

Usage as Complementary Correspondence Analysis and Logistic Regression in a Scientific Survey on Self Healing Methods

Zerrin Asan Greenacre, Levent Terlemez, Sevil Sentürk

Statistic Department, Science Faculty, Anadolu Univesity, Eskisehir, Turkey

Email: zasan@anadolu.edu.tr, lterlemez@anadolu.edu.tr, sdeligoz@anadolu.edu.tr

Received 24 September 2014; revised 22 October 2014; accepted 12 November 2014

Copyright © 2014 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

The aim of this study is to show complementary usage of logistic and correspondence analysis in a research subject to self-healing methodologies. Firstly, the number of the variables is reduced by logistic regression according to relationship between dependent and independent variables and then research carries on searching variables. The relationship among the behaviours of individuals and their demographic characteristics is modelled by logistic regression and shown graphically by correspondence analysis. In application, first of all, the effect of age, sex, marital status, education level, occupation and income level and present health condition, on appreciating self-health, is explained by a model. As a result of that model, it can be said that the effect of age, occupation and present health condition is reasonable. After analysing that model, the relationship between categorical variables (age, sex, occupation, preferred precautions, and worth of personal health) is shown graphically by multiple correspondence analysis.

Keywords

Logistic Regression, Correspondence Analysis, Self Healing

1. Introduction

One of the oldest instincts which the human being has is protection, especially against diseases. Protection is one of the basic parts of self healing process. Self healing is an extensive process which comprises many methods from non medical mixtures or applications to homemade medicines. Self healing process can be defined as curing oneself own medical problems without any professional support [1]. The way—having a present medicine using way in the refrigerator or suggested by a friend—used in self healing process is not chosen randomly.

How to cite this paper: Greenacre, Z.A., Terlemez, L. and Sentürk, S. (2014) Usage as Complementary Correspondence Analysis and Logistic Regression in a Scientific Survey on Self Healing Methods. *Open Journal of Statistics*, 4, 912-920.

<http://dx.doi.org/10.4236/ojs.2014.411086>

There are many sociological variables which affect self healing process. And also they are worth to be subject of a survey. Because they can be used to guide people who face medical problems in a survey.

As in the many fields, especially medical science, it generally came across categorical variables. There are lots of techniques developed for multiple variable techniques for categorical data. Up to analysis type, those techniques may vary. It is sometimes difficult to selection variable for explaining relationship in self healing process. The multiple variable techniques are used as complementary in self healing survey to solve this problem.

The contribution of this study, complementary usage of multi-variable techniques may be at issue in the same research. For example, in some research, variables can be modeled first and then may have visual presentation. In some cases, there may be too many independent variables that explain dependent variable. In such researches firstly, the number of the variables is reduced and research carried on searching variables.

Logistic regression which is a technique for analysis of categorical variables allows us to classify variables as dependent and independent and models the relationship among the variables. Correspondence analysis is also another technique which we may use to handle categorical data and by which it is possible to analyze two or more categorical variable in a single step and to display the relationship graphically.

In this study, representation of complementary usage of which are widely used multivariate techniques, logistic regression and correspondence analysis in a research subject to self healing methodologies is aimed. Again in the study, the number of categorical variables which are considered as related with each other, is reduced and modeled using logistic regression, and then relationship between the related variables is presented with correspondence analysis graphically.

Logistic regression is summarized and correspondence analysis is summarized in followings. Factors about self healing subject in the application are first analyzed by logistic regression and correspondence analysis in the last section.

2. Logistic Regression Analysis

The use of logistic regression modeling has exploded during the past decade. Although it is firmly established within epidemiology research, the method is now commonly employed in many fields including but not nearly limited to biomedical research, business and finance, criminology, ecology, engineering, health policy, linguistic and wildlife biology [2] [3]. Logistic regression is part of a category of statistical models called generalized linear models. This broad class of models includes ordinary regression and ANOVA, as well as multivariate statistics such as ANCOVA and log-linear regression. An excellent treatment of generalized linear models is presented in Agresti (1996) [4].

Logistic regression is used to explain the relationship between the dependent variable and the independent variables, when the dependent variable is observed into two or more categories. The effects of independent variables the dependent variable is defined as probabilities.

Logistic regression's purpose is to estimate parameters by creating logistic models. It is also possible to add common variables to the models and so corrected Y estimates may be obtained according to common variables. Logistic regression is a statistical technique that calculates the estimates of dependent variable and classifies it by using probability rules. Row data sets or data in tables may be analyzed by this method [4]-[7]. Regression model in the logistic regression as follows:

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}} \quad (1)$$

$$\beta_0, \beta_1, \dots, \beta_p \text{ regression coefficient} \quad (2)$$

There are three main methods in logistic regression analysis: *Binary Logistic Regression Analysis*: It is used for dependent variables that have binary values. *Ordinal Logistic Regression Analysis*: It is used for dependent variable which is ordinal. The observed values must be at least in three categories. *Nominal Logistic Regression Analysis*: This method is suitable when dependent variable is nominal. The observed values must be at least in three categories [8] [9].

3. Correspondence Analysis

Correspondence analysis can be defined as the combination of mathematical and graphical techniques used for

explanation of a contingency table [9]. In this technique, single value separation is used to analyze the contingency tables composed of data multi terms of non negative values. The relationship between the rows and columns of the contingency table is shown by a graphic with fewer dimensions. It is also possible to show each row and column as a point in the Euclid space.

For two-way contingency tables, simple correspondence analysis is used. For more than two-way contingency tables, multiple correspondence analysis applied [10].

Row and column marginal vectors can be written $r = pe_c$, $c = p'e_r$ where $e_c (c \times 1)$ and $e_r (r \times 1)$ are vectors of unities. The vectors r and c are also referred to respectively as row and column masses. Diagonal matrices constructed from the row and column masses are denoted by $D_r (r \times r)$ and $D_c (c \times c)$ respectively [9].

For two dimensional contingency tables matrices of row and column profiles is shown by

$$R = D_r^{-1}P, \quad C = D_c^{-1}P \tag{3}$$

[11]. In correspondence analysis, the distance between the points of the data in the multi-dimensional space can be defined. In Euclidean space, the distance between two points is defined by Euclidean distance. The distance can be computed by a function known as χ^2 distance. They are expressed with related to the row and column profiles. χ^2 distance between the i th and i' th row profiles is shown by

$$d_c(i, i')^2 = \|a_i - a_{i'}\|_c^2 = \sum_{j=1}^c \frac{(n_{ij}/n_i - n_{ij}/n_{i'})^2}{n_j/n} \tag{4}$$

[12]. For matrix $(P - rc')$, the generalized singular value decomposition subject to the conditions $A'D_r^{-1}A = I$, $B'D_c^{-1}B = I$ is given by

$$(P - rc') = AD_\mu B' = \sum_{k=1}^K \mu_k a_k b'_k \tag{5}$$

where the columns of $A (r \times k)$ and $B (c \times k)$ are denoted by a_k and b_k respectively. $\mu_1, \mu_2, \dots, \mu_k$ are the diagonal elements of the diagonal matrix $D_\mu (K \times K)$. The dimension K is taken $\min[(r-1), (c-1)]$.

The vectors a_k , $k = 1, 2, 3, \dots, K$ are called the principals axes of the columns of $(P - rc')$; the vectors b_k , $k = 1, 2, 3, \dots, K$ are called the principals axes of the rows of $(k = 1, 2, 3, \dots, K)$. Diagonal elements $\mu_1, \mu_2, \dots, \mu_k$ of D_μ are called the singular values of $(P - rc')$.

The total inertia can therefore be written as

$$\text{in}(r) = \text{in}(c) = \sum_{i=1}^r \sum_{j=1}^c \frac{(P_{ij} - r_i c_j)^2}{r_i c_j} = \chi^2/n \tag{6}$$

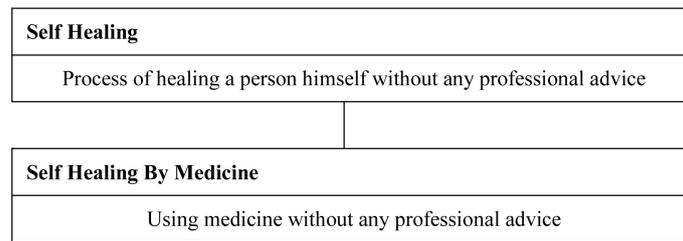
which is the sum of the squares of singular values [9].

Multiple correspondence analysis tackles the more general problem of associations among a set of more than two categorical variables [12]. Suppose the original matrix of categorical data is $N \times Q$, i.e. N cases and Q variables. Classical multiple correspondence analysis has two forms. The first form converts the cases-by-variables data to an indicator matrix Z where the categorical data have been recoded as dummy variables. The second form of multiple correspondence analysis calculates the Burt matrix $B = Z^T Z$ of all two-way cross-tabulations of the Q variables [13].

4. Application

The study about self healing which is figured out in **Figure 1** is planned since it is a vital matter for community health. In order to determine such methods used for public and personal health, a public survey was carried out on 750 individuals (age 15 - 56) who live in the province of Eskisehir in Turkey [13]. The priority of the alternatives, for example, visiting a pharmacy, applying a particular method or visiting a doctor was inquired.

First of all, frequencies about the individuals' demographic characteristics are shown in **Figure 1**. According to **Table 1**, 373 of individuals are male and 377 of individuals are female. 75.5% of individuals agglomerated in the first three age categories 15 - 25, 25 - 35 and 36 - 45 as 33.3%, 22.7% and 19.5% respectively. Also 94.1%

**Figure 1.** Self healing process [1].**Table 1.** Frequencies of individuals' demographic characteristics.

Demographic Variables	Categories	Individual Number	% (Percentage)
Sex	Male	373	49.7
	Female	377	50.3
Age	15 - 25	250	33.3
	26 - 35	170	22.7
	36 - 45	146	19.5
	46 - 55	97	12.9
	56+	87	11.6
Marital Status	Single	378	50.4
	Married	328	43.7
	Divorced	9	1.2
	Widow	35	4.7
Education Level	Literate Without a Diploma	13	1.7
	Primary School	79	10.5
	Junior-High School	421	56.1
	University and master, doctorate	237	31.6
Occupation	Student	100	13.3
	Housewife-Non Employed	79	10.5
	Officer	150	20.0
	Employee	155	20.7
	Tradesmen	77	10.3
	Retired	80	10.7
	Self-Employed Person	89	11.9
Other	20	2.7	

of the individuals agglomerated in the first two categories of the marital status as 50.4% single and 43.7% married. In the education level, this agglomeration showed itself as 56.1% Junior-High School and 31.6% university and master, doctorate. In occupation, this agglomeration differed from age, marital status and education level. Most of the agglomeration in occupation is in student, officer and employee categories as 20.7% employee, 20.0% officer and 13.3% student respectively.

Table 2 shows frequencies of individuals' responses about own health. There are two answers of individuals belong to two questions. They care for attaching importance to own health with 45.1% and physical health with 79.5 %.

Table 2. Frequencies of individuals' responses about own health.

	Categories	Individual Number	% (Percentage)
Attaching importance to own health	Much care	241	32.1
	Care	338	45.1
	Less care	130	17.3
	Do not care	30	4.0
	Never care	11	1.5
Caring Physical Health	Yes	596	79.5
	No	154	20.5

Table 3 shows frequencies of individuals' responses about physical health and precautions. They take mainly well nutrition for their physical health as precautions. The frequency table related to first method applied by 750 individuals in the case of an illness is illustrated in **Table 4**. According to table, the first method preferred by people is to rest by sleeping.

In this study, it has been used logistic regression and a complementary study, correspondence analysis, to show the relationship graphically. The relationship among the behaviors of individuals and their demographic characteristics was modeled by using logistic regression and correspondence analysis is applied as a complementary study.

In application, first of all, the effect of age, sex, marital status, education level, occupation and income level and present health condition, on appreciating self health, is explained by a model. After analyzing that model, the relationship between categorical variables (age, gender, marital status, education level, occupation, taking precaution, preferred precautions, and worth of personal health) is shown graphically by multiple correspondence analysis in the section followings.

4.1. Whether One Takes Precautions to Keep the Physical Health—Demographic Variables—Degree of Caring Physical Health

The relationship among the variables, which explain whether one takes precautions, gender, age, marital status, educational status, profession, income level and degree of caring physical health was examined by binary logistic regression since the question which supply the data is a yes - no question. One can find the results of the analysis in **Table 5**.

As a result of this analysis, age, marital status and the attaching importance to own health are obtained as significant variables. Then we go ahead by multiple correspondence analysis which enables us to display the graphical representation of the variables.

According to **Figure 2**, widowed individuals care less their physical health or even do not care. Married people in age groups 26 - 35 and 46 - 55 care physical health very much and take precautions. There isn't any significant data to say that people in 15 - 25 age group are caring or nor caring physical health.

4.2. Type of Precaution for Preserving Physical Health—Demographic Variables—Degree of Caring Physical Health

The relationship among the precaution types taken by 596 individuals to for keeping physical health, demographical variables and degree of caring physical health by logistic regression technique. Multi nominal logistic regression is applicable in this situation since the dependent variable is the type of the precaution (use medicine, well nutrition, high value nutrition, abstaining from detrimental materials, abstaining from bothering cases, other). The result of the analysis is seen in **Table 6**.

Age, educational status and profession are enough to explain the variation of precaution type. If multiple correspondence analysis with the chosen variables is applied, the graphic in **Figure 3** can be obtained. All those variables then can be analyzed in depth by correspondence analysis. As a result it can be said that officials and high-educated people (graduate and high) are in the age group 26 - 35 other group prefer the job from the other

Table 3. Frequency table about physical health and precautions.

Taking Precaution	%
Well Nutrition	31.4
Using Protective Medicine	25.2
To Abstain From Detrimental Materials	21.3
Having High Potent Nutrient	13.9
To Get Out Of Detrimental Conditions	6.7
Other	1.5

Table 4. Frequency table related to the frequency of first method preferred by people in aces of illness.

First Applied Method	%
Resting and Sleeping	24.4
Having A Kind Of Medicine Present At Home	23.5
Applying To A Doctor or Dentist	22.9
Having Home-made Natural Medicine	12.3
No Precaution	8.5
Applying To A Pharmacist	2.5
Having A Suggested Medicine	1.5

Table 5. Results of binary logistic regression analysis.

Variable	B	Standard Error	Wald	df	Sig.	Exp (B)
Constant	-3.212	0.325	97.560	1	0.000	0.040
Age	-0.326	0.111	8.599	1	0.003	0.722
Marital	0.456	0.165	7.600	1	0.006	1.578
Attaching Importance to Own Health	0.840	0.107	61.576	1	0.000	2.316

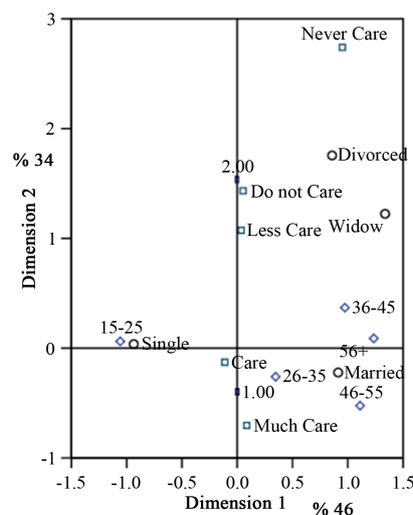


Figure 2. Correspondence analysis graphic among age-marital status—attaching importance to own health—caring physical health.

Table 6. Results of multi-nominal logistic regression analysis.

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	1214.520	0.000	0	
Sex	1222.999	8.478	5	0.132
Age	1246.045	31.525	20	0.049
Marital	1235.744	21.223	15	0.130
Education	1243.165	28.644	15	0.018
Occupation	1269.746	55.225	35	0.016
Income	1250.036	35.516	30	0.224
Attaching Importance to Own Health	1240.500	25.980	20	0.166

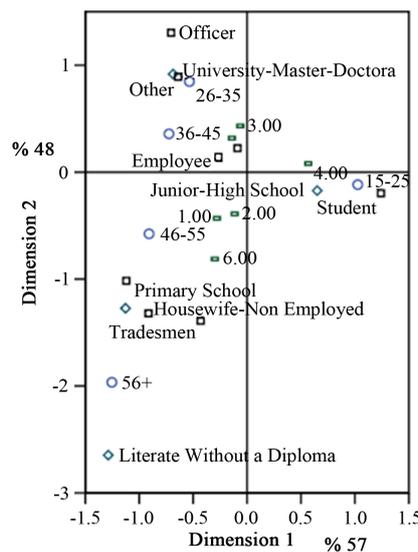


Figure 3. Graphic of correspondence analysis for age-education-occupation-type of precaution.

section. Workers of 36 - 45 age group are keeping away from and high nourished but it is not a matter for college students. People in 46 - 55, are having protective medicine, being well nourished and taking other precautions.

4.3. Action Preferred in Case of Illness—Demographic Variables and Degree of Caring Self Health

The relationship among the variables, which explain the methods preferred by people in case of illness, gender, age, marital status, educational status, profession, income level and degree of caring physical health, was examined by logistic regression. Multi nominal logistic regression is applicable in this situation since the dependent variable is the type of preferred method (No Precaution - Having Home-Made Natural Medicine - Having a Kind of Medicine Present at Home - Having a Suggested Medicine - Exercising - Resting and Sleeping - Applying to a Pharmacist - Applying to a Doctor or Dentist).

According to **Table 7**, it can be seen that gender, occupation and health status are enough to explain the first method preferred in case of an illness. After applying logistic regression, multiple correspondence analysis can be applied.

If multiple correspondence analysis is applied, the graph in **Figure 4** may be obtained. Up to the graph, officials are using homemade natural medicine. Workers prefers going to doctor, exercising and caring physical

Table 7. Results of multi-nominal logistic regression analysis.

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	1748.516	0.000	0	
Age	1783.669	35.153	28	0.165
Gender	1765.399	16.883	7	0.018
Marital	1780.007	31.491	21	0.066
Education	1765.540	17.023	21	0.710
Occupation	1823.466	74.950	49	0.010
Income	1780.691	32.175	35	0.605
Attaching Importance to Own Health	1831.510	82.994	28	0.000

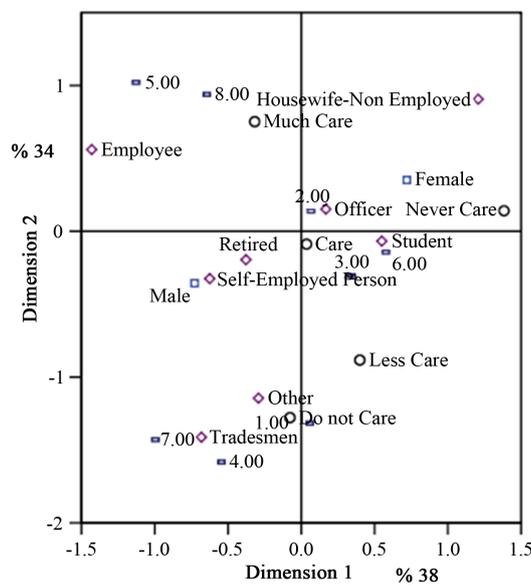


Figure 4. Correspondence analysis for the relation among gender-occupation-attaching importance to own health and preferred method.

health up to graph. People who are working in the other job groups are not caring physical health and do nothing in case of an illness. Tradesmen are having a suggested medicine or applying to a pharmacist while students are resting or having a medicine which is already at home in case of an illness.

5. Conclusions

In many field, techniques for categorical variables are applied as a complementary study. In the study, correspondence analysis and logistic regression analysis are applied at the same research. As a first step, the variables which are explaining the variation of dependent variable are determined by logistic regression and then correspondence analysis applied to show the relation of variables in the model, graphically. These techniques are used in this way because while there are too many explanatory variables explaining dependent variable, it causes complexity. In this situation, by using logistic regression best explanatory variables that describe dependent variable are selected and are modeled. Finally, by using correspondence analysis, this relationship is graphically shown. The relationship is analyzed both correspondence analysis and logistic regression which is a technique that allow us to categorize the variables as independent and dependent variables. By this method, using both methods simultaneously, we got more detailed results for relationship. Thus, in this study we try to show to use correspondence analysis and logistic regression together in variable selection. A survey aimed at usage of

self-healing methodologies is utilized to show how to use these techniques together.

The data of an inquiry, which was applied to 750 people health healing process, is analyzed in the study. First, whether taking precaution, type of precaution and type of method applied in case of an illness are determined as dependent variables. It is searched whether they are explained by demographic variables and degree of caring health. After that, determined variables and the other variables are analyzed by multiple correspondence analysis. It is seen that age is significant for dependent variables. It can be said that gender, marital status, education level and occupation are also significant. In addition, degree of caring physical health is considerable on dependent variables while income level is not significant.

References

- [1] Karaca, A.R. (1994) "Kendi Kendini Tedavi Reçetesiz İlaçve OTC. Turkish Pharmacists's Association News (TEB Haberler), No: 10, pp. 23-26.
- [2] Hosmer, David, W. and Lemeshow, S. (2000) Applied Logistic Regression. John Wiley & Sons, Inc., Hoboken.
- [3] O'Connell, A.A. (2007) Logistic Regression models for Ordinal Response Variables. Sage Publication, Inc., Thousand Oaks.
- [4] Agresti, A. (1996) An Introduction to Categorical Data Analysis. John Wiley and Sons, New York.
- [5] Agresti, A. (1990) Categorical Data Analysis. John Wiley and Sons, New York.
- [6] Andersen, E.B. (1990) The Statistical Analysis of Categorical Data. Springer-Verlag, Berlin.
<http://dx.doi.org/10.1007/978-3-642-97225-6>
- [7] Johnson, A.R. and Wichern, D.W. (1998) Applied Multivariate Statistical Analysis. Prentice-Hall.
- [8] Özdamar, K. (2004) Paket Programlar ile İstatistiksel Veri Analizi. Anadolu Üniversitesi Yayınları, Eskisehir.
- [9] Greenacre, M.J. and Hastie, T. (1987) The Geometric Interpretation of Correspondence Analysis. *American Statistical Association*, **82**, 437-447. <http://dx.doi.org/10.1080/01621459.1987.10478446>
- [10] Greenacre, M.J. (1984) Theory and Applications of Correspondence Analysis. Academic Press, London.
- [11] Greenacre, M.J. and Blasius, J. (2006) Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC, New York.
- [12] Greenacre, M.J. (2007) Correspondence Analysis in Practice. 2nd Edition, Chapman & Hall/CRC, New York.
- [13] Güler, E. and Aşan, Z. (2004) Kendi Kendine Tedavive Kişisel Sağlığın Önemi. *Meslekiçi Sürekli Eğitim Dergisi*, No: 7-8, pp. 80-91.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or [Online Submission Portal](#).

