 6 corresponding to the predictors 1,...,9 should be reversed). The one exception is that the LSE method does not indicate that the variable X_7 (="mmar") is significant. The reason for this becomes clear after looking at the collinearity exhibited by the 7-th column of the rotated matrix **D**: since X_7 is nearly collinear with X_4, and X_4 was accounted for by the LSE regression, thus there was no need to indicate for X_7 as significant in the established regression. However, when the variable X_4 was dropped from the LSE regression, then the variable X_7 was indicated as important, i.e. as significant.

The 9-th column of the matrix **D** exhibits a collinearity between the variables X_4, X_6 and X_9. Thus it can be expected that, after dropping the variables X_4 and X_6 from the equation established by column 8 of the rotated matrix **D**, the variable X_9 will appear as important in the considered regression.

Following these findings we have evaluated the goodness of fit of the LSE regression for some alternative subsets which were derived by exchanging variables as indicated by the collinearities found in Table 6. The goodness of fit (measured by RR, the squared multiple correlation coefficient) of the alternative regressions is shown in Table 5. One can see that the indications found from Table 6 yield in practice good alternative regressions.

5. Discussion and final remarks

We have demonstrated that establishing a dependence model between one specified predicted variable and several specified predictor variables is not so straightforward as could be judged from some introductory textbooks. In any

case we should not stop after carrying out a regression analysis, but follow Tukey's principle: *Look at the data and think what you are doing*!

As one of the referees has pointed out (thanks to him for his comments!) we should be aware that there is an important difference between analysing data to obtain indications and explanations about the importance of possible interrelationships, and analysing data to obtain good predictors.

The goal of our analysis was double. We wanted to obtain a subset of good predictors; however at the same time we wanted to obtain indications and explanations of the importance of the considered predictors, and of their mutual relationship.

Analysing the mortality data we have demonstrated that an established regression equation can have alternative subsets yielding very similar predictions, thus it is a matter of our choice – may be based on some additional reasoning – which subset to introduce into the proposed regression model.

We found that the method proposed by Hawkins (1973) is very convenient for identifying alternative subsets of predictors. Scatterplot matrices, biplots, and rotated loadings (derived from rescaled eigenvectors) can provide useful hints on the direction of mutual relationships amongst the considered variables.

We have reported in the paper our investigations on mortality of men. Similar analysis was carried out on mortality of women. The revealed interdependencies look very similar, although the variable X_3 (="divr") appears a little less, and the variable X_6 (="nmed") a little more important.

Acknowledgment

The work was partially sponsored by the KBN grant no. 0663/S4/93/04.

REFERENCES

Antoch J., Bartkowiak A. (1988). L-estimators in linear model. *Biometrical Letters* **25,** 3-24.

Bartkowiak A. (1992). An empirical study of robust regression in the presence of outliers blunders. Part 2. Iteratively reweighted regression. *Biocybernetics and Biomedical Engineering* **12**, 27-45.

Bartkowiak A. (1995). SFAX, a package for a specialistic fast exploratory analysis of data. *Advances in Modelling and Analysis* B, **36**, 57-64.

Bartkowiak A., Szustalewicz A. (1995). Konstrukcja biplotu i przykład jego zastosowania. In: A.Bartkowiak (ed.), *Lisp-Stat., Narzędzie eksploratywnej analizy danych,* Uniwersytet Wrocławski, Wrocław, 106-123.

Baskerville J.C., Toogood J.H. (1982). Guided regression modeling for prediction and exploration of structure with many explanatory variables. *Technometrics* **24**, 9-17.

Chatterjee S., Hadi A.S. (1988). *Sensivity Analysis In Linear Regression*. Wiley, New York.

Gabriel K.R. (1982). Biplot. In: S. Kotz and N.L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* V.1, Wiley, New York, 263-271.

Gabriel K.R., Odoroff Ch.L. (1990). Biplots in biomedical research. *Statistics in Medicine* **9**, 469-485.

Grizzle J.E., Sen P.K. (1983). Selection of informative variables in multivariate analysis: an analysis of covariance approach. In: P.K. Sen (ed.), *Essays in Honor of Norman L. Johnson*, North Holland, 195-204.

Hawkins D.M. (1973). On the investigation of alternative regressions by principal component analysis. *Applied Statistics* **22**, 275-286.

Jeffers J.N.R. (1981). Investigation of alternative regressions: Some practical examples. *The Statistician* **30**, 79-88.

Jolliffe I.T. (1986). *Principal Component Analysis*. Springer, New York.

Kaiser H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187-200.

Kane V.E., Ward R.C., Davis G.J. (1985). Assessment of linear dependencies in multivariate data. *SIAM J. Sci. Stat. Comput.* **6**, 1022-1032.

Krzanowski W.J. (1988). *Principles of Multivariate Analysis. A user's perspective*. Clarendon Press, Oxford (U.K.).

Kupść W., Jasiński B. (1991). *Sytuacja Epidemiologiczna w Zakresie Chorób Układu Krążenia w Polsce*. Część I. Wydawnictwo nr 16 w serii: Biblioteka Kardiologiczna. Instytut Kardiologii Warszawa.

Li G. (1985). Robust regression. In: Hoaglin D.C., Mosteller F., Tukey J.W. (Eds), *Exploring Data Tables, Trends and Shapes*. Wiley, New York, 281-343.

McKay R.J. (1979). The adequacy of variable subsets in multivariate regression. *Technometrics* **21**, 475-479.

Morrison D.F. (1967). *Multivariate Statistical Methods*. McGraw-Hill, New York.

Tierney L. (1990). *Lisp-Stat, an Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley New York.

Received 10 February 1995; revised 21 June 1995

Analiza regresji wyznaczającej umieralność mężczyzn w Polsce w latach 1989-91

Streszczenie

Badany jest związek między ogólną umieralnością mężczyzn w Polsce obserwowaną w 49 województwach i zespołem 9 zmiennych opisujących środowisko i socjo-ekonomiczny status każdego z województw. Standardowa analiza regresji wielokrotnej (metoda najmniejszych kwadratow) wykazała wyraźną zależność między umieralnością i zespołem badanych zmiennych (współczynnik korelacji wielokrotnej równy 0.75).

Jednakże wnioskowanie oparte na pełnym modelu regresji i testowanie hipotez o poszczególnych współczynnikach równania może nie doprowadzić do jednoznacznej oceny wpływu poszczególnych zmiennych objaśniających na umieralność. Fakt ten spowodowany może być przez

1) istnienie obserwacji wyraźnie "odstających", które mogą wpłynąć zasadniczo na postać otrzymanego równania regresji;

2) istnienie mniejszego (lub kilku mniejszych) podzbiorów zmiennych objaśniających "równie dobrze" opisującego badaną zależność;

3) występowanie silnych korelacji miedzy rozpatrywanymi zmiennymi, powodujących ich współliniowość i naruszających stabilność otrzymanych oszacowań współczynników regresji .

Dla wyjaśnienia sformułowanych zagadnień zastosowano techniki statystyczne eksploratywnej analizy danych, wyznaczono współczynniki regresji przy użyciu metod odpornych na zakłócenia – metodę alfa-obcięcia i metodę wag Hubera oraz metodę Hawkinsa (1973) badania współzależności statystycznych, pozwalającą na znajdowanie podzbioru zmiennych, które mogą zastępować się wzajemnie.

Słowa kluczowe: regresja metodą najmniejszych kwadratów, podzbiory predyktorów, alternatywne podzbiory, eksploratywna analiza danych, umieralność mężczyzn w Polsce.