International Journal of Community Medicine and Public Health
Dey A et al. Int J Community Med Public Health. 2016 Dec;3(12):3423-3429
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: http://dx.doi.org/10.18203/2394-6040.ijcmph20164268

Original Research Article

Application of latent class analysis to estimate susceptibility to adverse health outcomes based on several risk factors

Ankita Dey*, Arun K. Chakraborty, Kunal K. Majumdar, Asok K. Mandal

Department of Community Medicine, KPC Medical College & Hospital, Jadavpur, Kolkata, West Bengal, India

Received: 10 October 2016
Accepted: 26 October 2016

***Correspondence:**
Ms. Ankita Dey,
E-mail: ankitadey14@gmail.com

ABSTRACT

Background: The study demonstrates the use of latent class analysis (LCA) to segregate population in two latent classes e.g. susceptible or not susceptible to adverse health outcomes according to the observed risk factors as a method of medical diagnosis.
Methods: The present study uses a secondary data set on 420 patients referred to the University of California, Los Angeles (UCLA). Adult Cardiac Imaging and Hemodynamics Laboratories for Dobutamine stress echocardiography (DSE) between March1991 & March1996. LCA is used for estimating the individual item-response probabilities in each latent group and also the latent class sizes. The observed variables or indicators of the latent subgroups are the common risk factors viz. history of smoking, history of cardiac issues etc. The interaction effect of hypertension & diabetes is also included in the analysis.
Results: Based on the behaviour of the estimates of latent class model parameters, the unobserved groups are identified and named. Proportion of individuals falling in each latent class are approximately 0.20 & 0.80 respectively. The susceptibility to adverse health outcomes in future is the most in male individuals having a positive history of hypertension and/or diabetes, as the corresponding indicators have higher positive item-response probabilities (0.72 & 0.83 respectively) than the rest.
Conclusions: The study briefly explains the application of LCA for identifying subgroups according to susceptibility to adverse health effects in a large population. Assessment of common risk factors in predicting latent class sizes provides estimates of probabilities for being a member in each class. The importance of the combined effect of hypertension & diabetes in predicting future health problems related to cardiac issues is highlighted. Class assignments of individuals according to their pattern of response are also listed.

Keywords: Indicators, Interaction effect, Item-response probabilities, Latent class analysis, Medical diagnosis, Risk factors

INTRODUCTION

The concept of latent variable has emerged as an important tool for statisticians as well as social scientists. A latent variable may be defined as a random variable whose realizations are unseen or hidden from us.[1] Whereas the realizations of observed or manifest variables are observable in the study. A broad categorization of latent variable modelling is given by Biemer.[2] Latent class analysis(LCA) is considered to be an equivalent methodology for Factor Analysis, typically used for dichotomous or polytomous variables.[3-5] The parameters of interest in a typical problem of latent class analysis are the unobserved proportion or size of the latent classes and the conditional item-response probabilities given the membership in a latent class. Based on the observed data on manifest variables LCA provides a classification among the population. Manifest

variables are called the 'indicators' of a particular latent class.

The type of latent variable modelling in which both the latent and observed variables are categorical in nature is called LCA. The basis of the analysis lies in modelling the relationship between the latent variable and its indicators. LCA identifies unobservable (latent) subgroups within a population based on individuals' responses to different categorical observed variables. It is a technique which explains the relationships between manifest or observed variables (may be dichotomous or polytomous) with respect to some unobserved or latent variables (may be dichotomous or polytomous) on the basis of data obtained in various kinds of complex surveys. Given the response pattern of a respondent to various questions or items of a questionnaire, LCA performs grouping of individuals in two or more latent sub-populations. Latent class analysis was introduced in 1950 by Lazarsfeld PF, who performed LCA for building typologies (or clustering) based on dichotomous observed variables.[6]

The estimation of model parameters using maximum likelihood approach was proposed by Goodman.[7] Haberman obtained the relation between latent class models and log-linear models for frequency tables with missing (unknown) cell counts.[8]

Contributions of Vermunt, Hagenaars in formulating a general framework for the analysis encouraged further research in the area.[9,10] Many applications of LCA and theoretical developments have been proposed in the last two decades.

Latent class analysis encompasses a wide area of application in many disciplines such as survey research, health sciences, psychology, sociology, education, social sciences etc.[11-15] It provides a tool for clustering and classification of individuals in qualitative research. The identification of latent groups into which the respondents fall on the basis of their response pattern and estimating the class membership for each individual are the underlying objective of latent class analysis. LCA can be used in health research for identifying disease subtypes or diagnostic categories. LCA has been used for building typologies in medical conditions. Dunn et al.[16] categorized the low back pain through LCA and obtained four clusters viz. "persistent mild", "recovering", "severe chronic" and "fluctuating" representing different situations of back pain. Diagnosis of disease is based on the traditional method of assessment of the value of diagnostic indicators such as medical signs, symptoms and medical tests. To achieve this objective, the evaluation of sensitivity and specificity i.e. the probability that a person is positive on the indicator when the disease is present and the probability that a person is negative on the indicator when the disease is absent respectively is the common practice. But it requires the knowledge of the presence or absence of the disease. Latent class analysis provides a tool for correct diagnosis without the prior knowledge of whether the patient is suffering from the disease or not. The present work uses data on several risk factors of adverse health effects related to cardiac events to demonstrate the applicability of latent class analysis in establishing two unobserved latent clusters in the population of patients.

METHODS

The study is an analytical record based study, retrospective to a secondary source of data, describing the use of latent class analysis to categorize the individuals according to possibility of having adverse health outcomes in two latent classes based on several common risk factors as a method of medical diagnosis. LCA has been used as a multivariate statistical tool for the fulfilment of the objective. The mathematical model is described in the following section.

The latent class model

There are two types of probabilities in the latent class analysis model. The first type of probability indicates the likelihood of a response by respondents in each of the classes. This represents the probability of a particular response to a manifest variable, conditioned on latent class membership. This can be interpreted as factor loadings for factor analysis in which both the observed or latent variables are measured in a continuous scale. The other type of probability represents the latent class size or the proportion of individuals who are members of a particular latent class. Based on the observed response patterns of individuals (or the observed contingency table) LCA provides a clustering of individuals in a population.

The assumptions of the standard LC model are as follows:[2]

- Data is sampled without replacement from a large population units using simple random sampling.
- The usual latent class model assumes "local independence", i. e. variables are independent within a latent class.
- The response probabilities are same for any two individuals or units selected from the population.
- The indicators are univocal in a sense that they indicate to one and only one latent class.

Following the notations used by AB. Mooijaart[17], suppose there are k = 1,..., K latent classes and v = 1,...,V manifest or observed variables. s(v) represents the category number observed for the observed variable v which can take values from 1,...,I_v (polytomous response options). 's' be a vector of length V defined as (s(1), s(2), ... , s(V)) and it represents a response pattern for a particular individual.

π_s be the probability of outcome vector 's' and $\pi_s = \sum_{k=1}^{K} \pi_{s,k}$, where $\pi_{s,k} = \pi_k \sum_{v=1}^{V} \pi_{v,s(v)|k}$.

π_k is the size of class k and $\pi_{v,s(v)|k}$ is the probability of response s(v) in the v^{th} item (or variable) conditional on class k. Therefore, $\pi_{s,k}$ is the unobserved probability of simultaneously falling in the categories denoted by vector s and the latent class k.

The complete data likelihood can be written as, $L = \prod_{s,k}(\pi_{s,k})^{n_{s,k}}$, where $n_{s,k}$ is the number of individuals(unobserved) in the sample who have response pattern s and fall in class k.

LCA models differ mainly in the number of latent classes. There is always a trade-off between goodness of fit and parsimony of the latent class models. Usually, models with more parameters (i.e. more latent classes) provide a better "fit", and more parsimonious models tend to have a somewhat poorer fit. The best suited model describes the associations among the manifest variables in a more justiciable manner. The goodness-of-fit of an estimated LCA model can be tested with different information criteria. Bayesian Information Criteria (BIC) is commonly used to select the optimal number of latent classes in a model.[18] By comparing models with different number of latent classes, a model is selected with a lower BIC. Model parameters are estimated with suitable estimating algorithms such as Expectation-Maximization algorithm.[19]

Data description

The present study uses a secondary data set on patients referred to the UCLA Adult Cardiac Imaging and Hemodynamics Laboratories for Dobutamine stress echocardiography (DSE) between March 1991 and March 1996, who gave informed consent to have their medical data reviewed.[20] Study was designed to prospectively follow consecutive patients who underwent DSE at the UCLA School of Medicine during a five-year period. Patients who did not have either 12 months of follow-up or an event of cardiac malfunction within 12 months were not included in the final analysis. The dataset consists of 558 observations, of which 138 were deleted for data cleaning purpose. Observations on 420 individuals are considered for the analysis. There are 142 male and 278 female patients. The retrospective study has been conducted in March 2016 in the KPC Medical College and Hospital, Kolkata.

Variables on risk factors related to cardiac events viz. gender, history of hypertension, history of diabetes, history of smoking, history of Myocardial Infarction, history of angioplasty, history of bypass surgery are taken to be the observed or manifest variables for the analysis. All the manifest variables have binary response options (e.g. 1 as Yes and 2 as No, 1 as male and 2 as female). Consequently a two, three, four classes latent class models have been considered. Comparing the BIC values, a 2-class model is chosen. Based on all possible 2^7 response patterns, the population is segregated into two latent classes.

RESULTS

The membership probabilities estimated from the latent class model are summarized in the table 1. The probabilities of positive histories of the risk factors have higher values in latent class 1 compared to the second latent class in most of the cases. Therefore the underlying latent classes can be identified as "Susceptible to adverse health outcomes (Class 1)" and "not susceptible to adverse health outcomes (class 2)". Health issues of individuals having positive history of Myocardial Infarction include death as the final outcome. Proportions of individuals falling under each latent class are mentioned in the parenthesis in Table 1.

Table 1: Item-response probabilities of the observed variables.

Item-response probabilities	Latent class 1 (0.2451)	Latent class 2 (0.7549)
Gender	0.7274	0.2117
History of hypertension	0.6875	0.7072
History of Diabetes	0.3528	0.3680
History of Smoking	0.3901	0.2581
History of myocardial infarction	0.6653	0.1372
History of angioplasty	0.1950	0.0155
History of bypass surgery	0.4255	0.0259

The two indicators namely history of hypertension and history of diabetes are showing almost equal probabilities of positive response in class 1 and in class 2. So, to distinguish between the two classes with respect to these variables, a new variable has been formed by considering the interaction effect of hypertension and diabetes. The new variable is categorised into binary response options as "1" being the presence of at least one disease from hypertension and diabetes in the individual and "2" being the absence of both.

Table 2: Item-response probabilities including the combined history of hypertension and/or diabetes.

Item-response probabilities	Latent class 1 (0.2024)	Latent class 2 (0.7976)
Gender	0.7208	0.2202
History of hypertension and/or diabetes	0.8259	0.7546
History of smoking	0.3935	0.2587
History of myocardial infarction	0.6939	0.1351
History of angioplasty	0.2011	0.0159
History of bypass surgery	0.4310	0.0292

Table 2 shows the membership probabilities for the two latent classes including the new combined variable

instead of the individual variables. Unconditional probabilities of being in class 1 (i.e. Susceptible to adverse health outcomes) and class 2 are about 0.20 and 0.80 respectively. For the six indicators considered in the model are all showing higher probabilities of being a member in the latent class 1 than in the latent class 2. For example, the chance of occurrence of health issues in individuals having positive history of hypertension and/or diabetes is about 0.83 in the first class, which is higher than the same (0.75) in the second class. There are higher probabilities of being more susceptible to adverse health outcomes in individuals having history of smoking, MI, angioplasty and bypass surgery.

Selection of appropriate number of latent classes can be made by comparing the values of BIC, a suitable measure for variable selection. The data-set is best fitted for a 2-class model as the corresponding BIC value is the lowest. The data were inconsistent with several other possible models. Figure 1 shows the values of BIC up to 4-class model fitted to the data.

Table 3 gives the probabilities of membership in the latent classes for each response pattern in the observed data. The frequencies of the several response patterns are also listed in the table. Response patterns with zero observed frequencies are omitted. The first cell of probabilities (0.00332) indicates the probability of membership in the first latent class corresponding to the response pattern of six variables as (1, 1, 1, 1, 1, 1) i.e. male individuals with the positive histories in the five risk indicators.

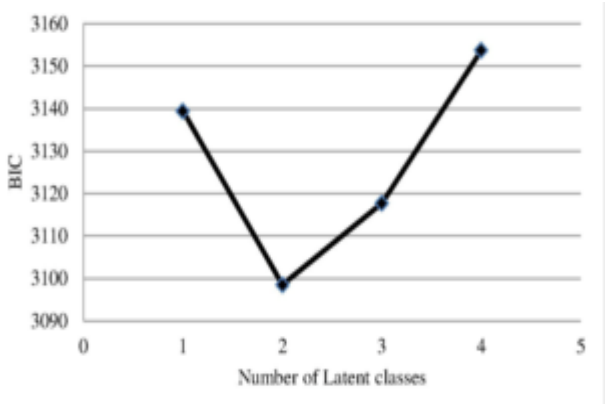

Figure 1: BIC for different number of latent classes.

Table 3: Assignment to latent classes for each pattern of indicators.

Gender	History of hypertension and/or diabetes	History of smoking	History of myocardial infarction	History of angioplasty	History of bypass surgery	Observed frequency	Probability of response pattern
1	1	1	1	1	1	2	0.003319835
1	1	1	1	1	2	1	0.004448083
1	1	1	1	2	1	8	0.013311785
1	1	1	1	2	2	6	0.021644721
1	1	1	2	1	2	3	0.002371786
1	1	1	2	2	1	2	0.006632957
1	1	1	2	2	2	21	0.034837941
1	1	2	1	1	1	2	0.00511891
1	1	2	1	1	2	4	0.006945811
1	1	2	1	2	1	9	0.020683465
1	1	2	1	2	2	15	0.038971966
1	1	2	2	1	1	1	0.002293643
1	1	2	2	2	1	4	0.011302581
1	1	2	2	2	2	26	0.089645638
1	2	1	1	2	1	1	0.002819909
1	2	1	1	2	2	1	0.005046645
1	2	1	2	1	2	1	0.000550125
1	2	1	2	2	2	8	0.010448843
1	2	2	1	1	2	2	0.001486264
1	2	2	1	2	1	2	0.004400613
1	2	2	1	2	2	4	0.009603152
1	2	2	2	2	1	4	0.002649422
1	2	2	2	2	2	15	0.027793744
2	1	1	1	1	2	2	0.001939102
2	1	1	1	2	1	3	0.005557103
2	1	1	1	2	2	13	0.021760502
2	1	1	2	2	1	1	0.005143059
2	1	1	2	2	2	32	0.099157287
2	1	2	1	1	2	1	0.003310241

2	1	2	1	2	1	5	0.009161639
2	1	2	1	2	2	23	0.053421401
2	1	2	2	1	1	1	0.001007616
2	1	2	2	1	2	4	0.005613724
2	1	2	2	2	1	4	0.011752671
2	1	2	2	2	2	131	0.280149037
2	2	1	1	2	2	2	0.006304218
2	2	1	2	2	2	15	0.03189986
2	2	2	1	1	2	1	0.000777293
2	2	2	1	2	2	5	0.016181266
2	2	2	2	2	1	3	0.003424134
2	2	2	2	2	2	32	0.090563486
1	1	1	1	1	1	2	0.003319835
1	1	1	1	1	2	1	0.004448083
1	1	1	1	2	1	8	0.013311785
1	1	1	1	2	2	6	0.021644721
1	1	1	2	1	2	3	0.002371786
1	1	1	2	2	1	2	0.006632957
1	1	1	2	2	2	21	0.034837941
1	1	2	1	1	1	2	0.00511891
1	1	2	1	1	2	4	0.006945811
1	1	2	1	2	1	9	0.020683465
1	1	2	1	2	2	15	0.038971966
1	1	2	2	1	1	1	0.002293643
1	1	2	2	2	1	4	0.011302581
1	1	2	2	2	2	26	0.089645638
1	2	1	1	2	1	1	0.002819909

DISCUSSION

The population of patients are divided into two latent groups according to their susceptibility to occurrence of adverse health outcomes in future based on the common risk factors. The model also provides risk probabilities of the risk factors in each of the latent groups. It may also be noted that the risk of having cardiac issues is increased for the interaction effect of hypertension and diabetes compared to that of individual disjoint effects. Moreover the male individuals having a positive history of hypertension and/or diabetes are more prone to have bad health in future, as the corresponding manifest variables have higher positive item-response probabilities (0.72 and 0.83 respectively) than the rest. The most important process lies in the diagnosis of disease in the individuals. Latent class model provides probabilities for each combination of response options in the manifest variables and those may be used for assignment of individuals in the latent classes. Observed data set contains only 41 out of total possible 2^7 response patterns.

The model describes the pattern of vulnerability among individuals having some history of cardiac issues in possibility of developing any adverse health effects, most commonly death. Individuals who have already incurred MI along with the positive histories on some risk factors of cardiac events are predisposed to occurrence of adverse health outcomes. The survival rate of those individuals will naturally be lower than that of the other group of individuals. Detailed analysis of the subgroups which are obtained by applying LCA may provide a further insight in the issue.

Another commonly used multivariate technique for determination of risk factors is logistic regression which only identifies the significant factors responsible for the outcome and the amount of dependence through the estimates of the regression coefficients. For instance, Panwar et al performed univariate and multivariate logistic regression to identify important cardiovascular risk factors in young patients with coronary heart disease in India in a case-control study.[21] To investigate potential risk factors for sports injury, multivariate statistical approach of logistic regression has been applied for determination of complex interaction of multiple risk factors and events.[22] To examine the association of Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories with coronary heart disease (CHD) risk age-adjusted Cox proportional hazards regression and risk analysis were used to test for the relation between various independent variables and the CHD outcome.[23]

Latent class model is a multivariate approach to categorical data, which provides a classification and discrimination of population in well defined clusters in probabilistic manner. It provides magnitudes of risks through the item-response probabilities in each class. This approach is more elaborate in nature.

The study demonstrates the use of latent class analysis in medical diagnosis. The analysis is based on a secondary data set referred to a period in the past (1991-96) and it is considered to be sufficient for the purpose of the study.[24] However it can be improved by taking more recent data containing more risk factors. The number of latent classes in that case may increase and interpretations can be made accordingly. The estimation of the parameters has been carried out using the 'poLCA' package of the R software.[25]

CONCLUSION

The study briefly explains the application of LCA for identifying subgroups according to susceptibility to adverse health effects in a large population. Assessment of common risk factors in predicting latent class sizes provides estimates of probabilities for being a member in each class. The importance of the combined effect of hypertension & diabetes in predicting future health problems related to cardiac issues is highlighted. Class assignments of individuals according to their pattern of response are also listed.

ACKNOWLEDGEMENTS

Authors acknowledge the contributions of Dr. Diganta Mukherjee, Indian Statistical Institute (ISI), Kolkata and Prof. Sugata Sen Roy, University of Calcutta, Kolkata in providing technical help and encouragement in the preparation of this manuscript. Data have been downloaded from the open source library in the home page of the website of the Department of Statistics, University of California, Los Angeles. The estimation of the parameters has been carried out using the 'poLCA' package of the R software.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Skrondal A, Rabe-Hesketh S. General latent variable modelling-multilevel, longitudinal, and structural equation models. Interdisciplinary statistics 1st ed., Chapman & Hall 2002; Boca Raton, Fla. ISBN 1584880007.
2. Biemer PP. Latent class analysis of survey error. A John Wiley and Sons, Inc. Publication; 2011.
3. Collins LM, Lanza ST. Latent class and latent transition analysis with applications in the social, behavioural and health sciences. A John Wiley and Sons, Inc. Publication; 2010.
4. Formann AK. Constrained latent class models: theory and applications. Br J Mathematical and Statistical Psychology. 1984;38: 87-111.
5. Formann AK. Linear logistic latent class analysis for polytomous data. J Am Statistical Association. 1992;87:418.
6. Lazarsfeld PF. The logical and mathematical foundations of latent structure analysis. S. A. Stouffer Ed.1950a; Measurement and Prediction, Volume IV of The American Soldier: Studies in Social Psychology in World War II, Princeton University Press: 362-472.
7. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika. 1974;61(2):215.
8. Haberman SJ. Analysis of Qualitative Data. New York: Academic Press 1979; Vol. 2: New Developments.
9. Vermunt JK. Log-linear models for event histories. Thousand Oaks, CA: Sage 1997.
10. Hagenaars J. Categorical longitudinal data: Log-linear panel, trend, and cohort analysis. Newbury Park, CA: Sage 1990.
11. Patterson BH, Dayton CM, Graubard BI. Latent class analysis of complex sample survey data. J Am Statistical Association. 2002;97:459 721-41.
12. Castle DJ, Sham PC, Wessely S, Murray RM. The sub-typing of schizophrenia in men and women: a latent class analysis. Psychological Medicine. 1994;24:41-51.
13. Rees KV, Vermunt J, Verboord M. Cultural classifications under discussion Latent class analysis of highbrow and lowbrow reading. Poetics. 1999;26:349-65.
14. Lin SW, Tai WC. Latent Class Analysis of Students' Mathematics Learning Strategies and the Relationship between Learning Strategy and Mathematical Literacy. Universal J Educational Research. 2015;3(6):390-5.
15. Bhatnagar A, Ghose S. A latent class segmentation analysis of e-shoppers. J Business Research. 2004;57:758-67.
16. Dunn KM, Jordan K, Croft PR. Characterizing the course of low back pain: a latent class analysis. American Journal of Epidemiology 2006;163:8.
17. Mooijaart AB. The EM algorithm for latent class analysis with equality constraints. Psychometrika. 1992;57(2):261-9.
18. Schwarz G. Estimating the dimension of a model. Ann Statistics. 1978;6(2):461-4.
19. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Statistical Society Series B (Methodological). 1977;39(1):1-38.
20. Krivokapich J, Child JS, Walter DO, Garfinkel A. Prognostic value of Dobutamine stress echocardiography in predicting cardiac events in patients with known or suspected coronary artery disease. J American College of Cardiology. 1999;33:3.
21. Panwar R, Gupta R, Gupta BK, Raja S, Vaishnav J, Khatri M, et al. Atherothrombotic risk factors and premature coronary heart disease in India: A case-control study. Indian J Med Res. 2011;134:26-32.

22. Bahr R, Holme I. Risk factors for sports injuries—a methodological approach. Br J Sports Med. 2003;37:384-392.
23. Wilson PF, D'Agostino R, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation-Journal of the American Heart Association. 1998;97:1837-47.
24. University of California, Los Angeles (UCLA), Department of Statistics [homepage on the internet]. Available from: http://www.stat.ucla.edu/ projects/ datasets/. Accessed 9 March 2016.
25. Linzer DA, Lewis JB. poLCA: An R package for polytomous variable latent class analysis. J Statistical Software. 2011;42:10.

Cite this article as: Dey A, Chakraborty AK, Majumdar KK, Mandal AK. Application of latent class analysis to estimate susceptibility to adverse health outcomes based on several risk factors. Int J Community Med Public Health 2016;3:3423-9.