ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 135, 2010

Iwona Kasprzyk[*]

LATENT CLASS MODELS IN THE R SOFTWARE

ABSTRACT. This paper presents traditional approach to latent class modelling and the latent class regression (LCR) as other type of latent class models. The latent class model was introduced by Lazarsfeld and Henry (1968). We can use poLCA package to estimate latent class and latent class regression models for polytomous outcome variables in the R statistical computing environment. In the basic latent class models it is assumed that all variables are mutually independent. The latent class regression (LCR) model allows to estimate the effects of covariates on predicting latent class membership.

The main aim of this article is identyfication of number of classes and next building latent class regression model.

Key words: latent class models, latent class regression models.

I. INTRODUCTION

Latent class model is becoming one of the most popular data analysis tools in social, marketing research. The method was introduced by Lazarseld and Henry (1968) for dichotomous variables. The latent class model assumes that the latent viariable is nominal. The methodology was formalized and extended to nominal variables (more than two variables) by Goodman (1974). Goodman (1974) popularized the maximum likelihood method, which nowadays is one of the most popular methods used in latent class software programs.

In this article we describe the traditional latent class and latent class regression model, where we can research influence of covariates on observed variables. The covariate for example are demographic variables such as: age, sex, place and other as: seniority. The example, which is discribed in this article gives answer to for example what is the odds of being the person, who has given status of employment in the given class in comparison to the person, who are in the first class.

II. LATENT CLASS MODELS

The latent class model assumes that each observation is a member of one and only one of W latent classes. Moreover, this method assumes local

[*] Ph.D. Student, Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

independence between the observed variables. The criterion of local independence contributes to variables its independence.

The aim of traditional latent class model is to determine the smallest number of latent classes W that is sufficient to explain the associations observed among the manifest variables.

The latent class model can be written as:

$$P(\mathbf{y}) = \sum_{w=1}^{W} \pi_w \prod_{i=1}^{I} \pi_{iy^{(i)}|w}, \tag{1}$$

where:

$\mathbf{y}$ – an object's scores on set of observed variables

$\sum_{w=1}^{W} \pi_w$ – the prior probability of belonging to w latent class,

$\pi_{iy^{(i)}|w}$ – the conditional probability that ith observation variable takes value $y^{(i)}$ on condition being in latent class w.

In the latent class model it is not only the estimation of the model parameters that is interesting, another important problem is the classification of observations into clusters. This can be based on the posterior class membership probabilities calculated according to Bayes' theory:

$$P(w|\mathbf{y}) = \frac{\sum_{w=1}^{W} \pi_w \prod_{i=1}^{I} \pi_{iy^{(i)}|w}}{\sum_{w'=1}^{W} \pi_{w'} \prod_{i=1}^{I} \pi_{iy^{(i)}w'}}. \tag{2}$$

The probabilities wth latent class have to sum to unity i.e.:

$$\sum_{w=1}^{W} \pi_w = 1, \quad \sum_{y^{(i)}=1}^{m^{(i)}} \pi_{iy^{(i)}|w} = 1. \tag{3}$$

As mentioned ealier, the most popular method to estimate parameters of the latent class model is to maximize likelihood of the log:

$$\ln L = \sum_{i=1}^{n} \ln \sum_{w=1}^{W} \pi_w \prod_{i=1}^{I} \pi_{iy^{(i)}|w} \tag{4}$$

Fitting the function into data is realized by using EM algorithm (Dempster, Laird and Rubin, 1977).

The latent class regression model generalizes the basic latent class model by permitting the inclusion of covariates to predict individuals latent class membership (Dayton and Macready, 1988; Hagenaars and McCutcheon, 2002). The latent class model can usually be reformulated as a the latent class regression model.

Let β_w denote the vector of coefficients corresponding to *w*th latent class and X_i represent the observed covariates for variable *i*. The first class is used as reference, $\beta_1 = 0$ is fixed by definition. The resulting LCR has the form:

$$\ln(\pi_{wi} / \pi_{1i}) = \beta_w X_i + \beta_{0w} \quad (w = 2,...,W) \tag{5}$$

where β_w and β_{0w} are the class-specific regression coefficients.

Latent Class Models show specific properties, which are significant from application point of view, namely:

– make identification of classes possible on the basis of observed variables or dependent variables,

– contain one categorical latent class (number of categories is equal to number of class),

– the basis of classification of variables are estimated on the basis of model posterior class membership probability,

– observed variables can be measured mainly on nominal scale,

– to model which can include covariaties.

III. APPLICATION

The illustration of the latent class model using the dataset *German credit*, can be freely obtained from UCI Machine Learning Repository (Blake and other, 1998). In this dataset there are 1000 individuals, that are described by 21 variables. This dataset classifies people described by a set of attributes as good or bad credit risks. At the conference SKAD 2007 it was shown how using the classification trees we can research on the influence of observed variables on the dependent variable i.e. the credit application acceptance or non-acceptance. As a result of making analysis five variables having the significant influence on the dependent variables i.e. status of existing checking account, duration in months, credit history, purpose of a credit and savings account/bonds were defined. In this example the covariate variable is the present employment. Five independent variables were taken to this analysis and the variable describes the status of application i.e. good and bad.

Table 1. List of variables used in the research

Name of variable	Categories of variable
Status of existing checking account (A1)	A11 <0 DM
	A12 (0,200> DM
	A13 ≥200 DEM/salary assignments for at least 1 year
	A14 no checking account
Duration in months * (A2)	A21 <21month
	A22 over 21 months
Credit history (A3)	A30 no credits taken/ all credits paid back duly
	A31 all credits at this bank paid back duly
	A32 existing credits paid back duly till now
	A33 delay in paying off in the past
	A34 critical account/ other credits existing (not at this bank)
Purpose of credit (A4)	A40 car (new)
	A41 car (used)
	A42 furniture/equipment
	A43 radio/television
	A44 domestic appliances
	A45 repairs
	A46 education
	A47 vacation
	A48 retraining
	A49 business
	A410 others
Savings account/bonds (A6)	A61 < 100 DM
	A62 <100,500) DM
	A63 <500,1000) DM
	A64 ≥ 1000 DM
	A65 unknown/ no savings account
Status of application (A21)	A211 bad
	A212 good
covariates: job	A71 unemployed
	A72 < 1 year
	A73 <1,4) years
	A74 <4,7) years
	A75 > 7 years

* variable was discritized.

Source: UCI Machine Learning Repository (1998).

In poLCA package there are two information criteria available: Akaike information criterion (AIC) and Bayesian information criterion (BIC), which are especially useful in comparing models. The most widely used in the latent class analysis is the BIC statistic. A model with a lower BIC value is preferred to a model with a higher BIC value. A more general definition of BIC is based on the log-likelihood and the number of parameters.

It is necessary to mention that the AIC criterion has tendency to choose models, which is too complex even when the simple is large. Whereas the BIC criterion chooses models less complex, i.e. with less number of parameters.

The BIC statistic suggests that the 3-class model is preferred over the 4-class model.

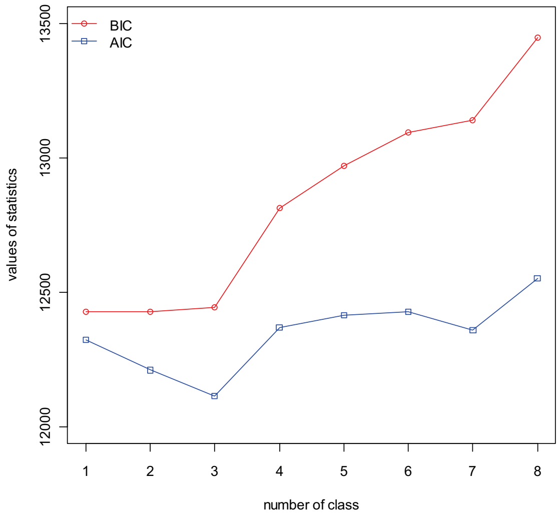

Fig. 1. Values of AIC and BIC criteria for number of class

Source: own research

By examining the estimated class-conditional response probabilities, we confirm model with three class according to BIC crierion.

```
$A1
           Pr(1)  Pr(2)  Pr(3)  Pr(4)
class 1:  0.5363 0.3122 0.0636 0.0879
class 2:  0.0784 0.1767 0.0705 0.6744
class 3:  0.1925 0.5256 0.0272 0.2547

$A2
           Pr(1)  Pr(2)
class 1:  0.5984 0.4016
class 2:  0.6114 0.3886
class 3:  0.1343 0.8657
```

```
$A3
           Pr(1)  Pr(2)  Pr(3)  Pr(4)  Pr(5)
class 1:  0.0396 0.0705 0.6513 0.0388 0.1998
class 2:  0.0084 0.0164 0.4969 0.0696 0.4086
class 3:  0.1833 0.1165 0.2349 0.3502 0.1151

$A4
           Pr(1)  Pr(2)  Pr(3)  Pr(4)  Pr(5)  Pr(6)  Pr(7)
Pr(8)  Pr(9) Pr(10)
class 1:  0.2997 0.0603 0.0086 0.2840 0.2248 0.0178 0.0319
0.0525 0.0073 0.0129
class 2:  0.1980 0.1413 0.0000 0.1374 0.3659 0.0100 0.0181
0.0461 0.0124 0.0710
class 3:  0.1555 0.0872 0.0781 0.0000 0.0963 0.0000 0.0034
0.0583 0.0000 0.5211

$A6
           Pr(1)  Pr(2) Pr(3)  Pr(4)  Pr(5)
class 1:  0.7764 0.0779 0.028 0.0246 0.0931
class 2:  0.4666 0.0941 0.100 0.0763 0.2629
class 3:  0.5811 0.2345 0.025 0.0065 0.1531

$A21
           Pr(1)  Pr(2)
class 1:  0.5531 0.4469
class 2:  0.0377 0.9623
class 3:  0.5522 0.4478

Estimated class population shares
 0.3998 0.4908 0.1094

Predicted class memberships (by modal posterior prob.)
 0.411 0.495 0.094
```

The first and third class is described by the borrowers, who do not receive the credit. The first class (39,98%) represents the borrowers with the status of existing checking account below 0 DM (53,63%), the time duration below 21 months (59,84%). 65,13% of the people belong to the class ‘existing credits paid back duly till now’ and their purpose of credit is buying a used car.

Up to 10,94% of the credit applications belong to the third class and this class usually characterizes the people whose status of existing checking account is in range 0-200 DM (52,56%). Morover, the people would like to allocate credit for other purpose (52,11%) and the duration time exceeds 21 months (86,57%). It is worth stressing that these people delayed in paying off in the past (35,02%).

The second class, the most numerous (49,08%), represents the people whose purpose of applying for credit is to buy domestic appliances (36,59%). These borrowers have no account *a vista* in the bank where they try to get the credit.

It turned out that the variable describing savings account/bonds do not disperse these classes.

In addition to the information for the basic model, the poLCA output also includes the estimated coefficient on the covariates in the latent class regression model, and their standard errors.

```
=========================================================
Fit for 3 latent classes:
=========================================================
2 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -1.30823     0.31021   -4.217     0
A7              0.45036     0.08098    5.561         0
=========================================================
3 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -1.95791     0.51376   -3.811         0
A7              0.20749     0.13626    1.523         0
=========================================================
number of observations: 1000
number of estimated parameters: 70
residual degrees of freedom: 930
maximum log-likelihood: -6517.552
AIC(3): 13175.10
BIC(3): 13518.65
X^2(3): 5295.757 (Chi-square goodness of fit)
```

To interpret the estimated generalized logit coefficients, we can calculate the prior probability of class membership. By using the follow commands in R software we can plot predicted values of the present employment:

```
data <- read.csv2("german.csv", header=TRUE)
names<- cbind(A1,A2,A3,A4,A6)~A7
compute_class <- poLCA(names, data, nclass=3)
pidmat <- cbind(1,c(1:5))
exb <- exp(pidmat %*% compute_class$coeff)
c<-(cbind(1,exb)/(1+rowSums(exb)))
matplot(c(1:5),c,ylim=c(0,1),pch=1, type="o", main=
"Probability of latent lass membership", ylab="",
xlab="present employment ")
text(4,0.38,"CLASS I")
text(4.2,0.15,"CLASS II")
text(3.5,0.6,"CLASS III")
```

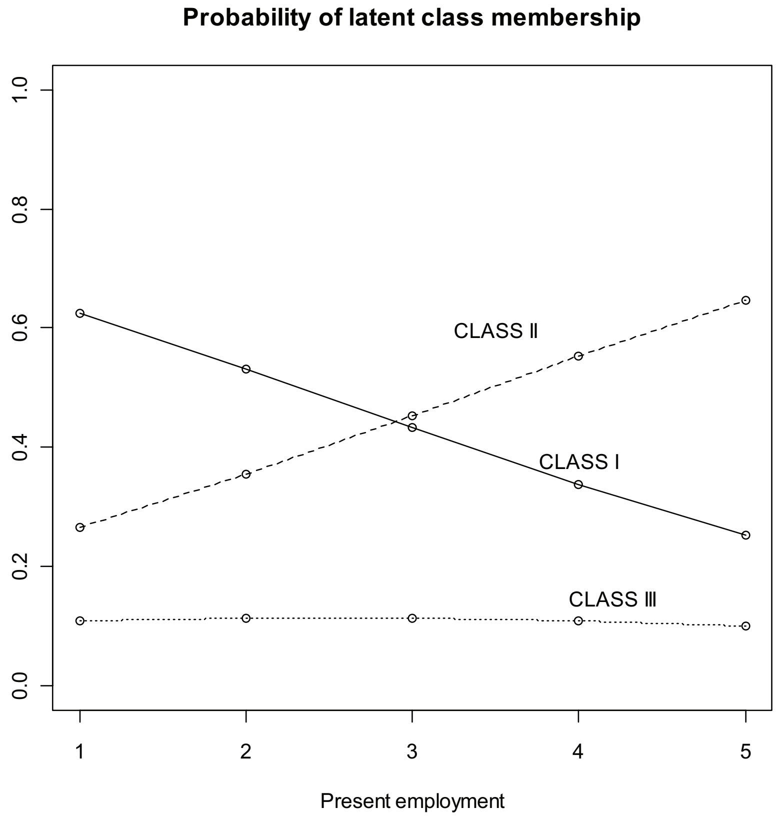

Fig. 2. Probability of latent class memebership of the present employment

Source: own research

The odds ratio:

```
               class II/I  class III/I
unemployed     0.4240642   0.1737010
< 1 year       0.6653065   0.2137545
<1,4) years          1.0437869   0.2630438
<4,7) years          1.6375777   0.3236986
>=7 years            2.5691650   0.3983398
```

The person, who has worked for at least seven years has twice the odds of being in class 2 versus class 1, the unemployment has 82,6% less the odds of being in class 3 in comparison to class 1.

We can also show results in the form of barplots for every class:

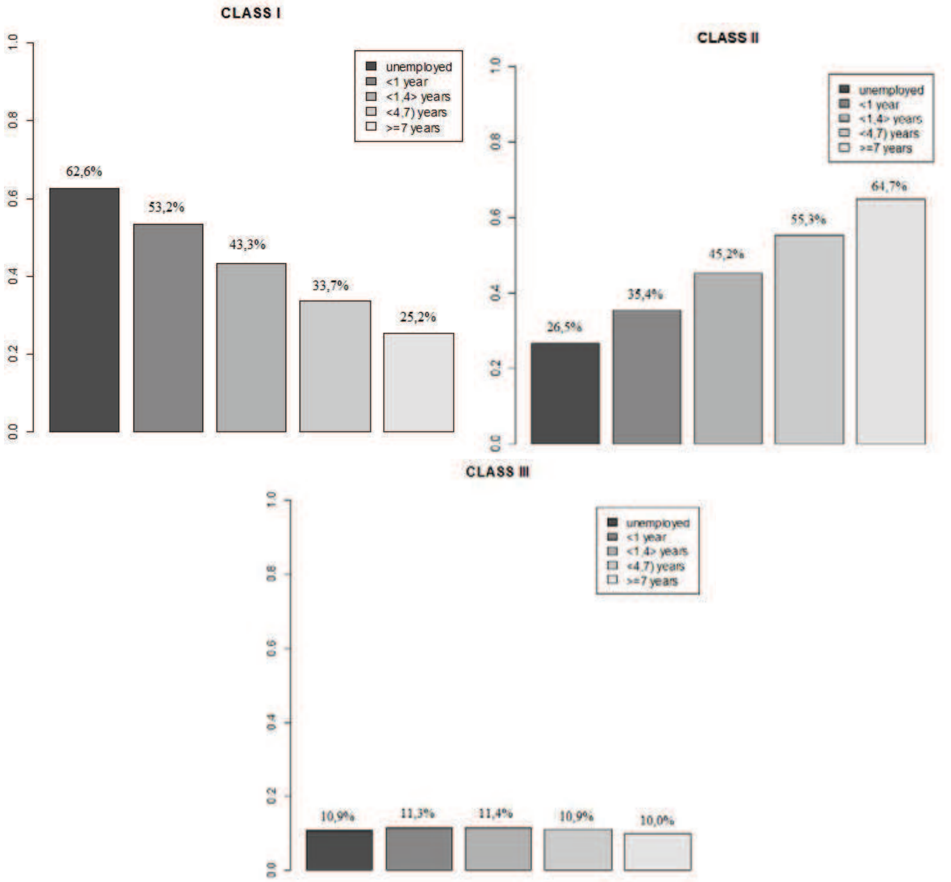

Fig. 3. Probability of latent class memebership of the present employment by using barplots
Source: Own research

IV. SUMMARY

The latent class regression model allows to show the influence of the covariate variable (present employment) on observed variables and predict membership of potential borrowers in the class. By means of the latent class model the unknown structure of classes was identified (separating three classes).

The open problem still remains the construction of poLCA package that allows to estimate only the odds ratio of individual classes in relation to the first class.

REFERENCES

Akaike H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control,* 19 (6), 716–723.

Blake C., Keogh E., Merz C.J.(1998). *UCI Repository of Machine Learning databases*, Department of Information and Computer Science, University of California, Irvine, www.ics.uci.edu /~mlearn/MLRepository.html.

Bozdogan H. (1987), Model selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions*, Psychometrica,* 52, 345-370.

Dayton, C. Mitchell and George B. Macready (1988), Concomitant Variable Latent Class Models, *Journal of the American Statistical Association* 83(401), 173-178.

Dempster A. P., Laird N. M. and Rubin D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm (With Discussion), *Journal of the Royal Statistical Society, Series B,* 39, 1-38.

Goodman L. (1974), Exploratory Latent Structure Analysis using both Identifiable and Unidentifiable Models, *Biometrika,* 61, 315-331.

Hagenaars J. A. and McCutcheon A. L. (2002). *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.

Lazarsfeld, P.F., Henry N.W. (1968), *Latent structure analysis*, Boston: Houghton Mill.

Linzer D.A., Lewis J. (2006), poLCA: Polytomous Variable Latent Class Analysis, http://userwww.service.emory.edu/~dlinzer/poLCA

McCutcheon A. L. (1987). *Latent Class Analysis*. Newbury Park: SAGE Publications.

Iwona Kasprzyk

MODELE KLAS UKRYTYCH Z WYKORZYSTANIEM PROGRAMU R

Artykuł przedstawia tradycyjne podejście do analizy klas ukrytych oraz analizę regresji klas ukrytych. Modele klas ukrytych zostały zaproponowane przez Lazarsfeld, Henry (1968). W tego rodzaju modelach zakłada się, że wszystkie zmienne są niezależne. Natomiast analiza regresji klas ukrytych jest uogólnieniem modeli klas ukrytych i pozwala poprzez włączenie tzw. zmiennych kontrolowanych na predykcję przynależności kategorii zmiennej do klasy ukrytej.

Obliczenia zostały wykonane za pomocą pakietu `poLCA` w programie R.