MULTIVARIATE DATA ANALYSIS TECHNIQUES TO DETECT EARLY WARNINGS OF ELDERLY FRAILTY[1]

Aldo Corbellini, Marco Riani

Sezione di Statistica e Informatica, Dipartimento di Economia
Università degli studi di Parma – Parma, Italy
aldo.corbellini@unipr.it – mriani@unipr.it

Andrea Donatini

Direzione Sanitaria, Azienda USL di Parma – Parma, Italy, adonatini@ausl.pr.it

Abstract

In this paper we outline the results of a project intended to detect the most important leading indicators of elderly frailty in a health district of the province of Parma (Italy). Through the joint application of several multivariate data analysis techniques we found a series of factors which are predictive of loss of functional and cognitive ability and we set up a model which provides early warnings of cases of potential elderly frailty. This allows social and health care workers of the Local Health Unit (LHU) of Parma and local administrators to take preventive action, in order to avoid cases of worsening of the physical and mental status.

Keywords: Generalized linear model, Cluster analysis, Principal Component Analysis, Ordinal Regression, Correspondence Analysis.

1. INTRODUCTION

The words "*elderly frailty*" are used to denote the state of "older persons (usually over the age of 75 years) who are afflicted with physical or mental disabilities that may interfere with the ability to independently perform activities of daily living" (Source: on line Medical dictionary[2]).

---

[1] 1 We are grateful to Professors Anthony Atkinson and Sergio Zani for providing helpful suggestions on a preliminary version of the paper. This work was partially supported by the grant "Metodologie statistiche per l'analisi di impatto e la valutazione della regolamentazione" of Ministero dell'Università e della Ricerca – PRIN 2006, (principal investigator C.N. Lauro).

[2] http://medical-dictionary.thefreedictionary.com/frail+elderly

According to what is known in the literature as the Fried definition (Fried *et al.* 2001), frailty is not considered as a disease, but rather as a sort of intermediate state between being functional and nonfunctional, and between being healthy and being sick. Frailty is a vitally important issue in the treatment of the elderly. It is something that most people who live to an advanced age will probably face (Wolf *et al.*, 2001). Frailty can strongly affect how an elderly person will respond to medical treatment, as well as how long and how well he/she will live. Doctors should try to recognize frailty early and institute appropriate therapies. This alone would greatly enhance the quality of life of many older persons. Statistically, a fairly large percentage of those who meet the definition die within five years (Morley *et al.,* 2001). With this in mind, it is important to consider what are the early warnings of frailty. Surprisingly, though common, frailty remains poorly understood (Morley, 2009).

Central to the clinical concept of frailty is that "no single altered system alone defines frailty, but that multiple systems are involved" (Lang *et al*. 2009). This implies that the effective detection of frailty can take place only if we consider a large set of potential explanatory variables and a set of different statistical tools in order to investigate its different aspects. The purpose of this paper is to show how the combined use of several multivariate statistical techniques can help to develop a model to detect early warnings of "elderly frailty" in order to avoid the worsening of the physical and mental status. This research may be considered in the context of AIR (*Analysis of Impact of Regulation*[3]) because it deals with the ex ante evaluation of the decision of the public administrations. Clearly the possibility of understanding in a deeper way what are the main factors of predictive of loss of functional and cognitive ability can help social care workers and local administrators to take preventive actions and introduce new models of care for frail older people (Senin *et al*., 2003).

The paper is structured as follows. In section 2 we introduce the survey which has been conducted and the characteristics of the available database. In section 3 we present the results from applying bivariate analysis, symmetric and non symmetric correspondence analysis techniques. In section 4 we briefly recall the generalized model of ordinal analysis and we perform a variable selection method in order to understand what are the most important indicators to detect early frailty. In section 5 we discuss the implications of the results which have been obtained. Section 6 concludes and provides food for thought, not only for further research, but also for those with responsibility in the LHU (Local Health Unit) of Parma and for the local authorities.

---

[3] *The Italian law 28 November 2005, n. 246 (resuming the dictation of law n. 50 of 1999) defines the Analysis of Impact of Regulation (AIR) as "the preventive appraisal of the effects of normative regulations concerning the activities of the citizens and the enterprises and the organization and the operation of the Public Administrations, by means of comparisons of alternative options".*

2. THE AVAILABLE DATA

The data we use come from a survey promoted by the Social Policy Department of the Province of Parma and by the LHU of Parma. The data were collected in two phases. In the first, which was conducted during the first half of the year 2007, demographic, social and health data were collected on the elderly population in a health district of the province of Parma. All the available information was directly collected by doctors working in the same municipality, because they seemed the most appropriate experts to evaluate the situations of the elderly people.

Given the multidimensional characteristic of the "elderly frailty phenomenon", the doctors were asked to provide detailed information about a plethora of characteristics of elderly people ranging from "demographic aspects", to "housing conditions", from "help and support" variables to "social relationships" and from "accessibility" to "actions". For example, under the group "actions", the ability of the subjects under study to carry out basic activities of daily life was investigated in terms of more than 20 variables dealing with physical mobility, cognitive aspects, memory and orientation, ability to use money, etc. In order to have an idea of the completeness of the survey which was carried out, Table 1 shows the classification of the collected indicators into homogeneous groups and an example of the variables falling inside each group.

Tab. 1: Classification of the variables into homogeneous groups.

Set of indicators	*Examples of the indicators falling under each set*
Demographic	Age, sex, marital status, …
Pharmacological behaviour	Numbers of medications per day, number of hospitalizations, medications taken correctly, …
Housing conditions	Presence of architectural barriers, presence of bathroom inside the house, quality of the house, …
Help and support	Presence of caregiver, effectiveness of the caregiver, presence of an accompanying allowance, …
Social relationships	Degree of isolation, presence of sons or relatives, degree of participation in social life, …
Accessibility	Road Accessibility, Accessibility to public transports, Distance from the first village, …
Actions	Ability of washing, getting up, walking, going to the bathroom, …

At the end of the questionnaire the doctors were also asked to provide a final classification of each subject in terms of his/her state of health (dependency status) using the categories shown in Table 2.

Tab. 2: Classification of the elderly people in terms of their state of health and relative score.

State of health	*Score*
Self dependent	1
Partially self dependent (slightly dependent)	2
Dependent	3
Completely dependent	4

In additional to all the variables which were collected by the doctors, the LHU of Parma added two indexes (namely Barthel and mini-mental[4]) derived from the set of variables of the questionnaire.

In the second phase of the survey, which took place at the end of year 2008 and at the beginnings of year 2009, a follow up of the population previously investigated was carried out in order to investigate the new "degree of dependency" for the subjects who were still alive adopting the same scoring system used in Table 2. Computing the differences between the scores in the two periods (given by the same doctor), we obtained values between -3 (which correspond to the maximum worsening of the conditions, that is to a subject who in the first wave was self dependent and in the second was completely dependent) and +3 (which correspond to the maximum improvement of the health condition[5]).

These scores have been used to obtain a new qualitative variable measured on an ordinal scale with the four modalities (classes) shown in Table 3 with their relative coding.

Tab. 3: Construction of the ordinal variable 'variation in health conditions'.

Difference in health conditions between the two surveys	**Difference between the scores in the two surveys**	**Final code assigned to the variation in health conditions**
Improved	<0	-1
Stable	=0	0
Slightly worse	=1	1
Serious deterioration	>1	2

4 *The Barthel index is a disability profile scale developed by D.W. Barthel in 1965 to evaluate a patient's self-care abilities in 10 areas, including bowel and bladder control. The patient is scored from 0 to 15 points in various categories, depending on his or her need for help, such as in feeding, bathing, dressing, and walking. The mini-mental index is a practical method for grading the cognitive state of a subject (Folstein et al., 1975).*

5 *In cooperation with the LHU we passed under close scrutiny the cases which in the first wave were classified as "completely dependent" and in the second as "self dependent". As a result of this we found that the two only one cases with difference in score equal to +3 were due to transcription errors.*

As concerns the size of the sample, the final number of the subjects under study was 2374, while the number of variables was greater than 100.

3. BIVARIATE ANALYSES

Due to the impossibility of considering simultaneously all the indicators and in order to preliminarily exclude the variables which had no effect on the "variations of the health conditions", we decided to carry out a bivariate analysis between each of the explanatory variables and the "variation of the health conditions". Namely, for each crossing we computed the usual association indexes (chi square, Cramer…) and in order to have a greater insight about the contribution of the single classes to the association, we performed a symmetric (Greenacre, 2007; Greenacre and Blasius, 2006), a non symmetric (D'Ambra and Lauro, 1992; Lauro and D'Ambra, 1984) and an ordinal (Beh, 2008) correspondence analysis.
As an illustration of what has been done for each pair of variables, consider the joint classification between the explanatory variable "degree of isolation" and the response "variation of health conditions".
Both the two chi-square indexes (traditional chi-squared and Goodman-Kruskal τ index) have a *p*-value smaller than 0.001. This clearly suggests that the relationship between "degree of isolation" and "variation of the health conditions" is significant. If we give a closer look to the table, we can see that the relative frequencies inside the profile "Stable" increase monotonically passing from "High isolation" to "Absence of isolation" (third column of Table 4). We can see the opposite phenomenon inside the profiles "Slightly worse" and "Serious deterioration" (fourth and fifth columns of Table 4).

Tab. 4: Cross tabulation between degree of isolation and variation of the health conditions (absolute and relative row frequencies).

	Improved	Stable	Slightly worse	Serious deterioration	RowTotal
	112	1318	187	69	1686
Absent	0.07	0.78	0.11	0.04	0.71
	55	258	45	24	382
Low	0.14	0.68	0.12	0.06	0.16
	43	152	43	30	268
Medium	0.16	0.57	0.16	0.11	0.11
	3	21	10	4	38
High	0.08	0.55	0.26	0.11	0.016
	213	1749	285	127	2374
ColTotal	0.09	0.74	0.12	0.05	1

To have a greater insight about the relationships between these two variables, we performed both a symmetric (CA) and asymmetric correspondence analysis (NSCA). As requested by the referee, in the Appendix we briefly recall the basic ideas of NSCA. Figure 1 and Figure 2 show respectively the correspondence analysis plot using the symmetric and asymmetric approaches. The two principal latent dimensions explain more than 99% of the association (inertia) between the two original variables. Therefore the 2 dimensional representation is highly reliable. In order to have a greater insight about the degree of overlapping of the different modalities of the rows, and in order to identify the classes which mostly contribute to the association between the variables, in the previous plots we also superimposed the circles associated with the 95% confidence intervals (Lebart, Morineau and Warwick, 1984). If the

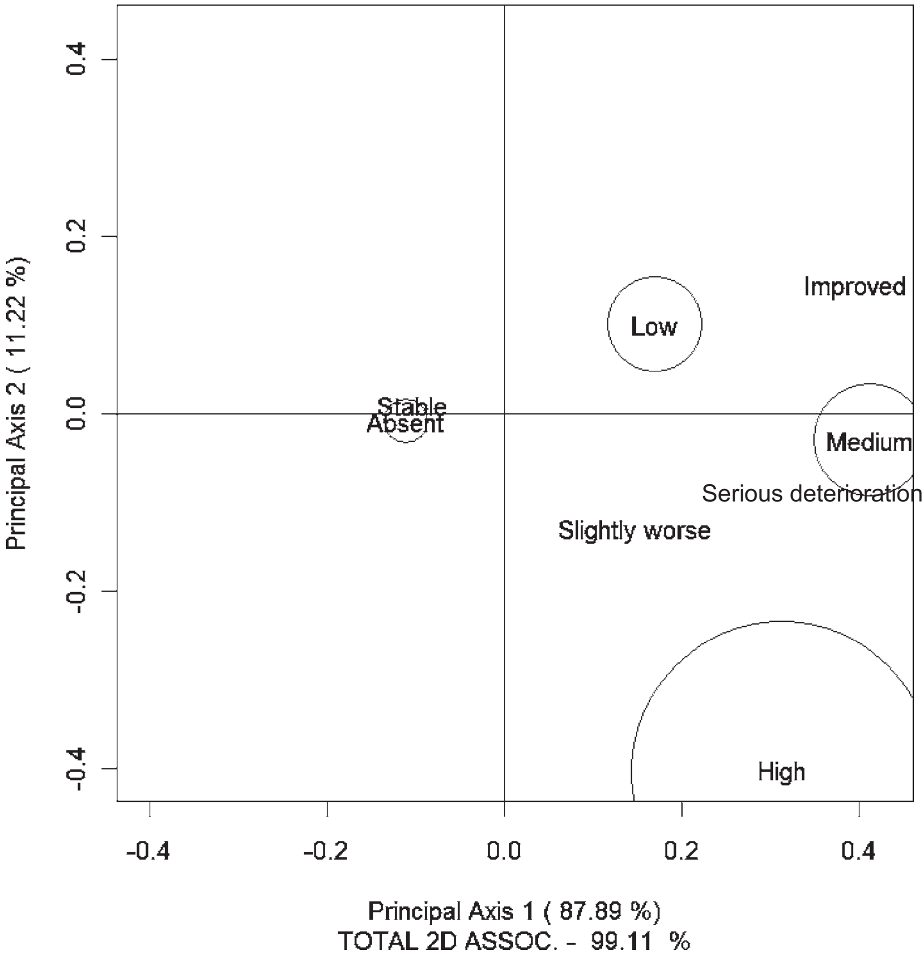

Fig. 1: Symmetric correspondence analysis: profile coordinates plot.

origin lies outside the confidence circle for a particular response variable, then that category contributes to the prediction of the response variable and vice-versa (Lombardo *et al*., 2007). Figure 1 and Figure 2 clearly show that people with an absence of isolation mainly have a stable pattern of the health conditions. On the other hand, people who have a medium or high degree of isolation are mainly associated with slight or serious deterioration of their health conditions. Given that all the confidence circles do not intersect the origin we can state that all categories of isolation contribute to explain the variations in health conditions. Finally, from the size of radii of the circles, we can easily infer that the classes of isolation which mostly contribute to the association among the two variables are "High" and to a lesser extent "Medium".

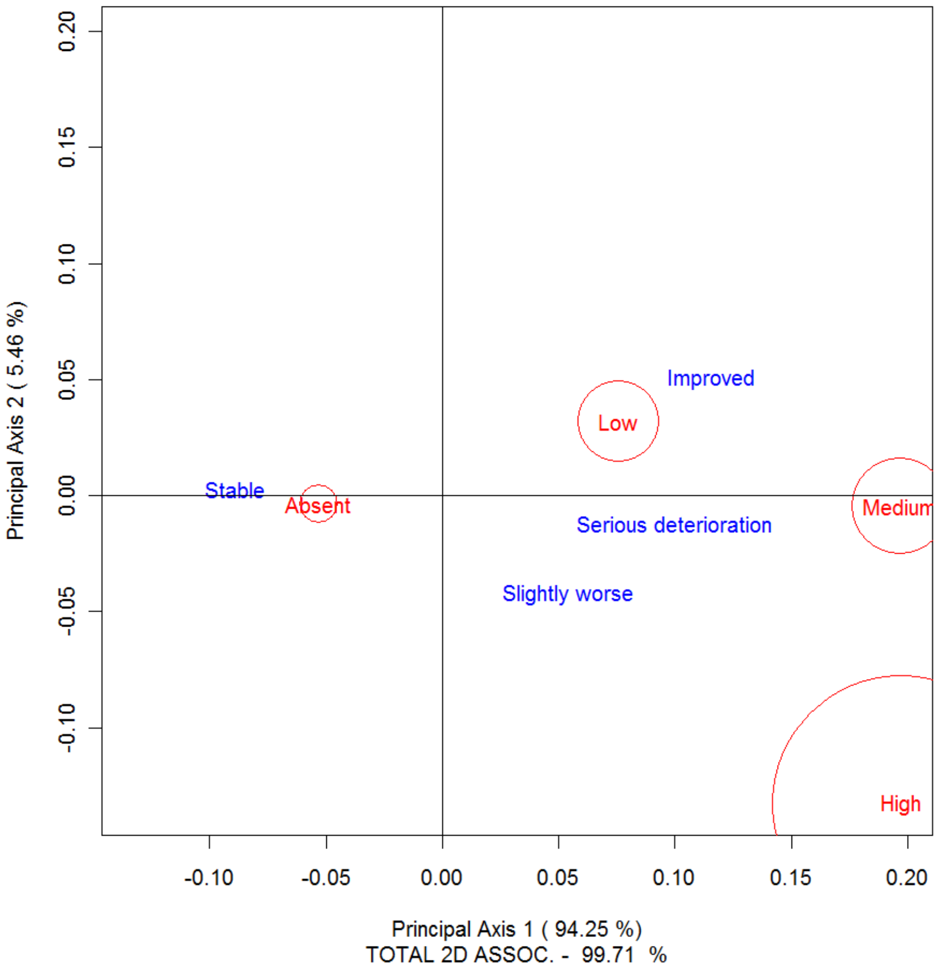

Fig. 2: Asymmetric correspondence analysis: profile coordinates plot.

A similar analysis to the one just seen for "Isolation", has been carried out for all the other pairs of variables. Table 5 shows the explanatory variables which turned out to be significant at the 0.10 level with the variation of the state of health between the two surveys, the modalities in which they are expressed and the sign of the relationship with the response. Of course, while certain relationships must be considered obvious (for example the relationship with age) others were unexpected. Surprisingly, there was no relationship between the presence of sons alive and the variation in the health condition. On the contrary, all the variables concerning the "degree of participation in social life" and "relationship with neighbours" were highly significant. Similarly, also the variables concerning marital status were significant. On the other hand, all the variables linked to the actions of the elderly were not significant and the medical indexes Barthel and mini-mental had a difficult interpretation (Table 5).

Remark: for each variable the doctors who collected the data had the possibility of indicating with the code NK (not known) the ignorance of the subject status with respect to that particular variable. For each explanatory variable with NK values, we repeated the analysis both with and without the modality NK. In this survey it is clear that NK cannot be considered a missing value at random, because these empty cells are associated mainly with patients who rarely see their doctor. The health conditions of these patients cannot be considered a random sample from the elderly population in the health district.

In conclusion, we can say that the set of bivariate analyses has fully confirmed results in the scientific literature about the elements which greatly contribute to elderly frailty: advanced age, presence of chronic pathologies, complex pharmaceutical treatments and inadequate socio-environmental conditions (for example Senin *et al.*, 2003). In addition, the bivariate analysis has stressed the importance of the variables linked to the degree of isolation, relationships with neighbours and participation in social life, in determining early warnings of the variation of the state of health.

Tab. 5: List of the variables significant at the 0.10 level for the variation of the state of health (response variable). NK stands for not known by the doctor.

Variables	Classes	Summary of the relationship with the response
Age	75-79, 80-84, 85-89, 90-94, 95+	As expected class "Serious deterioration" is associated with the class 95+, while class "stable" is mainly associated with 75-79.
Sex	Male, Female	The relative frequency of women who experience an improvement in health conditions is twice that of men.
Altimetric area	Plain, Hill, mountain	The highest percentage among those who had a serious deterioration can be found among subjects living in the mountains.
Marital status	Married, Unmarried, Widow, NK	Among those who were stable, the highest percentage was for married people. Among those with serious deterioration, the percentage of widows is twice that of married.
Typology of the family	Live alone, Couple, Family, NK	Among the 'Improved', the highest percentage by far can be found among those living in a family. On the other hand, the modality "seriously worse" is associated with those living alone.
Accom-panying subsidy	Yes, No	Among the subjects who showed an improvement, the percentage of those who had the accompanying subsidy was double that of those without it.
Relation-ship with neighbours	Frequent, Sporadic, Absent, NK	In the class "Serious deterioration" we can mainly find people with "Absent relationship". Among those who were stable, the most common modality is "Frequent".
Degree of participation in social life	High, Medium, Low, Absent, NK	In the classes "Slightly worse" and "Serious deterioration" the most common modality is "Absent". Among those who were "Stable" the most common modalities are "High" and "Medium" degree of participation in social life.
Presence of caregiver	Yes, No	The percentage of people with "improvement" and "Slightly worse" is by far greater among those who have a caregiver.
Possibility of receiving help from neighbours	High, Sufficient, Medium, Not sufficient, Very low, NK	In the class "Serious deterioration" we have mainly cases with "Very low" and "Not sufficient" possibility of receiving help. On the other hand, in the class "Stable" we can find more frequently cases with classes "High" and "Sufficient".
Degree of the relationship with the community	Adequate, Partially adequate, Inadequate, Absent	Among people with "Slight deterioration" we can find mainly cases with class "Inadequate". On the contrary, among those with class "Adequate", we can find mainly people who remained in stable conditions.
Accessibility	Good, Bad	Bad accessibility is mainly associated with subjects who had serious deterioration of health conditions.
Accessibility to public transport	Good, Bad, Absent	Stable profile is mainly associated with good accessibility to public transport. Class absent, on the other hand, is mainly associated with slight or serious deterioration of health conditions.

continues

continues

Typology of heating	Gas, wood, NK	Serious deterioration is mainly associated with "wood heating".
Hygienic conditions of the house	Good, Bad, NK	Bad hygienic conditions are associated with slight and serious deterioration of the health conditions.
Isolation	Absent, Low, Medium, High	There is a very strong relationship between degree of isolation and variation of health conditions. Medium and high isolation are mainly associated with strong deterioration. On the other hand, absent and low isolation is strongly associated with stable or improved conditions
Characte-ristics of the house	Adequate, Partially adeq., NK, Inadequate	Strong deterioration is mainly associated with Inadequate house conditions.
Number of hospital admissions	0, 1, 2, 3, 4, >4	The profile stable is mainly associated with 0 admissions. Inside the classes with more than 1 hospital admission there is a lot of variability.
Number of medicines per day	0, 1, 2-3,4-5, >5	The class ">5" has a strong association with "serious deterioration", while classes with values 0 and 1 are mainly associated with stable conditions.
Correct consump-tion of medication	Yes, No	Stable condition, as expected, is mainly associated with correct consumption of medication
Complexity of the therapeutic regime	Null, Low, Medium, High	Stable condition is mainly associated with class "Null". Among those with slight and serious deterioration, the most frequent class is "High".
Alimentary habits and hydration	Adequate, Not adequate	Stable condition is mainly associated with adequate alimentary habits and hydration.
Barthel (depen-dency) index	Self, Slight, Medium, Total dependence	Difficult interpretation because in the contingency table the joint frequency of cases with Barthel values equal to total dependence and medium dependence and "strong deterioration of health conditions" is zero.
Mini-mental (depen-dency) index	None, Slight, Serious	Difficult interpretation because among those who showed an improvement in health conditions and slight deterioration the most common profile for mini-mental index is "Slight". Stable conditions are mainly associated with class "None" for the mini-mental index.

4. ORDINAL REGRESSION TO PREDICT ELDERLY FRAILTY

In this section we apply the generalized linear model of ordinal regression using as response variable the "variation of the state of health in the two waves of the survey" and as potential explanatory variables all those which turned out to be significant using bivariate analysis (see Table 5). This section is structured as follows. In subsection 4.1 we briefly review the building blocks of ordinal regression in order to better understand the results we have obtained. In subsection 4.2 we discuss our model for elderly frailty prediction.

4.1 THEORY

In most cases it is implausible to assume the normality and homogeneity of variance for ordered categorical outcome when (as in the case of our application) the ordinal outcome contains merely a small number of discrete categories (Chu and Ghahramani, 2005). Thus, the ordinal regression model becomes a preferable modelling tool, because it does not assume normality and constant variance, but simply requires the assumption of parallel lines, that is the assumption that the location parameters are equivalent across the levels of the dependent variable (categorical outcome).

More precisely, the ordinal regression model (McCullagh, 1980) may be written in the following form:

$$link(\gamma_{ij}) = \alpha_j - [b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}], \; j=1, \ldots k-1, \text{ and } i=1, \ldots, n. \quad (1)$$

where γ_{ij} is the cumulative distribution function for the j-th category of the i-th case. In our case the number of categories k-1 is equal to 3 (-1=improved, 0=stable, 1=slightly worse). As usual, one of the classes of the response (in this case the category "serious deterioration") must be omitted from the model because it is redundant. For example in our case γ_{i2} is the cumulative probability that the i-th subject has had improved or stable health conditions.

$link(\gamma_{ij})$ is the so called "link function" which is typical of generalized linear models. In the case of ordinal regression, it is a transformation of the cumulative probabilities of the ordered dependent variable that allows for estimation of the model. The most commonly used links are Logit, Probit, Negative log-log, Complementary log-log and Cauchit (Cauchy). While the negative log-log link function is recommended when the probability of the lower category is high, the Complementary is particularly suitable when the probability of higher category is high. Finally, while the probit link is more suitable when a dependent variable is normally distributed, the Cauchit link is mainly used when extreme values are present in the data.

The parameter α_j is the threshold value of the j-th category. The α_j terms often aren't of much interest because their values do not depend on the values of the independent variables for a particular case. They are like the intercept in a linear regression, except that each j-th level of the response has its own value. The model is based on the assumption that there is a latent continuous outcome variable and that the observed ordinal outcome (which in our case is the variation in the health conditions) arises from discretizing the underlying continuum into k-ordered groups. The thresholds estimate these cut-off values.

The number of regression coefficients is p, $b_1, \ldots, b_p$ are a set of regression coefficients and $x_{i1}, x_{i2}, \ldots, x_{ip}$ are a set of p explanatory variables for the i-th subject.

As concerns the part of the model $-[b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{ip}]$, it is interesting to notice two aspects:

1. The negative sign before the square brackets ensures that larger coefficients indicate an association with larger scores. In our case, for example, a positive coefficient of a class of an explanatory variable means that people who are in that class have a greater probability of experiencing a worsening in health conditions.
2. What is inside the square brackets does not depend on j. In other words, if for example the link is logit, this implies that the effect of the independent variables is the same for all the logits. That means that the results are a set of parallel lines or hyperplanes, one for each category of the outcome variable. It is possible to test this assumption using the so called "test of parallel lines" against the alternative that the relationships between the independent variables and logits are not the same for all logits. If this test is rejected, it is necessary to include into the model a scale component:

$$link(\gamma_{ij}) = \frac{\alpha_j - \left[b_1x_{i1} + b_2x_{i2} + \cdots + b_px_{ip}\right]}{\exp\left(\tau_1 z_{i1} + \cdots + \tau_k z_{ik}\right)} \tag{2}$$

The numerator of equation (2) is known in the literature as the **location** of the model, while the denominator specifies the scale. The $\tau_1, \tau_2, \ldots, \tau_r$. are coefficients for the scale component and the $z_{i1}, z_{i2}, \ldots, z_{ir}$ are predictor variables for the i-th subject for the scale component (chosen from the same set of variables as the x's). The **scale component** accounts for differences in variability for different values of the predictor variables. For example, if certain groups have more variability than others in their ratings, using a scale component to account for this, generally improves the model. In our study the variability in health conditions among subjects who are in more advanced classes of age is much greater than that associated with people who are in class 75-79.

4.2 RESULTS FROM THE APPLICATION

The best link turned out to be the logit[6]. This is expected given that in our case the most likely category of the response is not one of the extreme values (see Figure 3).

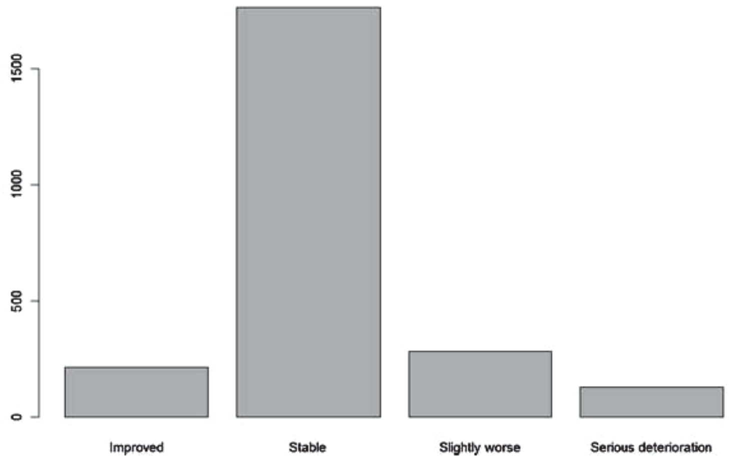

Fig. 3: barplot of the "variation in health conditions".

We initially tried model (1), but given that, as expected, the test of parallel lines was clearly rejected, we ended up using the specification with the scale component (equation 2).

The best model was chosen among all candidate models based on the model fitting statistics, the accuracy of the classification results, the validity of the model assumption, and the principle of parsimony. As concerns the variable selection process, in order to check the stability of the final model, we started the model selection procedure from different starting points. Table 6 shows the coefficients of the location component of the final generalized ordinal regression model. All the variables given in the table were significant at the 5% level. As is always the case with categorical predictors in models with intercepts, the number of coefficients displayed is one less than the number of categories of the variable. The variables which turned out to be significant in the scale component were: age, altimetric area, isolation and degree of participation in social life. The coefficients of the variables included in the scale component are not given here for lack of space.

---

6 *Very similar results to those obtained with the logit link have been found using the probit link specification.*

Tab. 6: Results of the ordinal generalized linear model: estimated thresholds, list of the variables which were significant in the location component and estimated coefficients.

	Estimated coefficients
Thresholds	
y=-1 (Improved)	-1.929
y=0 (Stable)	0.315
y=1 (Slightly worse)	1.184
y=2 (Serious deterioration)	-
Age	
75-79	-0.18
80-84	-0.083
85-89	-
90-94	0.081
95+	0.170
Civil status	
Widow	-0.123
Married	-0.139
Unknown	-0.495
Unmarried	-
Altimetrical area	
Plain	-0.228
Hill	-0.127
Mountain	-
Isolation	
High	0.195
Medium	-0.086
Low	-0.171
Absent	-
Degree of participation in social life	
Unknown	0.437
High	-0.170
Medium	-0.187
Low	-0.109
Absent	-
Number of hospitalizations	
Quantitative variable	-0.165

Table 7 gives a summary of the relationships between the set of the explanatory variables which were significant and the response.

Tab. 7 List of the variables which were significant in the final model and analysis of the sign of the relationship with variation of the health conditions.

Variables	*Description of the relationship with the response*
Age	As expected, with the increase of age, the probability of deterioration grows.
Altimetrical area	Those who live in the hills or in a mountain area have a probability of experiencing a deterioration in health conditions greater than those who live in the plain. In other words, the risk factor increases monotonically passing from plain, to hill and to mountain.
Civil status	The presence of a family kernel around the elderly or the presence of a couple of elderly reduces the probability of worsening in health conditions. On the other hand, keeping all the other conditions fixed, the condition unmarried is associated to a higher risk of worsening of health conditions.
Isolation	Those who suffer strong isolation have a higher probability of undergoing a worsening of the health state.
Degree of participation in social life	Elderly people who do not participate in social life have a high probability of experiencing a deterioration of health conditions.
Number of hospitalizations	The probability of an improvement of the health conditions increases with the number of hospitalizations

Finally, as concerns the model diagnostics for this generalized linear model (Atkinson and Riani, 2000), which are not given here for lack of space, we can say that: *a*) goodness of fit tests clearly revealed the joint importance of the selected set of variables in forecasting the variation of the state of health; *b*) The analysis of the residuals did not reveal clear cases of atypical observations and there were no systematic patterns.

5. IMPLICATIONS FOR LHU AND LOCAL ADMINISTRATORS

In order to have a more precise and quantitative idea of the increase of risk factors, we computed the theoretical probability of having a slight or a serious deterioration of the health conditions under particular common patterns. We considered the typical pattern of a person with age in class 75-79, who lives in a mountain area, is unmarried, has a high level of isolation, an absence of participation in social life and no hospitalizations. According to our model, this subject has a theoretical probability of experiencing in the next two years a serious deterioration of the health conditions equal to 0.129 and a generic worsening (slight or serious deterioration) equal to 0.380. In Table 8 and in Figure 4 we analyze how this probability (keeping all other factors fixed) varies with: age, altimetric area of residence, degree of participation in social life, isolation or civil status.

Quick facts which come out from Table 8 and from Figure 4 can be summarized as follows:

- the risk factor linked to serious deterioration of health conditions is 22.2% greater for those who are unmarried compared to those living in a couple;
- the risk factor of undergoing a worsening of the health conditions is 37.5% greater for those who live in a mountain area;
- those who have high isolation have a risk factor of worsening 42.6% greater than those who have low isolation;
- people who have an absence of participation in social life have a risk factor of serious deterioration 290.9% (generic worsening 62.4%) greater than those who have high relationships with the community.

These results have strong implications for doctors, LHU of Parma and in general for local administrators in terms of Analysis of Impact of Regulation (AIR) because, while the age of the person cannot be changed and the place where the elderly live generally cannot be changed, the possibilities of intervention to reduce the degree of isolation or to increase participation in social life are very high.

In conclusion, the joint application of different statistical techniques of analysis found predictive factors of functional and cognitive disability which can allow social and health care workers to take preventive action thus avoiding worsening of the physical and mental status of the elderly population.

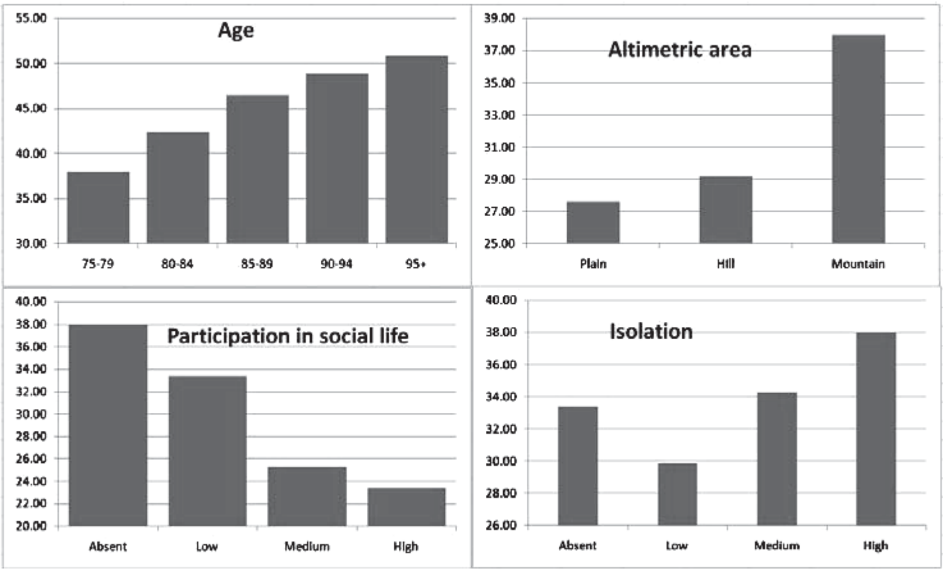

Fig. 4: Analysis of the relationship between the estimated probability of a worsening of health condition with: age, altimetric area, participation in social life and isolation.

Tab. 8: Monitoring of the estimated probabilities of serious deterioration of the health conditions or generic worsening (serious deterioration + slight worsening) for an unmarried person with age 75-79, who lives in a mountain area, has a high level of isolation and no participation in social life when every single factor varies.

Age changes	Serious deterioration	Worsening	*Area changes*	Serious deterioration	Worsening
75-79	12.9%	38.0%	Plain	7.3%	27.6%
80-84	16.4%	42.4%	Hill	6.4%	29.2%
85-89	23.9%	46.5%	Mountain	12.9%	38.0%
90-94	26.3%	48.9%			
95+	36.0%	50.9%			
Degree of participation in social life changes	Serious deterioration	Worsening	*Isolation changes*	Serious deterioration	Worsening
Absent	12.9%	38.0%	Absent	13.0%	33.4%
Low	10.4%	33.4%	Low	12.4%	29.9%
Medium	4.7%	25.3%	Medium	16.4%	34.3%
High	3.3%	23.4%	High	12.9%	38.0%
Civil status changes	Serious deterioration	Worsening			
Widow	10.8%	33.4%			
Married	10.6%	32.8%			
Unmarried	12.9%	38.0%			

7. CONCLUSIONS

Frailty has always been with us, though today it is increasingly seen not as an inevitable part of ageing, but as a condition that in many cases can and should be treated aggressively. Many of the causes of frailty are treatable and even reversible through a combination of appropriate medical treatment, maintenance of a good diet and a good exercise regime. It is extremely important therefore to detect the early symptoms of frailty in order to carry out the appropriate actions.

In this paper we have set up a generalized linear model based on ordinal regression which helps us to detect cases of potential frailty. Thanks to this model doctors can have, through the estimated frailty probability, an additional tool for early warnings about the potential frailty condition of each person who is visited.

Finally, we have been able to outline in a quantitative way how the risk factor of frailty changes with aspects in which the LHU responsible can intervene like the "degree of isolation" and "the degree of participation in social life".

The methodological output which can be extracted from this paper is that in presence of a very large number of potential explanatory discrete variables it is better to start reducing their number using the results of the chi square tests between each of them and the response before starting the model selection procedure. From the applied point of view and the implications in terms of Analysis of Impact of Regulation (AIR), the results of this paper have confirmed the general impression existing in the literature that increasing frailty is distinguishable from ageing and, as a consequence is potentially reversible, provided appropriate actions are taken by the doctors, the LHU responsible and the local administrators.

APPENDIX: basic concepts in non symmetric correspondence analysis

Traditional correspondence analysis considers the partition of the chi-squared statistic in two way contingency tables to graphically describe the association of categorical variables (e.g. Greenacre, 2007). On the other hand at the heart of the non symmetric correspondence analysis there is the partition of the Goodman-Kruskal τ index (Goodman and Kruskal, 1954). More precisely, let us consider a two-way contingency table of dimension $I \times J$ and denote the joint relative frequencies with p_{ij} so that $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$ and with p_{i*} and p_{*j} respectively the row and marginal frequencies. Without loss of generality, let us suppose that the variable in the row is the response while the variable in the column is the predictor. The conditional probability that an individual (unit) is classified into row i given that it belongs to column j is p_{ij}/p_{*j}. The difference between the conditional prediction p_{ij}/p_{*j} (column profile) and the unconditional marginal prediction p_{i*} (row marginal proportion) for row i and column j can be written as $\pi_{ij} = p_{ij} / p_{*j} - p_{i*}$. The proportional reduction of error in the prediction of the response variable, given the presence of a predictor (column) variable, can be evaluated used using the τ index defined as follows:

$$\tau = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{*j}(p_{ij} / p_{*j} - p_{i*})^2}{1 - \sum_{i=1}^{I} p_{i*}^2}$$

The numerator of this index is the overall measure of predictability of the rows given the columns, while the denominator measures the overall error in prediction. Under the null hypothesis that all distributions of the predictor variable are identical to the overall marginal distribution, namely $H_0 : p_{ij} / p_{*j} = p_{i*}$, the quantity $(n\text{-}1)(I\text{-}1)\tau$ is asymptotically chi-squared distributed with $(I\text{-}1)(J\text{-}1)$ degrees of freedom (Agresti, 1990). In order to obtain a graphical representation of the predictability of the response categories given the predictor categories, in NSCA the singular value decomposition is applied to the matrix whose generic element is $\pi_{ij} = p_{ij} / p_{*j} - p_{i*}$. For more details on the theory and application of NSCA the reader can see D'Ambra and Lauro (1992). An extension of NSCA to the case in which the two variables in the contingency table have an ordinal structure can be found, for example, in Lombardo *et al.* (2007).

REFERENCES

ARMSTRONG B.G., SLOAN M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology*, *129*(1), 191-204.

AGRESTI A. (1990). *Categorical Data Analysis*, Wiley, New York.

ATKINSON A.C., RIANI M., (2000). *Robust Diagnostic Regression Analysis*, Springer Verlag, New York.

BEH E. (2008). Simple correspondence analysis of nominal-ordinal contingency tables, *Journal of Applied Mathematics and Decision Sciences*, Volume 2008, Article ID 218140

CACCIATORE F., ABETE P., MAZZELLA F., VIATI L., DELLA MORTE D., D'AMBROSIO D., ET AL. (2005). Frailty predicts long-term mortality in elderly subjects with chronic heart failure. *European Journal of Clinical Investigation*; 35(12), 723-30.

CHU, W., GHAHRAMANI, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, *6*, 1019-1041.

D'AMBRA L., LAURO, N.C. (1992). Non symmetrical Exploratory Data Analysis, *Statistica Applicata*, 4, 511-530.

FOLSTEIN M.F., FOLSTEIN S.E., McHUGH P.R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*., 12, 189-198.

FRIED L.P., TANGEN C.M., WALSTON J., NEWMAN A.B., HIRSCH C., GOTTDIENER J., SEEMAN T., TRACY R., KOP W.J., BURKE G., MCBURNIE M.A. (2001). Frailty in older adults: Evidence for a phenotype. *J. Gerontology*, 6(3), 146-56.

GILLICK M. (2001). Pinning down frailty. *J. Gerontology*, 56(3), 134-5.

GOODMAN L.A., KRUSKAL, W.H. (1954), Measures of association for cross classifications, *Journal of the American Statistical Association.* 49, pp. 732–764

GREENACRE M. (2007). *Correspondence Analysis in Practice* (Interdisciplinary Statistics), 2nd edition, Chapman & Hall/CRC, Boca Raton, FL.

GREENACRE M., BLASIUS J. (eds) (2006). *Multiple Correspondence Analysis and Related Methods*, Statistics in the Social and Behavioral Sciences, Chapman & Hall/CRC, Boca Raton, FL.

LALL, R., CAMPBELL, M.J., WALTERS, S.J., MORGAN, K. (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research*, 11(1), 49-67.

LANG P.O., MICHEL J.P., ZEKRY D. (2009). Frailty Syndrome: A Transitional State in a Dynamic Process, *J. Gerontology*, 55(5), 539-49.

LAURO N.C., D'AMBRA L. (1984), L'analyse non symetrique des correspondances. In: E. Diday *et al.* (eds.), *Data Analysis and Informatics*, III, North-Holland, Amsterdam.

LEBART L., MORINEAU A., WARWICK K.M. (1984). *Multivariate Descriptive Statistical Analysis*, John Wiley, New York.

LOMBARDO R., BEH E., D'AMBRA L. (2007). Non-symmetric Correspondence Analysis with ordinal variables using orthogonal polynomials. *Computational Statistics and Data Analysis*, 52(1), 566-577.

MCCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, *42*(2), 109-142.

MORLEY J.E., BAUMGARTNER R.N., ROUBENOFF R., MAYER J., NAIR K.S. (2001). Sarcopenia. *J. Lab. & Clin. Med.*, 137(4), 231-43.

MORLEY J.E. (2009). Developing Novel Therapeutic Approaches to Frailty, *Current Pharmaceutical Design*, 15(29), 3384-3395, Bentham Science Publishers.

SENIN U., CHERUBINI A., MECOCCI P. (2003). "Impatto dell'invecchiamento della popolazione sull'organizzazione socio-sanitaria: necessità di un nuovo modello di assistenza continuativa", *Annali Italiani di Medicina Interna*, 18, 6-15.

WOLF S.L., SATTIN R.W., O'GRADY M., FRERET N., RICCI L., GREENSPAN A.I., XU T.S., KUTNER M.A. (2001). Study design to investigate the effects of intense Tai Chi in reducing falls among older adults transitioning to frailty, *Controlled Clinical Trials* 22(6), 689-704.

TECNICHE DI ANALISI MULTIVARIATA DEI DATI PER SCOPRIRE SEGNALI ANTICIPATORI DI FRAGILITÀ DEGLI ANZIANI

Riassunto

In questo articolo si riportano i risultati di un progetto di rilevazione delle principali caratteristiche demografiche, abitative, sociali, sanitarie e di autosufficienza della popolazione anziana residente in un Distretto dell'Azienda USL di Parma. L'analisi del grado di autosufficienza della popolazione è stato oggetto di un follow-up che ha consentito di costruire un indicatore associato alla variazione delle condizioni di salute. L'applicazione congiunta di svariate tecniche statistiche ha consentito di evidenziare le variabili più importanti che definiscono segnali anticipatori della variazione dello stato di salute e la perdita di autosufficienza nella popolazione anziana. Tutto ciò consente alle unità sanitarie locali ed agli amministratori di prendere opportune azioni preventive in modo di evitare che le famiglie si rivolgano ai servizi sociosanitari solo quando la situazione precipita verso la non autosufficienza per effetto di un evento grave e improvviso.