Quaderni di Statistica
Vol. 11, 2009

Evaluating public services through multivariate linear regression analysis

Mirko Di Martino
Servizio Controllo Strategico e Statistica, Regione Emilia-Romagna
E-mail: MiDiMartino@Regione.Emilia-Romagna.it

Giuliano Galimberti Gabriele Soffritti
Dipartimento di Scienze Statistiche, Università di Bologna
E-mail: giuliano.galimberti@unibo.it, gabriele.soffritti@unibo.it

Summary: Data obtained from the most recent Italian multipurpose survey *Health conditions and the use of health services* is analysed through multivariate linear regression methods in order to study the influence of some geographic, social, economic, demographic and health factors on the Italian public opinion about the Italian health system and other public services (postal services, rail transports, education system, public television, phone services, public utilities). Recently-proposed multivariate linear regression methods that take account of the possible presence of skewness and/or heavy-tails in the distribution of the error terms are employed. A best subset selection of the relevant regressors for the multivariate linear regression model is performed. The analysis allows to detect the factors that mainly affect public opinion and the way such factors influence satisfaction with the investigated public services.

Keywords: Finite mixture model, Italian multipurpose survey, Multivariate linear regression.

1. Introduction

The Italian multipurpose survey on families *Health conditions and the use of health services* is periodically performed and processed by the National Institute of Statistics (ISTAT). The 2004-2005 survey collected information about respondents' opinions on the Italian health system and

some other public services (ISTAT, 2006).

The aim of this paper is to evaluate the dependence of the Italian public opinion about such services on some geographic, social, economic, demographic and health factors. Namely, the factors considered in this paper are: educational level, income, gender, age, physical and psychological perceived health status, health service consumption and presence of chronic diseases. The analysis is performed at a local health unit[1] level, by using aggregated variables.

This aim is pursued through multivariate linear regression methods. Such methods are widely used in many branches of science to predict values of D responses from a set of P regressors, where $D \geq 1$ and $P \geq 1$. They are based on a statistical model in which the error terms are generally assumed to be independent and identically distributed random variables. As far as the error distribution is concerned, it is usually considered to be multivariate normal with a zero mean vector and a positive definite covariance matrix (see, for example, Mardia *et al.*, 1979, Srivastava, 2002). However, if the error distribution is skewed and/or heavy-tailed, the assumption of multivariate normality will ignore such important features of the data and the results will be not completely adequate. Many solutions to this problem have been discussed and proposed in literature (see, for example, Zellner, 1976; Sutradhar and Ali, 1986; Galea *et al.*, 1997; Liu, 2002; Diaz-Garcia *et al.*, 2003; Ferreira and Steel, 2004). However, they rely on parametric models that may still be incorrectly specified.

A proposal that allows to overcome this drawback is based on a mixture model framework: the unknown distribution of the error terms is modelled using a finite mixture of Gaussian D-dimensional components (Soffritti and Galimberti, 2009). Finite mixture modelling represents a convenient framework in which to model unknown distributional shapes. It is well known that, through an appropriate choice of its components, a finite mixture model is able to model quite complex distributions, including skewed and/or heavy-tailed distributions, and can handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data (McLachlan and Peel, 2000). In

[1] ASL: *Azienda Sanitaria Locale*.

the special case of a single Gaussian component, this approach coincides with the classical one. In order to deal with the possible presence of skewness and/or heavy tails in the distribution of the error terms, we have performed the analysis of the data using this approach.

The remainder of the paper is organized as follows: in Section 2 we give some information about the survey *Health conditions and the use of health services* and the data analysed in this paper; in Section 3 we describe the multivariate linear regression models employed in this study to evaluate opinions about public services; in Section 4 we illustrate the main results of the data analysis; finally, in Section 5, we present a short discussion and some concluding remarks.

2. The 2004-2005 Italian multipurpose survey and the analysed dataset

The multipurpose survey on *Health conditions and the use of health services* is periodically conducted by ISTAT on a sample of Italian families. In the most recent survey (2004-2005) an entire section was dedicated to opinions on the Italian health system and some other public services. ISTAT survey data collection is based on local administrative offices spread all over Italian territory at regional and municipal level. From ISTAT central office to the local network of interviewers two main steps are performed: at first, ISTAT central office contacts the ISTAT regional offices[2] to provide information on the survey (questionnaires, instructions, time schedule); then, every regional office contacts selected municipalities (first stage sampling units) to coordinate second stage sampling selection (households), interviewers' recruiting and field-work performance.

In order to provide accurate estimates for sub-regional levels, the sample size of the 2004-2005 survey was increased up to more than 50,000 families, approximately 128,000 subjects. The families were selected in a representative way for the whole Italian population. The survey was con-

[2] Each Italian region has its own ISTAT regional office apart from Trentino-Alto Adige, that has two ISTAT offices, one for each autonomous province. The ISTAT offices of Piemonte and Valle D'Aosta are formally considered as a unique regional office.

ducted by both face-to-face interviews and self-administered questionnaires with all family members in a sampled family. The face-to-face questionnaire is composed of two sections: one concerns each family member while the other refers to the household. The individual section includes demographic, social and economic factors, working conditions, weight, height and diet, prevention and physical activity, physical examinations or checkups, hospitalizations, health-rehabilitation services, disabilities, expenditures on health and social services, alternative medicine. The household section collects information about aids and services for the family, the house in which the family lives and resources of the family. The self-administered questionnaire focuses on perceived state of health, chronic or long-lasting illness, drug consumption, tobacco consumption, opinions on some public services, pregnancy and breastfeeding. If a family member was absent a proxy was interviewed. Nearest-neighbour imputation methods were applied for missing data. Items related to opinions on public services were restricted to respondents older than 17 years of age.

The World Health Organization defines health as being "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity". In order to gather quantitative information about this multi-dimensional concept, the 2004-2005 Italian multipurpose survey included the 12-item Short-Form Health Survey questionnaire (SF-12), which is one of the most widely-used tools to measure the health-related quality of life. The SF-12 was originally developed in the United States to provide a shorter alternative to a similar 36-item questionnaire (the SF-36, see Ware *et al.*, 1992), for use in large-scale health measurements whose focus is on overall physical and mental health outcomes. The SF-12 was subsequently validated in European countries (see, for example, Gandek *et al.*, 1998). It generates two summary scores: the physical component summary (PCS) and the mental component summary (MCS). These scores are computed by weighting data about physical and social habits, limitations due to physical, emotional and mental problems, bodily pain, vitality and perceptions of general health.

Two indexes were derived from the recorded data: the chronicity index and the health services consumption index. The chronicity index is

based on the number of self-reported chronic conditions and their impact on perceived health status. Scores range from 0 (absence of long-lasting illness) to 100. The health services consumption index is a weighted sum of the health resources recently used by respondents (hospital admissions, medical examinations, checkups, rehabilitation services, drug consumption). Weights are proportional to the economic value of these services. Scores range from 0 (no service consumption) to 100. For further details see, for example, Gargiulo *et al.* (2008).

Analyses were performed at a local health unit level (189 units), by using the following aggregated variables as measures of the opinions on public services:

- *Trend*, percentage of respondents who think that the public health system is getting worse;
- *Health_Grade*, percentage of respondents giving a rating lower than or equal to 5 on the local health system (on a 1-10 rating scale, 1 worst judgement, 10 best judgement);
- *Other_Grade*, percentage of respondents giving an average rating lower than or equal to 5 (on a 1-10 rating scale, 1 worst judgement, 10 best judgement) on other public services (postal service, rail transports, education system, public television, phone service, public utilities).

These three variables were analysed as response variables, while the candidate explaining factors were:

- *Male*, percentage of male respondents;
- *Elder*, percentage of respondents older than 64 years of age;
- *PCS*, average physical component summary;
- *MCS*, average mental component summary;
- *Chron*, average chronicity index;
- *Cons*, average health services consumption index;

- *Educ*, percentage of respondents with low educational level;
- *Econ*, percentage of respondents belonging to households with low (scarce or absolutely insufficient) economic resources;
- *Area*, geographic area (North-West, North-East, Centre, South, Islands).

The effects of these explaining factors on the three response variables were studied within the framework of multivariate linear regression. This framework allows to take into account the relationships among the response variables, thus leading to a single, simultaneous model for jointly evaluating the investigated Italian public services.

3. Multivariate linear regression methods

Multivariate linear regression models (see, for example, Mardia *et al.*, 1979; Srivastava, 2002) are generally based on the assumption that the joint dependence of D response variables on P regressors is linear. Furthermore, it is usually assumed that the error terms are independent and identically distributed random vectors, whose distribution is assumed to be multivariate Gaussian with a D-dimensional zero mean vector and a positive definite covariance matrix $\mathbf{\Sigma}$. Namely:

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \;\; \boldsymbol{\epsilon}_i \sim MVN_D(\mathbf{0}, \mathbf{\Sigma}), \tag{1}$$

where $\mathbf{Y}_i$ and $\boldsymbol{x}_i$ represent the D-dimensional random column vector of the response variables and the P-dimensional column vector of the fixed regressor values for the ith sample unit, respectively; $\boldsymbol{\beta}_0$ is a D-dimensional column vector containing the intercepts for the D responses; $\boldsymbol{B}$ is a matrix of dimension $P \times D$ whose (p, d)th element, β_{pd}, is the regression coefficient of the pth regressor on the dth response; finally, $\boldsymbol{\epsilon}_i$ denotes the D-dimensional column random vector of the error terms corresponding to the ith observation.

Model (1) has been recently generalised by assuming that the distribution of the error terms is a mixture of K D-dimensional Gaussian

components (Soffritti and Galimberti, 2009):

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \;\; \boldsymbol{\epsilon}_i \sim \sum_{k=1}^{K} \pi_k MVN_D(\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k), \tag{2}$$

where $\pi_k > 0 \; \forall k$, $\sum_{k=1}^{K} \pi_k = 1$, and $\sum_{k=1}^{K} \pi_k \boldsymbol{\nu}_k = \mathbf{0}$.

Clearly, when $K = 1$ model (2) coincides with model (1). When $K > 1$, model (2) is able to model quite complex distributions and can handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. Model (2) also represents a generalization of a model proposed for dealing with non-normal error terms in the multiple linear regression analysis, that is when $D = 1$ (Bartolucci and Scaccia, 2005).

Given equation (2), the probability density function of the ith observation of the D response variables, $\boldsymbol{y}_i$, is

$$\sum_{k=1}^{K} \pi_k \phi_D(\boldsymbol{y}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k), \;\; \boldsymbol{\mu}_{ik} = \boldsymbol{\nu}_k + \boldsymbol{\beta}_0 + \boldsymbol{B}'\boldsymbol{x}_i, \tag{3}$$

where $\phi_D(\boldsymbol{y}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k)$ is the density of the D-dimensional Gaussian distribution $MVN_D(\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k)$ evaluated at $\boldsymbol{y}_i$. According to equation (3), model (2) can also be seen as a mixture of K restricted multivariate linear regression models with Gaussian error terms, whose generic component takes the form

$$\mathbf{Y}_i = \boldsymbol{\lambda}_k + \boldsymbol{B}'\boldsymbol{x}_i + \tilde{\boldsymbol{\epsilon}}_{ik}, \;\; \tilde{\boldsymbol{\epsilon}}_{ik} \sim MVN_D(\mathbf{0}, \boldsymbol{\Sigma}_k), \tag{4}$$

where $\boldsymbol{\lambda}_k = \boldsymbol{\beta}_0 + \boldsymbol{\nu}_k$, for $k = 1, \ldots, K$. Thus, the components have different intercepts for the D responses and different covariance matrices for the error terms, but the K matrices of the regression coefficients are restricted to be equal. This alternative representation of model (2) allows to highlight a further property: in model (2) the set of regressors also includes one latent categorical variable with K categories. It affects both the conditional expected value and the conditional covariance matrix of the dependent variables, but it does not interact with the other regressors.

This categorical latent variable is also assumed to be independent of all regressors.

Provided that the $I \times P$ matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_i$ for $i = 1, \ldots, I$ has full column rank, it is possible to prove that, apart from the well-known label-switching problem (see, for example, McLachlan and Peel, 2000), model (2) is always identifiable. This proof is similar to the proof of Proposition 2 in Yakowitz and Spragins (1968) and exploits the fact that matrix $\boldsymbol{B}$ does not depend on k (for further details see Soffritti and Galimberti, 2009).

Maximum likelihood estimation of the parameters of model (2) may be carried out through the well-known Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). This is a general-purpose algorithm for maximum likelihood estimation in a wide variety of situations best described as incomplete-data problems. A comprehensive account of the EM algorithm, including the special case of parameter estimation for mixture models, can be found in McLachlan and Krishnan (2008). As far as model (2) is concerned, the model log-likelihood given a random sample of I observations is

$$l = \sum_{i=1}^{I} \log \left(\sum_{k=1}^{K} \pi_k \phi_D(\boldsymbol{y}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k) \right). \tag{5}$$

Let $\boldsymbol{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_I')$ be the matrix of dimension $I \times D$ with the values of the D response variables for the I sample units. Let z_{ik} be a binary variable equal to 1 when the error term for the ith observation has been generated from the kth component, and 0 otherwise, for $k = 1, \ldots, K$. Thus, $\sum_{k=1}^{K} z_{ik} = 1$. Furthermore, let $\boldsymbol{z}_i$ be the K-dimensional column vector whose kth element is z_{ik}. Since vectors $\boldsymbol{z}_i$'s are unknown, the observed data $\boldsymbol{Y}$ can be considered incomplete, and equation (5) represents the incomplete-data log-likelihood. If we know both the observed data and the component-label vectors $\boldsymbol{z}_i$'s, we can obtain the so-called complete log-likelihood of the model. For random samples, it is appropriate to assume that the component label vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_I$ are observed values of I independent and identically distributed random vectors whose unconditional distribution is multinomial consisting of one draw on K categories with probabilities $\pi_1, \ldots, \pi_K$. Up to a constant factor, the complete

log-likelihood of the model is equal to

$$l_c = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}[\log \pi_k + \log \phi_D(\boldsymbol{y}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k)] = l_{c1} + l_{c2}, \qquad (6)$$

where

$$\begin{aligned} l_{c1} &= \sum_{k=1}^{K} z_{\cdot k} \log \pi_k, \\ l_{c2} &= -\frac{1}{2}\sum_{k=1}^{K} z_{\cdot k} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}(\boldsymbol{y}_i - \boldsymbol{\mu}_{ik})'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{ik}), \end{aligned}$$

with $z_{\cdot k} = \sum_{i=1}^{I} z_{ik}$, and $|\boldsymbol{A}|$ denotes the determinant of matrix $\boldsymbol{A}$.

Function l_{c1} depends only on the parameters π_k's and can be maximized simply by letting π_k equal to $\hat{\pi}_k = z_{\cdot k}/I$, $k = 1, \ldots, K$.

In order to show how l_{c2} can be maximized, it is convenient to express such a quantity in the matrix notation obtained as follows. Let $\boldsymbol{\Gamma} = (\boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_K', \boldsymbol{B})$ whose dimensions are $(K + P) \times D$. Moreover, let $\boldsymbol{z}_k = (z_{1k}, \ldots, z_{Ik})'$ and $\boldsymbol{\mu}_k = (\boldsymbol{\mu}_{1k}', \ldots, \boldsymbol{\mu}_{Ik}')$, whose dimensions are $I \times 1$ and $I \times D$, respectively. Note that the latter may be expressed as $\boldsymbol{\mu}_k = \boldsymbol{X}_k\boldsymbol{\Gamma}$, where $\boldsymbol{X}_k = (\boldsymbol{O}_k\ \boldsymbol{X})$, $\boldsymbol{O}_k$ is a matrix of dimension $I \times K$ with all the elements equal to 0 apart from those of column k which are equal to 1. As a consequence of these relations, it is possible to write

$$l_{c2} = -\frac{1}{2}\sum_{k=1}^{K} z_{\cdot k} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2}\sum_{k=1}^{K} tr(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{D}_k), \qquad (7)$$

where $\boldsymbol{D}_k = (\boldsymbol{Y} - \boldsymbol{X}_k\boldsymbol{\Gamma})'diag(\boldsymbol{z}_k)(\boldsymbol{Y} - \boldsymbol{X}_k\boldsymbol{\Gamma})$, and $diag(\boldsymbol{z}_k)$ is the $I \times I$ diagonal matrix whose main diagonal equals vector $\boldsymbol{z}_k$.

Function l_{c2} defined by equation (7) depends on the parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$. It can be maximized by evaluating its first differential, by setting the first derivatives computed with respect to all the parameters equal to 0, and by solving the resulting equations (for further details see Soffritti and Galimberti, 2009). Provided that matrix $\boldsymbol{M} = \sum_{k=1}^{K} \boldsymbol{\Sigma}_k^{-1} \otimes [\boldsymbol{X}_k'diag(\boldsymbol{z}_k)\boldsymbol{X}_k]$ is non-singular, the solutions are

$$vec(\hat{\boldsymbol{\Gamma}}) = \boldsymbol{M}^{-1}\boldsymbol{N}vec(\boldsymbol{Y}), \tag{8}$$

$$\hat{\boldsymbol{\Sigma}}_k = z_{.k}^{-1}\boldsymbol{D}_k, \;\; k = 1, \ldots, K, \tag{9}$$

where $\boldsymbol{N} = \sum_{k=1}^{K} \boldsymbol{\Sigma}_k^{-1} \otimes [\boldsymbol{X}'_k diag(\boldsymbol{z}_k)]$, $vec(\boldsymbol{A})$ denotes the vector formed by stacking columns of the matrix $\boldsymbol{A}$, one underneath the other, and $\otimes$ is the Kronecker product operator. From $\hat{\boldsymbol{\Gamma}}$ we directly obtain $\hat{\boldsymbol{\lambda}}_1, \ldots, \hat{\boldsymbol{\lambda}}_K$ and $\hat{\boldsymbol{B}}$. We may also obtain $\hat{\boldsymbol{\mu}}_{ik}$ as $\hat{\boldsymbol{\lambda}}_k + \hat{\boldsymbol{B}}'\boldsymbol{x}_i$ for $k = 1, \ldots, K$ and $i = 1, \ldots, I$. Furthermore, $\hat{\boldsymbol{\beta}}_0$ is obtained as $\sum_{k=1}^{K} \hat{\pi}_k \hat{\boldsymbol{\lambda}}_k$, and $\hat{\boldsymbol{\nu}}_k$ as $\hat{\boldsymbol{\lambda}}_k - \hat{\boldsymbol{\beta}}_0$ for $k = 1, \ldots, K$.

As equation (8) depends on the $\boldsymbol{\Sigma}_k$'s, and equation (9) depends on $\boldsymbol{\Gamma}$, the maximization of function l_{c2} with respect to such parameters can be obtained by iteratively updating the estimate of $\boldsymbol{\Gamma}$ given an estimate of the $\boldsymbol{\Sigma}_k$'s, and vice versa. Since the z_{ik}'s are missing, in the EM algorithm they are substituted with their conditional expected values. More specifically, the EM algorithm consists in iterating the following two steps until convergence:

Step E On the basis of the current estimate of the model parameters, the expected value of the complete log-likelihood given the observed data, $E(l_c|\boldsymbol{Y})$, is computed. In practice, this consists of substituting any z_{ik} in equation (6) with its conditional expected value

$$p_{ik} = E(z_{ik}|\boldsymbol{Y}) = \frac{\hat{\pi}_k \phi_D(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_{ik}, \hat{\boldsymbol{\Sigma}}_k)}{\sum_{h=1}^{K} \hat{\pi}_h \phi_D(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_{ih}, \hat{\boldsymbol{\Sigma}}_h)}. \tag{10}$$

Step M $E(l_c|\boldsymbol{Y})$ is maximized with respect to the model parameters as follows:

1. the estimate for π_k is updated by computing $\frac{1}{I}\sum_{i=1}^{I} p_{ik}$ ($k = 1, \ldots, K$);
2. the estimates of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$, are iteratively updated, until convergence, through equations (8) and (9) respectively, where any z_{ik} is substituted with the corresponding p_{ik} defined in equation (10).

The iterative estimation process requires a set of starting values for the model parameters. A possible solution can be obtained as follows: for $\boldsymbol{B}$ simply compute the ML estimate under assumption (1). For $\boldsymbol{\lambda}_k$ use $\tilde{\boldsymbol{\beta}}_0 + \tilde{\boldsymbol{\nu}}_k$ ($k = 1, \ldots, K$), where $\tilde{\boldsymbol{\beta}}_0$ is the estimate of $\boldsymbol{\beta}_0$ under assumption (1), and $\tilde{\boldsymbol{\nu}}_k$ the estimate of $\boldsymbol{\nu}_k$ obtained by fitting a multivariate Gaussian mixture model to the residuals computed under assumption (1), using for example the `R` package `mclust02` (Fraley and Raftery, 2002; 2003). This also provides the starting estimates of the parameters $\boldsymbol{\Sigma}_k$ and π_k, $k = 1, \ldots, K$. As far as the choice of the unknown value of K is concerned, model-selection techniques that take into account both the fit and complexity of a model can be employed (see, for example, McLachlan and Peel, 2000).

4. Results

Given the data described in Section 2, the multivariate linear regression model defined by equation (2) was estimated for values of K from 1 to 3. No restriction was imposed on the error covariance matrices throughout the analysis. All calculations were performed in the `R` environment (R Development Core Team, 2008). A specific function implementing the maximum likelihood estimation of the model parameters through the EM algorithm described in Section 3 was used. In order to avoid difficulties when applying linear regression models to bounded-range dependent variables, logit transformation was applied to each of the $D = 3$ dependent variables before estimating the models, as suggested, for example, in Montgomery *et al.* (2006). One local health unit was excluded from the analysis due to an outlying value for one of the dependent variables, thus considering $I = 188$ units. Furthermore, the explaining factor *Area* was recoded into the following four dummy indicators, using the North-West area as a reference category: *Area_NE* (North-East), *Area_C* (Centre), *Area_S* (South), *Area_IS* (Islands), thus producing 12 candidate explaining factors.

Since a crucial point of a regression analysis is the choice of the relevant regressors (see, for example, Miller, 2002), we performed a best subset selection in which the whole set of the possible regressors was

composed of the factors described in Section 2. Only main effects were considered; thus, $2^{12} = 4096$ different regression equations were estimated and fitted to the dataset for each value of K, with a different subset of regressors each. Estimation was performed using a PC with Windows XP Professional operating system, an Intel Core 2 QUAD processor and 4 GB RAM. The convergence criteria used in the analysis were: the increment in the log-likelihood value between two consecutive steps lower than 0.0005 for the EM algorithm (with a maximum number of iterations equal to 300); the Euclidean distance between two consecutive model parameter estimates, divided by the total number of estimated parameters, lower than 0.0005 for the M step within the EM algorithm (with a maximum number of iterations equal to 100). Fitting the 12288 models took 108 hours and 37 minutes: 9 minutes for models with $K = 1$, 32 hours and 8 minutes with $K = 2$, and 76 hours and 20 minutes with $K = 3$.

The choice of the best model among the fitted ones was performed using the Bayesian Information Criterion (Fraley and Raftery, 2002; Fraley and Raftery, 2003):

$$BIC_M = 2 \max [\log L_M] - npar_M \log(I),$$

where $\max [\log L_M]$ is the maximum of the log-likelihood of a model M for the given sample of I units and $npar_M$ is the number of independent parameters to be estimated for that model. This criterion enables us to trade-off the fit and parsimony of a given model: the greater the BIC, the better the model (Schwartz, 1978).

Figure 1 summarizes the distributions of the BIC values taking into account the number of components K and the number of regressors P. If we compare models with the same number of regressors, it is possible to see that the median BIC values tend to decrease as K increases for $P = 1, \ldots, 11$. However, as far as the maximum BIC values are concerned, they correspond to models with one component only when $P \leq 5$. When $P > 5$, the maximum BIC values are obtained using models with $K > 1$ components.

The best models were selected and examined for $K = 1, 2, 3$. Their BIC values are highlighted in Figure 1. Table 1 reports the three subsets of regressors corresponding to these three best models, together with the

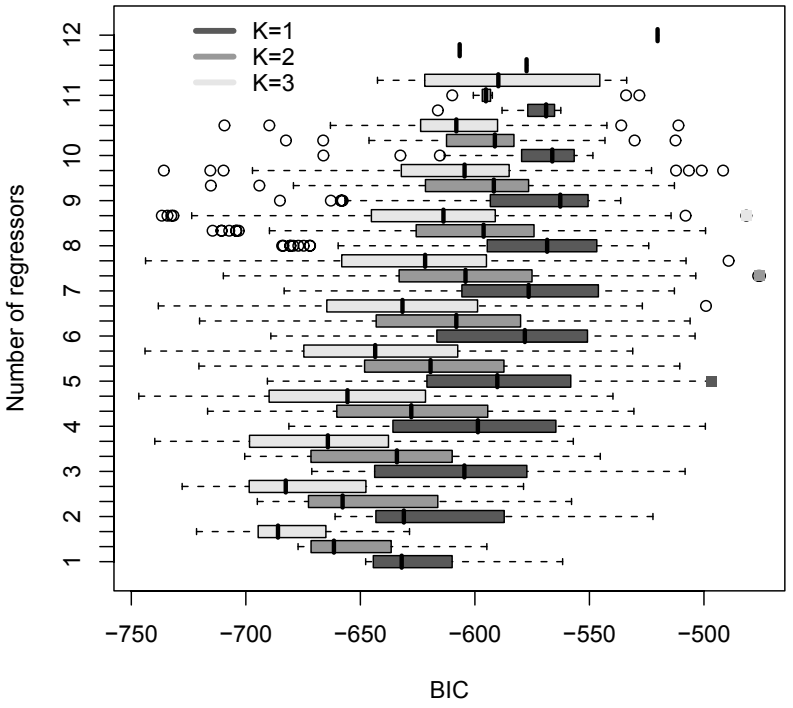

Figure 1. Boxplots of the BIC statistic distributions by number of regressors P and number of components K

values of the BIC statistic for the models with $K = 1, 2, 3$ components. A different subset of relevant regressors is obtained for each value of K. The regressors that result to be relevant in all the selected models are: $Econ$, $Area\_S$ and $Area\_NE$. According to the BIC the best model is the one with $K = 2$, that is, a model with a mixture of two Gaussian components for the error terms. In the following we focus on the results obtained from this particular model.

The estimates of the two mixing proportions are $\hat{\pi}_1 = 0.084$ and $\hat{\pi}_2 = 0.916$. Table 2 shows the estimates of the remaining model pa-

Table 1. Best subsets of regressors for each value of K, and corresponding BIC values for models with $K = 1, 2, 3$ (for each subset the largest BIC value is reported in bold type)

$K=1$	$K=2$	$K=3$	Regressors
−496.84	−531.17	−569.21	$Chron, Econ, Area\_NE,$ $Area\_S, Area\_IS$
−522.39	**−475.53**	−557.34	$Chron, Cons, Educ, Econ,$ $Area\_NE, Area\_S, Area\_IS$
−552.75	−535.79	**−481.50**	$Male, Elder, PCS, Cons,$ $Educ, Econ, Area\_NE, Area\_S$

Table 2. Estimates of the parameters $\boldsymbol{\nu}_k$ and $\boldsymbol{\Sigma}_k$ for the best model (estimated correlation coefficients between dependent variables in brackets)

Estimate	$logit(Trend)$	$logit(Health\_Grade)$	$logit(Other\_Grade)$
$\hat{\boldsymbol{\nu}}_1'$	−0.284	−0.375	−0.427
$\hat{\boldsymbol{\nu}}_2'$	0.026	0.034	0.039
$\hat{\boldsymbol{\Sigma}}_1$	0.070	0.024	−0.014
	(0.637)	0.020	−0.023
	(−0.305)	(−0.931)	0.032
$\hat{\boldsymbol{\Sigma}}_2$	0.116	0.078	0.028
	(0.579)	0.158	0.068
	(0.210)	(0.440)	0.149

rameters that depend on the components, that is, $\boldsymbol{\nu}_k$ and $\boldsymbol{\Sigma}_k$. An interesting result concerns some correlations between the dependent variables: within the first component $logit(Other\_Grade)$ is negatively correlated with both $logit(Trend)$ and $logit(Health\_Grade)$, while within the second component the same correlations reverse.

Table 3 shows the estimates of the model parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{B}$ together with their 95% confidence intervals obtained using the parametric bootstrapping residual method (Efron and Tibshirani, 1993). Such intervals do not contain the zero value for some of the model parameters (see the bold-faced entries in Table 3). Thus, not all the model parameters may be considered significant. As far as the effects of the selected explaining

Table 3. Estimated intercepts and regression coefficients of the best model (bootstrap 95% confidence intervals are reported in brackets)

	$logit(Trend)$	$logit(Health\_Grade)$	$logit(Other\_Grade)$
$\hat{\beta}_0'$	**−1.19**	**−1.01**	**−1.70**
	(−1.82, −0.50)	(−1.79, −0.26)	(−2.40, −0.91)
$Chron$	**0.17**	**0.16**	**0.19**
	(0.10, 0.24)	(0.06, 0.25)	(0.09, 0.26)
$Cons$	−0.08	−0.29	−0.32
	(−0.37, 0.23)	(−0.61, 0.09)	(−0.65, 0.04)
$Educ$	**−1.01**	**−1.08**	**−1.08**
	(−1.88, −0.18)	(−2.11, −0.03)	(−2.10, −0.08)
$Econ$	**0.79**	**1.62**	0.45
	(0.15, 1.49)	(0.77, 2.48)	(−0.44, 1.32)
$Area\_NE$	**0.25**	−0.09	0.06
	(0.10, 0.38)	(−0.26, 0.07)	(−0.10, 0.20)
$Area\_S$	0.11	**0.65**	**0.19**
	(−0.03, 0.26)	(0.46, 0.80)	(0.01, 0.34)
$Area\_IS$	0.04	**0.51**	**0.29**
	(−0.18, 0.24)	(0.26, 0.77)	(0.03, 0.49)

factors are concerned, dissatisfaction with health and other public services is positively related with the presence of chronic conditions, while it decreases when the percentage of people with low educational level increases. Low economic resources seem to be significantly related only with the dissatisfaction with health services. Moreover, South Italy and the Islands show markedly lower satisfaction with both local health system and other public services compared to the North-West area, while in those areas opinions about the trend in the public health system are similar. North-East Italy is characterized by more pessimistic views about this trend than the North-West, while the satisfaction levels with the public services are similar. Regarding health services consumption index, dissatisfaction with health system seems to be lower (the regressor coefficients are nearly statistically significant) in local health units where residents use more extensively health care services.

Table 4. Estimated intercepts and regression coefficients of the model that uses the same subset of regressors as the best one but with only one component

	$logit(Trend)$	$logit(Health\_Grade)$	$logit(Other\_Grade)$
$\hat{\boldsymbol{\beta}}_0'$	**−1.41**	**−1.53**	**−2.17**
$Chron$	**0.16**	**0.12**	**0.15**
$Cons$	−0.07	−0.24	−0.26
$Educ$	−0.57	−0.11	−0.19
$Econ$	**0.81**	**1.70**	0.51
$Area\_NE$	**0.22**	−0.15	0.01
$Area\_S$	0.09	**0.61**	0.14
$Area\_IS$	0.04	**0.51**	**0.31**

A deeper insight into the effects of the error term specification on the selection of the relevant regressors emerges from the comparison between the best selected model and the model that has the same subset of relevant regressors but with $K = 1$ component. Table 4 shows the estimated intercepts and regression coefficients of this second model (bold-faced entries denote the estimates whose 95% bootstrap interval does not contain the zero value). The main difference with respect to the results reported in Table 3 is related to the effect of the percentage of respondents with low educational level: assuming normal error terms for the linear regression model leads to regression coefficients which are closer to zero and also not significant.

5. Discussion and concluding remarks

The set of geographic, social, economic, demographic and health factors identified by the analysis described in this paper may be used to explain Italian public opinion about the Italian health system and other public services. The estimated effects are sensible and coherent with previous knowledge and conjectures on the phenomenon. The multivariate linear regression model selected using the BIC statistic is characterized by non-Gaussian error terms. However, it is worth noting that the selected model

is not completely satisfactory. If we compute the proportion of the total sum of squares of each response explained by the selected model (by using the posterior estimated mixing proportions), the estimated model accounts for only 21%, 55% and 22% of the deviance of $logit(Trend)$, $logit(Health\_Grade)$ and $logit(Other\_Grade)$, respectively. Furthermore, the use of a different model selection criterion could lead to the selection of a different model, with respect to both the error term distribution and the relevant regressors. Finally, it should be noted that this analysis is performed on aggregated variables. This aggregation over respondents implies a reduction of the information originally present in the collected data. Thus, the obtained results do not take into account any differences between individual responses. In order to fully exploit this information, regression models for ordinal responses could be considered (see, for example, Faraway, 2006). This approach could also take account of the hierarchical structure of the dataset (respondents are nested within local health units) by considering multilevel models (see, for example, de Leeuw and Meijer, 2008).

The multivariate linear regression methods used in this paper seems to represent a useful and flexible strategy for evaluating public services by handling possible non-normal error terms in the regression model. However, some theoretical aspects are still under study, namely the Hessian of the log-likelihood function, the asymptotic covariance matrix of the model estimators and the inclusion of restrictions on the estimation of the regression coefficients. In particular, the development of an estimation procedure that allows for linear restrictions on the regression coefficients could be very useful in practical applications: allowing for different sets of restrictions for the D dependent variables may lead to the selection of a different subset of relevant regressors for each response and, thus, to the identification of more parsimonious and more flexible models.

References

Bartolucci F., Scaccia L. (2005), The use of mixtures for dealing with non-normal regression errors, *Computational Statistics and Data Analysis*, 48, 821–834.

de Leeuw J., Meijer E. (Eds.) (2008), *Handbook of multilevel analysis*, Springer, New York.

Dempster A.P., Laird N.M., Rubin D.B. (1977), Maximum likelihood for incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1–22.

Diaz-Garcia J.A., Rojas M.G., Leiva-Sanchez V. (2003), Influence diagnostics for elliptical multivariate linear regression models, *Communications in Statistics - Theory and Methods*, 32, 625–642.

Efron B., Tibshirani R.J. (1993), *An introduction to the bootstrap*, Chapman & Hall, London.

Faraway J.J. (2006), *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, Chapman & Hall, Boca Raton.

Ferreira J.T.A.S., Steel M.F.J. (2004), Bayesian multivariate skewed regression modeling with an application to firm size, in: Genton M.G. (ed.), *Skew-elliptical distributions and their applications: a journey beyond normality*, Chapman & Hall, Boca Raton, 174-189.

Fraley C., Raftery A.E. (2002), MCLUST: software for model-based clustering, density estimation and discriminant analysis, Technical Report No. 415, Department of Statistics, University of Washington.

Fraley C., Raftery A.E. (2003), Enhanced software for model-based clustering, *Journal of Classification*, 20, 263–286.

Galea M., Paula G.A., Bolfarine H. (1997), Local influence in elliptical linear regression models, *Statistician*, 46, 71–79.

Gandek B., Ware J.E.Jr., Aaronson N.K., Apolone G., Bjorner J.B., Brazier J.E., Bullinger M., Kaasa S., Leplege A., Prieto L., Sullivan M. (1998), Cross-validation of item selection and scoring for the SF12 health survey in nine countries: results from the IQOLA project, *Journal of Clinical Epidemiology*, 51, 1171–1178.

Gargiulo L., Iannucci L., Quattrociocchi L., Sebastiani G., Tinto A. (2008), Innovazioni di processo nell'indagine Istat sulla salute, *I quaderni di Monitor*, Trimestrale dell'Agenzia nazionale per i servizi sanitari regionali, Terzo supplemento al numero 22, 16–27.

ISTAT (2006), Il sistema di indagini sociali multiscopo, *Metodi e Norme*, 31.

Liu S. (2002), Local influence in multivariate elliptical linear regression models, *Linear Algebra and its Applications*, 354, 159–174.

Mardia K.V., Kent J.T., Bibby J.M. (1979), *Multivariate analysis*, Academic

Press, London.

McLachlan G.J., Krishnan T. (2008), *The EM algorithm and extensions, Second edition*, Wiley, Chichester.

McLachlan G.J., Peel, D. (2000), *Finite mixture models*, Wiley, Chichester.

Miller A. (2002), *Subset selection in regression, Second edition*, Chapman & Hall, Boca Raton.

Montgomery D.C., Peck E.A., Vining G.G. (2006), *Introduction to linear regression analysis, Fourth edition*, Wiley, New York.

R Development Core Team (2008), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. `URL http://www.R-project.org.`

Schwartz G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, 6, 461–464.

Soffritti G., Galimberti G. (2009), Multivariate linear regression with non-normal errors: a proposal based on mixture models (submitted for publication).

Srivastava M.S. (2002), *Methods of multivariate statistics*, John Wiley & Sons, New York.

Sutradhar B.C, Ali M.M. (1986), Estimation of the parameters of a regression model with a multivariate t error variable, *Communications in Statistics - Theory and Methods*, 15, 429–450.

Ware J.E., Sherbourne C.D. (1992), The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, 30, 473–483.

Yakowitz, S.J. and Spragins, J.D. (1968), On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39, 209–214.

Zellner A. (1976), Bayesian and non-Bayesian analysis of the regression model with multivariate student-t error terms, *Journal of the American Statistical Association*, 71, 400–405.