Clustering of Inflammatory Skin Disease Patients Using Latent Class Analysis

Rupesh K Khare† & Gauri Gupta‡

† Hewitt Associates, Gurgaon, India.

e-mail: rupesh_khare@hotmail.com

‡ MICA, Ahmedabad, India.

e-mail: gauri8@mica.ac.in

Key Words: Clustering, Latent Class Analysis, Latent Gold, Consumer Behavior & Attitudes

ABSTRACT

This paper highlights the concept, advantages and application of Latent Class Analysis (LCA). The first section presents a preview of LCA, a clustering technique, to underline the method's suitability for various research and analytics work. The second section highlights the relevance of LCA in light of the limitations encountered by other frequently used clustering techniques such as K Means and hierarchical clustering. Subsequently, the third section underscores the application of LCA by presenting a real life project executed by the authors while they were working with marketRx, a consulting company in pharmaceutical analytics.

1.0 INTRODUCTION

Cluster analysis is a class of statistical techniques that can be applied to data that exhibit "natural" groupings. According to Kaufman and Rousseeuw (1990), cluster analysis is the classification of similar objects into groups, where the number of groups, as well as their forms are unknown. In marketing research, Cluster Analysis is an extremely popular classification technique. The technique divides a set of items into two or more mutually exclusive unknown groups based on combinations of interval variables. The goal of cluster analysis is to organize items into groups in such a way that the degree of similarity is maximized for the items within a group and minimized between groups.

In practice, clustering is performed primarily using traditional methods such as K Means and hierarchical clustering. Lazarsfeld, in 1950, introduced another technique as Latent Class Analysis (LCA). He used the technique as a tool for building clustering based on dichotomous observed variables. More than 20 years later, Goodman (1974) made the model applicable in practice by developing an algorithm for obtaining maximum likelihood estimates of the model parameters. However, lately due to development of extended algorithms which allow today's technology to perform LCA on data containing more than just a few variables, there has been significant increase in interest in LCA.

LCA is a probabilistic technique, which can be applied for cluster, factor, or regression analytic purposes. The analysis evaluates a possible association between a set of observed categorical variables and unobserved latent variable. The technique assumes that although each object belongs to one class or cluster, but still that there is uncertainty about an object's class membership. The basic premise of LCA models is that the covariation actually observed among the observed variables is due to each observed variable's relationship to the latent variable (McCutcheon, 1987). Latent classes are defined by the criterion of "conditional independence," meaning that, within each latent class, each variable is statistically independent of every other variable. LCA (Goodman, 1978; McCutcheon, 1987) enables estimations of the probability that any given individual would fall into a specific latent class. Given the class that the individual is in, LCA also estimates the probability that the individual would endorse a certain behavior or trait.

The latent class model with three observed variables A, B and C can be written as:

$$\Pr(A=a,B=b,C=c) = \sum_{x=1}^{N} \Pr(A=a|X=x)\,\Pr(B=b|X=x)\,\Pr(C=c|X=x)\,\Pr(X=x)$$

Where X (x =1…..n) denotes latent class.

Most statisticians credit Lazarsfeld and Henry (1968) with the origins of latent class analysis and Goodman (1974) with the computational breakthroughs that made it practical. Goodman's maximum likelihood approach (1974, pp. 216-218; Goodman, 2002; see also McCutcheon, 1987, pp. 21-27) remains the standard way to estimating parameters in the latent class model.

The likelihood function is built on the product of the joint probabilities of each observed response profile, given the assignment of the hypothesized latent class (X=n). These joint probabilities are summed across latent classes, so the likelihood function ultimately contains information from the indicators, which are observed, and from latent class membership, which is not (Dayton & Macready, 2002). Likelihood functions can, therefore, be extremely complex.

A way around the difficulty of constructing a likelihood function from partly unobserved information involves estimating the missing information on latent class membership, then maximizing the likelihood for the provisional but complete 'data.' The approach involves alternating steps that first calculate the likelihood function's *expected* value, and then find the parameter values that *maximize* the function. The expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) first calculates the likelihood function's expected value, given observed counts for a complete (though provisional) contingency table, and given provisional values for the latent class model's parameter estimates. Next, it maximizes the likelihood function and, in so doing, updates the parameter estimates.

The E-M cycle must begin with the arbitrary assignment of each response profile to a specific latent class X. This random assignment, which "fills in" missing information on latent class membership, constitutes the first *expectation step*. Goodman's seminal (1974) approach, which incorporates the EM algorithm, updates its parameter estimates in proportion to the observed marginal probabilities for each response profile. The algorithm repeats until expected and observed values converge (Uebersax & Grove, 1990, equations 3 and 4, p. 569; Goodman, 2002, p. 29) and the parameter estimates stabilize.

The E-M algorithm has a significant limitation in that it does not produce standard errors for the parameter estimates, so one cannot test hypotheses or estimate confidence intervals that relate to the parameters. Techniques proposed to overcome this limitation (Goodman, 1972; Louis, 1982) are complex and obtain estimates for the entire covariance matrix only for latent class models with certain kinds of manifest indicators.

The "goodness" of a latent class model's fit to observed data is conventionally evaluated by Pearson and likelihood ratio (LR) chi-square statistics. These statistics compare observed counts with the counts expected under the assumption of conditional independence. The statistics' associated degrees of freedom are equal to the number of non-redundant observed counts in the relevant contingency table, minus the number of LC parameters that the model estimates. Akaike and Bayes information criteria, which are based on the likelihood ratio chi-square, and which

account for the number of estimated parameters, are also calculable to compare models with differing restrictions on the parameters.

2.0 LCA IN COMPARISON WITH TRADITIONAL CLUSTERING TECHNIQUES

LCA differs from traditional cluster analysis algorithms, which group cases near each other by some definition of distance. The latent class approach defines one cluster per latent class, using model-based probabilities to classify cases. The technique has an advantage of using data from both simple and complicated distributional forms for clustering. Moreover, LCA does not restrict the analysis to a predetermined number of clusters as does K-Means clustering.

The table below summarizes the differences between latent class segmentation and traditional clustering methods:

	Latent Class Modeling	**Traditional Clustering Approaches**
Models	Classifies cases into clusters based upon membership probabilities estimated directly from the model.	Identifies clusters by grouping cases that are "near" each other according to an ad-hoc definition of "distance" (e.g., "Euclidian distance").
Variable Use	Variables may be continuous, categorical (nominal or ordinal), counts or any combination of these.	Traditional clustering methods like K-means or Hierarchical accept categorical or continuous variables only, not even both at the same time.
Covariates	Covariates and demographics can be used. Covariates are variables used to describe or predict (rather than to define or measure) the latent variable.	Does not utilize covariates.

Data Assumptions	The model accommodates data that violates normality, linear relationship between variables and homogeneity of variance.	Variables must be normally distributed, have a linear relationship and be homogeneous. Distance methods highly affected by outliers.
Formal statistical procedures for determining the # of clusters	BIC (Bayes Information criterion), AIC (Aikike Information criterion) and CAIC (Consistent AIC) allow for easier comparison between models.	F ratios to measure the between group variance to Within group variance are used as goodness of fit measures. The objective of traditional clustering methods is to "maximize" the between group variation and minimize the "within" group variation

An important difference between standard cluster analysis techniques and LCA is that the latter is a model-based clustering approach. This means that a statistical model is postulated for the population from which the sample under study is coming. More precisely, it is assumed that the data is generated by a mixture of underlying probability distributions. When using the maximum likelihood method for parameter estimation, the clustering problem involves maximizing a log-likelihood function. This is similar to standard non-hierarchical cluster techniques in which the allocation of objects to clusters should be optimal according some criterion. These criteria typically involve minimizing the within-cluster variation and/or maximizing the between-cluster variation. An advantage of using a statistical model is, however, that the choice of the cluster criterion is less arbitrary. Nevertheless, the log-likelihood functions corresponding to LC cluster models may be similar to the criteria used by certain non-hierarchical cluster techniques like k-means.

LC models do not rely on the traditional modeling assumptions which are often violated in practice (linear relationship, normal distribution, homogeneity). Hence, they are less subject to biases associated with data not conforming to model assumptions. In addition, LC models have

recently been extended (Vermunt and Magidson, 2000a, 2000b) to include variables of mixed scale types (nominal, ordinal, continuous and/or count variables) in the same analysis. Also, for improved cluster or segment description the relationship between the latent classes and external variables (covariates) can be assessed simultaneously with the identification of the clusters. This eliminates the need for the usual second stage of analysis where a discriminant analysis is performed to relate the cluster results to demographic and other variables.

3.0 APPLICATION OF LCA

3.1 Project Background:

Anonymous Inc. (name changed to maintain confidentiality), a US based pharmaceutical company experienced a number of additions to its dermatology product line. The company wanted to develop an understanding of the products in dermatology market in US from both physician and the consumer perspectives. With this intention, the company approached marketRx to conduct a two steps primary research in the year 2007. In the first step of the research, which is out of the scope of this paper, Anonymous Inc. specifically sought to develop a comprehensive understanding of the attitudes and behaviors of physicians regarding the products, including OTC products that are used to treat a particular disease pertaining to skin of the patients. The second step of the study focused on the consumer ATU and segmentation to develop an understanding of the consumer's decision making and process flow that leads to selecting a treatment. A quantitative assessment of attitudes and behaviors of consumers regarding the products that are used for treatment, including OTC products was also desired. The scope of this paper is limited to the illustration of consumer segmentation component of the second step.

3.2 Objective of the Research:

Anonymous Inc wanted to understand the various segments that may exist within the patient population and gain insight into what the differing attitudinal perceptions and informational needs:

- In order to identify who are the likely best potential targets of the product, as well as how best to position those products.
- It is assumed that the segmentation schema will be for the actual patients themselves rather than the parents or guardians on those that suffer from the disease.

3.3 Methodology:

For better project management, survey administration and efficiency, only one questionnaire was designed to capture all of the metrics necessary to conduct an attitudinal segmentation of consumers as well as the traditional ATU metrics that can then be examined on a consumer segment basis. The segmentation analysis performed based on the responses of the questions included segmentation metrics as:

3.31 Demographic Metrics:

Age, gender, geography, climate, social security, extracurricular activities, physical activity, socio-economic status, reimbursement coverage, shaving habits for men, amount of exposure to sun, familial support if under 18, schooling/ employment status, etc.

3.32 Self Perception Metrics:

Attitudes about personal appearance, attractiveness, confidence and self-esteem, etc

3.33 Disease Related Metrics:

Knowledge about disease, physicians seen for treatment, compliance with dosing schedule, severity of disease, etc

3.4 Data Collection and Segmentation Variables Creation:

This questionnaire was programmed to a 45 minutes long Internet-based survey and conducted among consumers. The survey was fielded for four weeks and the data was collected in SPSS. Survey participants were recruited through various methods, namely:

- Purchased lists of patients
- Social networking websites
- On-line forum posts
- Anonymous Incs' opt-in database

The study targeted three distinct populations:

- Teens (age 13-17) who have taken prescription drugs in the last 12 months
- Adults (age 18-25) who have taken prescription drugs in the last 12 months
- Parents of teens who have taken prescription drugs in the last 12 months

Selection of clustering variable is the next step in the process. This selection is very crucial in the light of producing meaningful clusters and subsequently enabling organization to take actions based on these clusters. Though differences across segments exist for other variables of interest, ultimately the variables included in the segmentation model were based on responses to the certain questions of the questionnaire asked in the survey. Parts of a few such questions are presented below :

- Attribute importance ratings:

…… Using the table below, please rate how important it is for a drug to have the following qualities. Use a scale from 1 to 7 where 1 is "Somewhat important" and 7 is "Most important"……

- Responsiveness to various personalities recommending/ representing the product:

…….. Please indicate how interested you would be in a drug if the following people recommended/represented it. Use a scale from 1 to 7 where 1 is "Not at all interested" and 7 is "Extremely interested"……..

- Preference for the tone an advertisement:

…….How appealing would you find the following types of drug advertisements? Use a scale of 1 to 7 where 1 is “Not at all appealing” and 7 is “Very appealing”……

3.5 Model Estimation Using Latent Gold:

The modeling process starts with the preparation of clustering variables. The Variables are identified using various methods like Correlation Matrix Analysis, Factor Analysis, regression Analysis etc. These final variables are used in latent class segmentation to obtain various latent classes segments.

Number of clusters is determined using Model chi-square (L2) and p-value. L2 represents the amount of the relationship between the variables that remains unexplained by a model. It is assumed that the lower the value, the better the model fits the data. However, a good fit model has a p-value greater than 0.05.

Model Validation is done by running nearly five hundred iterations and re-estimating p-value. With minimal standard error, the new estimate of p-value should be similar or greater than the old estimate. The model is validated if it qualifies this condition.

3.6 Segmentation for Different Populations :

3.6.1 Adult Population :

Segment 1:

They are predominantly slightly older females who are well educated and have high earnings. Respondents from this segment carry very positive attitude and are very particular about their appearance. They are quite aware of available OTC products for the medication of the disease. This product knowledge and self consciousness propel them to use these OTC products. Consumers from this segment quickly respond to the promotions of the products but carefully

evaluate the efficacy and the side effects of these products. They sometimes visit doctors incase these products don't produce desirable results.

Segment 2:

Consumers in this segment are primarily young students from both the genders. These consumers belong to all income groups and educational backgrounds. They are not very conscious about their appearance and also don't carry a high self perception. They are also moderately responsive to the promotions of the products. Self-consciousness and OTC failures were main motivators to see a doctor however efficacy of the product is important to this group also.

Segment 3:

This group contains more females than males who are mainly young students. The education level and household incomes are low for the respondents from this segment. This group also are not very conscious about their appearance and also don't carry a high self perception. They are very dependent on physicians for selecting treatment choice. Their general response to product advertisement is not high.

3.6.2 Teen Population:

Segment 1:

This group highly uses the OTC products and efficacy is the most important product attribute for them. They have moderately low responsiveness to drug promotion. The respondents from this group does not belong to any particular educational background or parental income. They are not quite concerned about their looks and carry low self perceptions.

Segment 2:

No product attribute stands out as more important to this segment. They show low interest to drug promotion but their parents have large income. They exhibit low concern about their looks and carry low self perceptions. Incase if the drug does not give them desired results, they are less likely to visit doctors.

Segment 3:

This group is internet savvy and their parents have good range as far as income is concerned. This group highly uses the OTC products and are moderately high concern about looks. Though the parents of the members of this segment influence the decision to see a doctor but still these members respond to drug promotion especially phone calls or advertisements.

Segment 4:

Seven out of ten members of this segment are females who carry highly particular about their looks and self esteem. They have low educational and income profile. They also have very good knowledge of OTC products and very responsive to drug promotion.

4.0 CONCLUSIONS

This clustering exercise produced accurate and actionable segmentation of different populations. The analysis resulted in memberships of respondents in each segment. In Adult population, three distinct segments were identified on the basis of demographic, self perception and disease related metrics. Interestingly, the membership distribution among these segments was not significantly different from each other. On the other hand, analysis produced four segments in Teens population. The membership distribution had significant variation. Segment 1 had the maximum membership with nearly one third of population while Segment 4 had minimum membership with approximately one fifths of population. This accurate assessment of model fit is not available in K means. Additionally, LCA produced undoubtedly better segments on the basis of behavioral traits of consumers. These superior results are primarily because LC approach allows cases to be classified into clusters using model based posterior membership probabilities estimated by maximum likelihood (ML) methods.

REFERENCES

1. Heinen, Ton, Latent class and discrete latent trait models, Sage Publications, 1996
2. Hagenaars, AJ. & Mccutcheon, AL., Applied Latent Class Analysis, Cambridge University Press, 2002
3. Dayton, CM, Latent Class Scaling Analysis, Sage Publications, 1998
4. Magidson, Jay & Vermunt, JK, Latent class models for clustering: A comparison with K-means, Canadian Journal of Marketing Research, Volume 20, 2002
5. Thompson, DM, Latent Class Analysis in SAS®: Promise, Problems, and Programming , SAS Global Forum, 2007
6. Vermunt, JK & Magidson, J, Latent Class Cluster Analysis, Articles at Statistical Innovations, Inc, 2002.
7. Consumer Acne- Attitudes, Trial and Usage Study, A Primary Research conducted and published by marketRx, 2007