A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 269, 2012

Aneta Rybicka[*], *Marcin Pełka*[**]

EVALUATION OF COLLEGE STUDENTS PREFERENCES WITH APPLICATION OF LATENT CLASS ANALYSIS

Abstract. In social sciences, especially in economy, to reveal relations between variables it's easy to apply many known statistical tools when we deal with observable (measureable) variables. The problems appear when dealing with latent variables – that are not directly observed and they are of subjective matter. It's also an important issue to measure relations between latent variables.

The example of latent variables are preferences. The preferences play a very important role in economy. Very often real market decisions, choices (or answers in a questionnaire) are described by non-metric variables (nominal and ordinal). These variables are also called qualitative.

The latent class analysis allows to reveal hidden relations between observable variables. The observable variables allow, with a specified probability, to find a non-observable phenomenon. The latent class analysis allows to analyze the qualitative data [see: McCutcheon 1987, p. 7; 11; Hagenaars 1993, p. 21–23]. LCA was introduced by Lazarsfeld in 1950 [1968].

The paper presents evaluation of college students with application of latent class analysis. To obtain such a goal data collected (winter recruitment of 2008/2009) by a college in Walbrzych was used.

Key words: Latent class analysis, segmentation, evaluation of college students

I. INTRODUCTION

Statistical research with application of latent class analysis assumes that within sample there is a finite number of relatively similar groups (segments) of objects (consumers). There are some important differences within those groups. These groups are not *prior* known, they are latent because group memberships and number of groups are unknown [Bąk 2004, pp. 134].

Latent class models, as a part of multivariate statistical methods, are a part of finite mixture models [Domański, Pruska 2000, pp. 30–36]. The share of each element within the mixture is determined by mixing parameter. Sum of those parameters is equal to one.

When latent class models applied in segmentation researches, mixing parameter is interpreted as the size of a segment.

[*] Ph.D., Chair of Econometrics and Informatics, Wroclaw University of Economics.
[**] Ph.D., Chair of Econometrics and Informatics, Wroclaw University of Economics.

Applications of latent class models and latent class regression models (especially in satisfaction surveys) are presented in: LaLonde S.M. [1996]; Colias J., Horn B., Wilkshire E. [2007]; Cooil B. et. al. [2007]; Hill N., Roche G., Allen R. [2007]; Allen D.R. [2004]. Other applications of latent class models are presented in: Shen J., Sakata Y., Hashimoto Y. [2006]; DeSarbo W. S., Ramaswamy V., Cohen S. H. [1995]; Moore W. L., Gray-Lee J., Louviere J. [1996], Green W. H., Hensher D. A. [2002], Pacifico D. [2009].

The paper presents evaluation of college students with application of latent class analysis. To obtain such a goal data collected (winter recruitment of 2008/2009) by a college in Walbrzych was used.

II. LATENT CLASS ANALYSIS

Procedure of construction and estimation of a latent class model consists of following steps [Bąk 2004, pp. 134–135]:

- determine the conditional distribution for the respondent,
- determine distribution for the respondent (non conditional) – it's the weighted sum of conditional distribution, where weights are estimated probabilities of respondents' segment membership,
- forming maximum likehood function – it's the product of individual distributions in condition that they are independent,
- model estimation (parameters, segments size),
- estimation of *posterior* probabilities of respondents segment membership.

Latent class models have some important formal properties, which are very important in segmentation [Bąk 2004, p. 141; Cameron, Trivedi 2005, p. 621–625]:

- they allow to identify segments (based on observed variables or dependent variable),
- they have one categorical latent variable (the number of categories is equal to the number of segments),
- the estimated cluster membership is based on probabilities,
- observed variables can be either nominal, ordinal, interval or ratio,
- model can include concomitant and explanatory variables as well.

The goal of latent class analysis is to find the size of each latent class and the estimated probabilities of occurrence for each category of each variable, within particular latent class. Goodness of fit is typically tested by calculating a chi square value, based on actual versus fitted cell frequencies. It is of further interest to note that latent class analysis, unlike conventional factor analysis:

1. Avoid the computation of correlation measures, such as phi or the tetrachoric measure.
2. Does not assume linearity or even monotonicity of relationships among the qualitative variables.

3. Is not constrained to the use of pairwise associations. From a somewhat more philosophical viewpoint latent class analysis provides a perspective on "causality" via the local independence assumption. That is, under this view the association between two (or more) variables has been "explained" when their joint probability of occurrence within the latent class is a product of the respective marginals. In effect, this says that their partial correlation is zero, within the latent class [see: Green, Carmone, Wachspress 1976, s. 171–172].

There are three main types of latent class models [see: Magdison, Vermunt 2003, pp. 2]:

- *Latent Class Cluster Models* (LCCM).
- *Latent Class Factor Models* (LCFM).
- *Latent Class Regression and Choice Models* (LCRM).

III. ESTIMATION OF PARAMETERS

The latent class model can be defined as follows [see: DeSarbo i Wedel 1994, Virens 2001]:

$$f(\mathbf{y} \mid \mathbf{\Phi}) = \sum_{c=1}^{C} \pi_c f(\mathbf{y} \mid \mathbf{\theta}_c), \tag{1}$$

where: $f(\mathbf{y} \mid \mathbf{\Phi})$ – function of observation distribution; $\sum_{c=1}^{C} \pi_c$ – distribution of non conditional probabilities which represents the membership to latent clusters; $f(y \mid \mathbf{\theta}_c)$ – function representing conditional distributions; $\mathbf{\Phi} = (\pi, \mathbf{\theta})$ – all unknown model parameters; $\mathbf{\theta}_c$ – vector of unknown parameters for c-th cluster.

On the basis of the latent class model (equation 1) parameters of each segments are estimated with application of maximum likehood method. The maximum likehood function for a sample of S consumers can be defined as follows:

$$L(\mathbf{y}; \mathbf{\Phi}) = \prod_{s=1}^{S} f(y_s \mid \mathbf{\Phi}). \tag{2}$$

The estimation of function’s parameters is done with application of Newton-Raphson or EM algorithm.

IV. PREFERENCE ANALYSIS

One of Walbrzych college schools asked their students, while winter 2008/2009 recruitment, to indicate the factors that had influence on choosing this school and specialization chosen. Students could choose one of following factors:

- X_1 – place of learning,
- X_2 – learning without a fee,
- X_3 – good school opinion,
- X_4 – additional courses without a fee,
- X_5 – willingness to learn,
- X_6 – need to raise qualifications,
- X_7 – the possibility of postponing army service,
- X_8 – the possibility of getting certificate of being a student.

The respondents indicated any number of factors that had influence on choice they have made by placing "X" next to the factor.

The **R** software allows to estimate latent class models with application of poLCA function from poLCA package. The poLCA package is designed to estimate latent class models with dichotomous and polytomous outcome variables, as well as models with covariates.

poLCA uses the assumption of local independence to estimate a mixture model of latent class models with application of multi-way tables. The number of clusters is specified by the user. Estimated parameters include the class-conditional response probabilities for each manifest variable, the "mixing" proportions denoting population share of observations corresponding to each latent multi-way table, and coefficients on any class-predictor covariates, if specified in the model.

poLCA uses EM and Newton-Raphson algorithms to maximize the latent class model log-likelihood function. Depending on the starting parameters, this algorithm may only locate a local, rather than global, maximum.

The choice of number of clusters depends on values of AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) functions [see: Kasprzyk 2009, p. 292–294]. The number of clusters is usually indicated by lowest value of BIC. Different number of cluster (from 2 up to 8) have been considered. Values of AIC, BIC, χ^2 (Pearson Chi-square goodness of fit statistic for fitted versus observed multiway tables) and G^2 (Likelihood ratio/deviance statistic) for estimated model and number of clusters are presented in tab. 1.

Table 1. Choosing the number of clusters

Crite-rion	Number of clusters							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
AIC	1927,54	**1893,79**	1888,13	1891,25	1900,26	1902,51	1905,95	1927,54
BIC	1988,20	**1986,57**	2013,02	2048,26	2089,38	2123,75	2159,30	1988,20
χ^2	177,99	**126,24**	102,58	87,70	78,70	62,96	48,40	177,99
G^2	291,57	**211,22**	149,37	131,21	102,25	73,04	44,29	291,57

Source: own computations.

The lowest value of BIC is reached for two clusters (the decision is the same when considering AIC).

Table 2. Estimated probabilities for answers 1 and 2

Factor and estimated probabilities			
$X_1		$ X_2	
Pr (1)	Pr (2)	Pr (1)	Pr (2)
class 1: **0.6088**	0.3912	class 1: **0.5005**	0.4995
class 2: 0.1721	**0.8279**	class 2: 0.0000	**1.0000**
class 3: **0.6522**	0.3478	class 3: 0.4304	**0.5696**
$ X_3		$ X_4	
Pr (1)	Pr (2)	Pr (1)	Pr (2)
class 1: **0.6911**	0.3089	class 1: **0.9054**	0.0946
class 2: 0.1954	**0.8046**	class 2: 0.0132	**0.9868**
class 3: **0.7296**	0.2704	class 3: **0.9213**	0.0787
$ X_5		$ X_6	
Pr (1)	Pr (2)	Pr (1)	Pr (2)
class 1: **1.0000**	0.0000	class 1: **0.5847**	0.4153
class 2: 0.0384	**0.9616**	class 2: **0.8670**	0.1330
class 3: 0.0000	**1.0000**	class 3: **0.9587**	0.0413
$ X_7		$ X_8	
Pr (1)	Pr (2)	Pr (1)	Pr (2)
class 1: **0.9047**	0.0953	class 1: **0.6902**	0.3098
class 2: **1.0000**	0.0000	class 2: **0.9200**	0.0800
class 3: **0.9948**	0.0052	class 3: **0.8972**	0.1028

Highest values are bolded.
Source: own computation.

Pr (1) in tab. 2 is a probability that factor was chosen. Pr (2) is a probability that factor was not chosen by college students.

Estimated class population shares are following: 0.1601 for cluster 1, 0.1039 – cluster 2 and 0.736 – cluster 3. Predicted class memberships (by modal posterior probabilities) for these clusters are: 0.1603 for cluster 1, 0.1183 for cluster 2 and 0.7214 for cluster 3.

Number of observations in estimated model is equal to 262, the number of estimated parameters reached 26, residual degrees of freedom 229. Maximum log-likehood is equal to 920.8976.

V. FINAL REMARKS

Latent class analysis allowed to detect an unknown structure of three classes of students (participants) of Walbrzych college school. The largest class shares are estimated for class 3 – representing 73.6 of the population. Taking into account the probability distribution of responses according this class the biggest role in the choice of school were: learning without a fee, willingness to learn, need to raise qualifications, the possibility of postponing army service.

Latent class analysis can be a useful tool to detect and study consumer preferences. The **R** software allows an easy and efficient estimation of latent class model with dichotomous and polytomous outcome variables as well as latent class models with covariates

BIBLIOGRAPHY

Allen D.R (2004), *Customer satisfaction research management*, Quality Press, Milwaukee.

Bąk A. (2004), *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*. Wyd. AE we Wrocławiu, Wrocław.

Cameron A. C., Trivedi P. K. (2005), *Microeconometrics. Methods and Applications,* Cambridge University Press, New York.

Colias J., Horn B., Wilkshire E. (2007), *Improving customer satisfaction and loyalty with time-series cross-sectional models*, [URL:] www.decisionanalyst.com.

Cooil B., Keiningham T.L., Aksoy L., Hsu M. (2007), *A longitudinal analysis of customer satisfaction and share wallet: Investigating the moderating effect of customer characteristics*, ``Journal of Marketing``, Vol. 71, No. 1, s. 67–83.

Domański C., Pruska K. (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.

DeSarbo W. S., Ramaswamy V., Cohen S. H. [1995], *Market Segmentation With Choice-Based Conjoint Analysis*, Maketing Letters, 6, 137–148.

DeSarbo M., Wedel W.S. (1994), *A review of recent developments in latent class regression models.* [In:] R.P. Bagozzi, (Ed.), *Advanced Methods of Marketing Research*, Blackwell Publishers, Cambridge, pp. 352–388

Green P.E., Carmone F.J., Wachspress D.P., *Consumer segmentation via latent class analysis*, "Journal of consumer research" 3, pp. 170–174.

Green W. H., Hensher D. A. [2002], *A latent class model for discrete choice analysis: contrast with mixed logit*, Institute of transport Studies, Working Paper, [URL:] pages.stern.nyu.edu.

Hill N., Roche G., Allen R. (2007) ,*Customer satisfaction: The customer experience through the customer's eyes*, Cogent Publishing, London.

Kasprzyk I. (2009), Wykorzystanie konfiguracyjnej analizy częstości w analizie klas ukrytych, PN UE we Wrocławiu nr 51, pp. 36–45.

LaLonde S.M. (1996), *Key driver analysis latent class regression*, Proceedings of the Survey Research Methods Section, American Statistical Association, s. 474–478 [URL:] www.amstat.org.

Magidson J., Vermunt J.K. (2003), *A nontechnical introduction to latent class models*. „DMA Research Council Journal", [URL:] statisticalinnovations.com.
McCutcheon, A.L. (1987), *Latent class analysis*. Newbury Park: SAGE Publications, 1987.
Moore W. L., Gray-Lee J., Louviere J. [1996], *A Cross-Validity Comparision of Conjoint Analysis and Choice Models at Different Levels of Aggregation*, Working Paper, University of Utah, November.
Pacifico D. [2009], *Modelling unobserved heterogeneity in discrete choice models of labour supply*, MPRA Paper No. 19030, [URL:] mpra.ub.uni-muenchen.de.
Shen J., Sakata Y., Hashimoto Y. [2006], *A Comparision between Latent Class Model and Mixed Logit Model for transport Mode Choice: Evidences from Two Datasets of Japan*, Discussion paper In Economics And Business, Discussion Paper 06–05, January 2006, [URL:] http://www2.econ.osaka-u.ac.jp.
Virens M. (2001), *Market segmentation. Analytical developments and application guidelines*, Millward Brown IntelliQuest.

Aneta Rybicka, *Marcin Pełka*

OCENA PREFERENCJI UCZNIÓW SZKOŁY POLICEALNEJ Z WYKORZYSTANIEM ANALIZY KLAS UKRYTYCH

W ekonomii do badania zależności między zmiennymi łatwo jest zastosować metody statystyczne, gdy mamy do czynienia z obserwowalnymi cechami mierzalnymi. Problem pojawia się natomiast w przypadku „cech ukrytych", czyli takich, których nie da się bezpośrednio zmierzyć, a ich ocena jest subiektywna. Również istotnym zagadnieniem jest badanie charakteru i siły zależności między cechami niemierzalnymi (ukrytymi).

Przykładem zmiennych ukrytych są m.in. preferencje. W ekonomii preferencje konsumentów zajmują ważne miejsce. Bardzo często wybory, czyli decyzje podejmowane na rynku (np. odpowiedzi w badaniu ankietowym) przez konsumentów są opisywane przez zmienne niemetryczne (nominalne i porządkowe), które czasem nazywa się zmiennymi jakościowymi.

Analiza klas ukrytych pozwala na odkrywanie nieobserwowalnych zależności pomiędzy zmiennymi obserwowalnymi. Zmienne obserwowalne pozwalają z określonym prawdopodobieństwem stwierdzić zaistnienie zjawiska nieobserwowalnego. Analiza klas ukrytych pozwala na analizę danych jakościowych [zob. McCutcheon 1987, s. 7; 11; Hagenaars 1993, s. 21–23]. Metoda ta została wprowadzona przez Lazarsfeld`a w 1950 roku [1968].

W artykule przedstawiono zastosowanie analizy klas ukrytych w badaniach preferencji na przykładzie preferencji uczniów szkoły policealnej. W tym celu wykorzystano dane zebrane przez szkołę policealną w trakcie naboru zimowego roku 2008/2009.