Available online at www.jpsscientificpublications.com
Volume – 2; Issue - 4; Year – 2016; Page: 681 – 686
DOI: 10.21276.iajmr.2016.2.4.5

Indo – Asian Journal of Multidisciplinary Research (IAJMR)
ISSN: 2454-1370

A REVIEW ON APPLICATIONS OF LATENT CLASS ANALYSIS

Ankita Dey*

Assistant Professor and Statistician, Department of Community Medicine, KPC Medical College and Hospital, Kolkata, West Bengal, India.

Abstract

Latent class analysis (LCA) is considered to be an equivalent methodology for Factor Analysis, typically used for dichotomous or polytomous variables. 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 present paper provides a review on the theory of latent class analysis and its wide area of applications in various disciplines such as Survey research, Medical diagnostics, Economics, Education, Psychology, Sociology, etc.

Key words: Latent Class Analysis, Item-response probabilities and Indicators.

1. 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 (Skrondal, 2004). Whereas, the realizations of observed or manifest variables are observable in the study. Relevant methodologies demonstrating various uses of latent variable in multi-disciplinary research can be seen in the literature. The analysis regarding building models on latent variables varies over the nature of study, research questions and type of data. Modelling the latent truth in different phenomena leads to different types of latent variable modelling. A broad categorization of latent variable modelling is given by Biemer (2011). Factor analysis is used when both the latent variable and indicator variables are continuous. Latent trait analysis is suitable in case the categorical manifest variables are indicators of one or more continuous latent variables. If a discrete number of subgroups determine the categories of a discrete latent variable whose classes are characterized by manifest variables measured on a continuous scale, the corresponding analysis is called latent profile analysis. LCA is appropriate when the latent variable and all the indicator variables are discrete. Factor analysis is a special kind of latent variable modelling where the latent and the indicator variables are continuous. Collins and Lanza (2010) have provided an analytical comparison between LCA and factor analysis. Latent trait analysis is useful if the observed categorical manifest variables are indicators of one or more continuous latent variables. Item response theory, a special case in such kind of modelling, the underlying trait is latent (eg. ability of a student). Latent profile analysis provides the categorization of latent variable into discrete latent classes whose indicator variables are measured on a continuous

Corresponding author*: **Ankita Dey
E. mail: ankitadey14@gmail.com
Received: 22.05.2016; *Revised*: 05.06.2016;
Accepted: 28.06.2016.

scale. There are several other latent variable models based on research design with a substantive theory. In this paper an overview of the particular type of latent variable modelling viz. latent class analysis is presented with the applications in various disciplines.

2. Latent Class Analysis

The type of latent variable modelling in which both the latent and observed variables are categorical in nature is called Latent Class Analysis (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 individual's 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. In addition, LCA will provide estimates of the probabilities of respondents being misclassified by the question. Latent class analysis was introduced in 1950 by Paul F. Lazarsfeld, who performed LCA for building typologies (or clustering) based on dichotomous observed variables. To study military personnel during World War II, the War Department assigned multidisciplinary teams including statisticians. The four volumes of The American Soldier: Studies in Social Psychology in WW II, (Stouffer, 1949 - 1950) have these case studies documented. Lazarsfeld constructed the theoretical aspects and applications of LCA in two chapters published in the Volume 4 of The American Soldier, entitled Measurement and Prediction in 1950. His work in the area of latent structure modelling continued further. The estimation of model parameters using maximum likelihood approach was proposed by Goodman (1974). Haberman (1979) obtained the relation between latent class models and log-linear models for frequency tables with missing (unknown) cell counts. Construction of linear logistic latent class analysis (Formann 1984, 1992) for dichotomous and polytomous data was a further addition to the literature.

Contributions of Vermunt (1997), Hagenaars (1990) in formulating a general framework for the analysis encouraged further research in the area. Many applications of LCA and theoretical developments have been proposed in the last two decades. A closely related field to LCA is the field of finite mixture modelling which deals with continuous outcome variables. The underlying idea in both the cases is that the population contains some inherent groups based on the parameters of the statistical model of interest. One of the application areas of LCA lies in density estimation in which a complicated density can be approximated as a finite mixture of simpler densities.

2.1. 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:

- 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 Mooijaart (1992), suppose there are k = 1,...,K la-tent 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 = (probability of outcome vector 's') = $\sum_{k=1(1)K} \pi_{s,k}$

where $\pi_{s,k} = \pi_k \sum_{v=1(1)V} \pi_{v,s(v)|k}$

π_k is the size of the 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.

Under the assumption of a multinomial distribution, the kernel of 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.

Latent class analysis not only builds a classification model but a relation of the class membership to explanatory variables may also be an issue. Inclusion of covariates (Vermunt, 2010) in the model is an extension towards this goal. In this extension, latent class membership probabilities are predicted by covariates through a logistic link.

2.2. Parameter estimation

Dempster *et al*. (1977) provided maximum likelihood estimation in case of observed and missing data involved in the analysis. In LCA, observed data corresponds to the observed responses of individuals on each item which are considered as categorical observed variables and the missing data are the unobserved scores in a latent categorical variable whose categories are called the latent classes. The method consists of two steps: the expectation (E) and the maximization (M) steps. In the E-step, expectations of the unobserved complete data, i.e data for manifest variables and the latent variable (individual classification of to latent classes as well as their responses to items), conditional on the observed incomplete data on manifest variables and the current estimates of the parameters of interest are required. In the M-step, the provisionally complete log-likelihood obtained in the E-step is maximized as a function of the unknown model parameters latent class sizes and the item-conditional probabilities. Mooijaart (1992) presented the application of EM algorithm in latent class analysis in a very elaborative approach.

2.3. Model selection

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 justifiable manner. The goodness-of-fit of an estimated LCA model can be tested by the Pearson or the Likelihood-Ratio Chi-Square (asymptotically, it is distributed as a chi-square with degrees of freedom (df) equal to the number of cells in the frequency tables minus one, then minus the number of independent parameters). Different information criteria such as Bayesian Information Criteria (BIC) and Akaike Information Criteria (AIC) are commonly used to select the optimal number of latent classes in a model. By comparing models with different number of latent classes, a model selected with a lower AIC and BIC. Residuals also can be considered as an indicator of good or bad fit.

3. Applications of LCA

Latent class analysis encompasses a wide area of application in many disciplines such as Survey research, Health Sciences, Psychology, Sociology, Education, Social Sciences etc. 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*. (2006) 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. In sociological research, Rees *et al*. (1999) used LCA to identify three clearly distinguishable categories *viz*., non-readers, high readers, and low readers, in 3000 respondents (from 1990 Dutch Time Budget Survey) on the basis of a cultural activity "reading in leisure time". Sub-typing of Schizophrenia *viz*., "neuro developmental", "paranoid" and "schizoaffective" in men and women based on the presence or absence of some manifest variables such as Persecutory delusion, family history, early onset etc was identified using latent class analysis by Castle *et al*. (1994). The use of LCA in Education has already been a new focus area for the researchers. Lin and Tai (2015) performed latent class analysis to classify students mathematics learning strategies based on the Taiwan data from the Programme for International Student Assessment (PISA) 2012 and obtained the optimal four-class model corresponding to the classes termed as the memorization strategy, multiple strategies, control strategy and elaboration and control strategies.

3.1. Dietary Survey

The method of LCA has an application in dietary or food consumptions surveys in estimating the consumption of a particular nutrient or food among individuals. Patterson *et al*. (2002) used LCA to classify individuals according to regularity of consuming vegetables. The study is based on the data obtained from Continuing Survey of Food Intakes by Individuals (CSFII) held in 1985. The survey was conducted on women aged 19 - 50 living in private households in the 48 conterminous states which were further divided into 60 relatively homogeneous. From each stratum 2 primary sampling units (PSUs) were sampled. Six dietary recall interviews were conducted on each individual; first was collected in a face-to-face interview and the rest were telephonic. Recalls corresponding to previous 24 hours were collected at about 2 month intervals. The public-use data set contains those women who have participated in the face-to-face interview and completed at least three phone recall interviews. A particular food intake of an individual was recorded in the data set in grams (no intake means 0 gram). To dichotomize the intake variables, a respondent was assigned a value 1 if she consumed that particular food and 0 otherwise. Thus, individual responses on same food were recorded for 4 recall interviews. Based on these binary responses the individuals were categorised into the classes of latent variable "regularity of vegetable consumption". The parameters of the model here represent the proportions in "regular" and "non-regular" vegetable intake categories and the conditional probability of consuming at least one vegetable on the corresponding recall day given the membership in a particular latent class. So, LCA groups respondents into categories, or classes which define patterns of food consumption and provides estimates of class size. About eighteen percent of the population of women aged 1950 consumed a diet deficient in vegetables.

3.2. Medical Diagnostics

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. Rindskopf and Rindskopf (1986) demonstrated this use of LCA in a data set related to myocardial infarction (MI). Manifest variables used in the study were presence or absence of a positive new Q-wave in the ECG, presence or absence of a classical clinical history and two blood tests *viz.*, presence or absence of high CPK-MB and whether or not the patient had flipped LDH. These four observed variables have binary response options. A test of independence among the four observed variables indicated the presence of relationship among them which ensured that there are more than one latent class. In selection of the appropriate model (number of latent class), Chi-square goodness-of-fit test (4.9 with 8 df) gives the essential support in favour of two-class model to be the best fit for the data. The unconditional probabilities of being in each latent class 1 and 2 were found to be 0.542 and 0.458 respectively. The two latent classes 1 and 2 were identified as "No MI" and "MI" based on the fact that in class 1, the probabilities of being positive for the indicators were low compared to that of class 2. For example, the probability of presence of clinical history is 0.195 in class 1 and 0.791 in class 2 and probability of presence of positive Q-wave is 0.00 in class 1 and 0.767 in class 2. Thus Q-waves have a sensitivity of 0.77 for MI and a specificity of 1.0 for MI. If the disease is correctly diagnosed the results of the latent class analysis would coincide with the results of traditionally recommended methods.

3.3. Economic research

Latent class analysis may be applied to classify types of attitude structures from survey responses, consumer segments from demographic and preference variables, or categorize sub-populations based on their responses to test items. LCA has been applied to the field of economics in consumer behaviour, health economics etc as a relatively straightforward way to model unobserved heterogeneity in the population. Bhatnagar and Ghose (2004) applied LCA as a tool for segmentation of web shoppers according to their preferences over several product options. To investigate the nature of benefit segments that exist in online marketing, a Likert-type scaled data on several attributes for evaluating the web-based store were analysed. BIC values for models of 2, 3 and 4 latent classes are 8480.096, 8218.486 and 8228.013 respectively, which indicates that the sample has three distinct segments. The conditional response probabilities can be used to name the segments. The first segment mainly prefers buying computer hardware and software at online stores. The second segment have strong dislike for buying anything from an online store. The last segment buys mainly standardized products at lower prices at the online store.

4. Discussion and Future Research

The existing frame work of Latent Class Analysis can be further modified in view of more realistic and more generalized situations. Modifications in different aspects of the formulation of the model would lead to more generalized form of the analysis. Different sampling frames with different parameterization of the unknown parameters can be further explored. Other options for the imputation procedure can also be investigated.

Acknowledgement

I thank my supervisors Dr. Diganta Mukherjee and Prof. Sugata Sen Roy for their help and encouragement in the preparation of this manuscript

5. References

1) Bhatnagar, G., Amit Ghose and Sanjoy. 2004. A latent class segmentation analysis of e-shoppers. *Journal of Business Research*, 57: 758 – 767.

2) Biemer Paul, P. 2011. Latent Class Analysis of Survey Error, A John Wiley and Sons, Inc. Publication.
3) Castle, D. J., Sham, P. C., Wessely, S. and Murray, R. M. 1994. The sub-typing of schizophrenia in men and women: a latent class analysis. *Psychological Medicine*, 24: 41 - 51.
4) Collins, S., Linda, M., Lanza, J and Stephanie T. 2010. Latent Class and Latent Transition analysis with applications in the Social, Behavioural and Health Sciences, A John Wiley Sons, Inc., Publications.
5) Dempster, A. P., Laird, N. M and Rubin, D. B. 1977. Maximum Likelihood from Incomplete Data *via* the EM Algorithm. *Journal of the Royal Statistical Society*, 39 (1):1 - 38.
6) Dunn, S., Kate, M., Jordan Kelvin and Croft Peter R. 2006. Characterizing the Course of Low Back Pain: A Latent Class Analysis. *American Journal of Epidemiology*, 163: 8.
7) Formann, A. K. 1984. Constrained Latent Class Models: Theory and Applications, *British Journal of Mathematical and Statistical Psychology*, 38: 87 - 111.
8) Formann, A. K. 1992. Linear Logistic Latent Class Analysis for Polytomous Data, *Journal of the American Statistical Association*, 87: 418.
9) Goodman, D and Leo, A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61 (2): 215.
10) Haberman, S. J. 1979. Analysis of Qualitative Data (Vol. 2: New Developments), New York: Academic Press.
11) Hagenaars, J. 1990. Categorical longitudinal data: Log-linear panel, trend, and cohort analysis. Newbury Park, CA: Sage.
12) Lazarsfeld, P. F. 1950. The logical and mathematical foundation of latent structure analysis and the interpretation and mathematical foundation of latent structure analysis. S.A. Stou er *et al*. (eds.), Measurement and Prediction, 362 - 472. Princeton, NJ: Princeton University Press.
13) Lin Su-Wei, Tai and Wen Chun. 2015. Latent Class Analysis of Students' Mathematics Learning Strategies and the Relationship between Learning Strategy and Mathematical Literacy. *Universal Journal of Educational Research*, 3(6): 390 - 395.
14) Mooijaart, A. B. 1992. The EM Algorithm For Latent Class Analysis With Equality Constraints. *Psychometrika*, 57 (2): 261 - 269.
15) Patterson, D., Blossom H., Dayton Mitchell and Graubard Barry, I. 2002. Latent Class Analysis of Complex Sample Survey Data. *Journal of the American Statistical Association*, 97: 721 - 741.
16) Rees Kees Van, Vermunt Jeroen and Verboord Marc. 1999. Cultural classifications under discussion Latent class analysis of highbrow and lowbrow reading. *Poetics*, 26: 349 - 365.
17) Rindskopf, D and Rindskopf, W. 1986. The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, 5: 21 - 27.
18) Skrondal, A., Rabe-Hesketh, S. 2004. Interdisciplinary Statistics - General Latent Variable Modelling - Multilevel, Longitudinal, and Structural Equation Models, Chapman and Hall/CRC.
19) Vermunt, J. K. 1997. Log - linear models for event histories. Thousand Oaks, CA: Sage.
20) Vermunt, K and Jeroen K. 2010. Latent class modelling with covariates: Two improved three-step approaches. *Political Analysis*, 18: 450 - 469.