LATENT CLASS ANALYSIS IN THE EVALUATION OF ITEMS IN SURVEY RESEARCH

Justyna Brzezińska

University of Economics in Katowice, Katowice, Poland
e-mail: justyna.brzezinska@ue.katowice.pl

ORCID: 0000-0002-1311-1020

Abstract: Latent class analysis has been widely used in the measurement models. Models based on latent variables have a wide range of applications in the presence of repeated ob-servations, longitudinal data, and multilevel data. In this paper we present and apply log-linear analysis as a method for the analysis of multi-way tables. We also present a latent variable model based on a variable that is not directly observed. The basic model postulates an underlying categorical latent variable; within any category of the latent variable the manifest or observed categorical variables are assumed independent of one another (axiom of conditional independence). In this paper we present the results of a survey research based on categorical data and the author`s questionnaire. We present the results of the latent class analysis in the classification of respondents into clusters characterized by similar attitudes and features in economic research. We also conduct a prior log-linear analysis for a multi-way contingency table. All the calculations are conducted in R.

Keywords: latent class analysis, log-linear models, item response, survey questions.

1. Introduction and historical background

The measurement of human properties has been a long-lasting quest that originated in the 19th century and successively continued to this very day. Significant inroads to the development of the interest in the area of measurement may be traced back to the beginning of the twentieth century. Binet and Simon [1905], were the first to develop and use what we would nowadays call a standardized intelligence test, and in the same era Charles Spearman [1904] developed the concepts and methodology of what would later be called classical test theory (CTT) and factor analysis (FA). Although CTT was the dominant statistical test model in the first half of the 20th century [Guilford 1936; Gulliksen 1950], other developments were also taking place. In England, Walker [1931] set the stage for what later was to become known as Guttman

(1944, 1950) scaling, by introducing the idea that if a respondent can answer a harder question correctly then he or she should be able to answer easier questions on the same topic correctly as well. Walker also introduced the idea of a quantitative index measuring deviation from this deterministic model in real data [Loevinger 1948]. This was a deterministic approach without a random error component to the analysis of data collected by a set of items that are assumed to measure one psychological property in common. A little earlier, in the 1920s, Thurstone [1927] developed his statistical measurement method of comparative judgment. Thurstone's work may be viewed as the most important probabilistic predecessor of item response theory (IRT).

In the 1970s, item response theory became the dominant topic for study by measurement specialists, but the genesis of item response theory (IRT) can be traced back to the mid-thirties and early forties. In fact, the term *Item Char-acteristic Curve* (ICC), which is one of the main IRT concepts, can be attributed to Ledyard Tucker in 1946. Despite these early research efforts, interest in item re-sponse theory lay dormant until the late 1960s and took a backseat to the emerging development of strong true score theory. While true score theory developed rapidly and drew the attention of leading psychometricians, the problems and weaknesses inherent in its formulation began to raise concerns. Such problems as the lack of invariance of item parameters across examinee groups, and the inadequacy of clas-sical test procedures to detect item bias or to provide a sound basis for measure-ment in "tailored testing", gave rise to a resurgence of interest in item response theory. Impetus for the development of item response theory as we now know it was provided by Frederic M. Lord through his pioneering works [Lord 1952, 1953]. The progress in the fifties was painstakingly slow due to the math-ematical complexity of the topic and the nonexistence of computer programs. In-terest in Lord's work spread quickly, as evidenced in the increase in publications in this area. Allen Birnbaum wrote a series of technical reports on logistic test models and model parameter estimation in 1957 and 1958. George Rasch [1960] published his book proposing several models for item responses.

2. Analysis of cross-classified tables

Log-linear models are a standard tool to analyze structures of dependency in multi-way contingency tables. The criteria to be analyzed are the expected cell frequencies as a function of all the variables in the survey. There are several types of log-linear models depending on the number of variables and interactions included. A satu-rated model for a three-way $H \times J \times K$ ($h = 1, ..., H$, $j = 1, ..., J$, $k = 1, ..., K$) table including all the possible effects in multiplicative form for three variables is given as [Christensen 1997]:

$$m_{hjk} = \eta \, \tau_h^X \, \tau_j^Y \, \tau_k^Z \, \tau_{hj}^{XY} \, \tau_{hk}^{XZ} \, \tau_{jk}^{YZ} \, \tau_{hjk}^{XYZ} \, , \qquad (1)$$

where: m_{hjk} – expected cell counts for the contingency table.

By taking the natural logarithms we have an additive equation given as:

$$\log\!\left(m_{hjk}\right) = \lambda + \lambda_h^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hj}^{XY} + \lambda_{hk}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{hjk}^{XYZ}, \tag{2}$$

where: λ represents an overall effect or a constant, λ_h^X, λ_j^Y, λ_k^Z represents the effect of the row, column and layer variable X, Y, Z, λ_{hj}^{XY}, λ_{hk}^{XZ}, λ_{jk}^{YZ} represents the interaction between two variables XY, XZ, YZ and λ_{hjk}^{XYZ} is an interaction term between XYZ.

A saturated model reproduces perfectly the observed cell frequencies through the theoretical frequencies and such a model is meaningless since the aim is to find a more parsimonious model with less parameters. In order to find the best model from a set of possible models, some additional measures have to be considered. A rule of thumb to determine the degrees of freedom df is number of cells – number of free parameters [Agresti 2002]. The starting point is the saturated model. Thus the aim of the researcher is to find a reduced model. A reduced model is a more parsimonious model with fewer parameters and thus fewer dependencies and effects. The hierarchy principle reveals that a parameter of a lower order cannot be removed when there is still a parameter of higher order that concerns at least one of the same variable.

A unique set of ML estimates for every cell can be derived from the sufficient statistics alone with the use of iterative proportional fitting [Deming, Stephan 1940]. The main goal of log-linear analysis is to find the smallest model that fits the data. The overall goodness-of-fit of a model is assessed by comparing the expected frequencies to the observed cell frequencies for each model. The goodness of fit of a log-linear model is usually tested using either the Pearson chi-square test statistic or the likelihood ratio statistic:

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{hjk} \ln\!\left(\frac{n_{hjk}}{\hat{n}_{hjk}}\right), \tag{3}$$

where: n_{hjk} – observed cell counts for a three-way table.

Therefore, larger G^2 values indicate that the model does not fit the data well and thus the model should be rejected. This strategy is the opposite of the usual chi-square test of independence, where we seek to reject the null hypothesis of no association. Yet in trying to find the best fitting log-linear model to describe the cross-table we hope to accept the hypothesized model, hence we want to find a low G^2 value relative to df [Knoke, Burke 1980]. The likelihood ratio can also be used to compare an overall model within a smaller, nested model (i.e. a saturated model with one interaction or main effect dropped to assess the importance of that term). The equation is $\Delta G^2 = G_2^2 - G_1^2$ with: $\Delta df = df_2 - df_1$, where 2 is the nested model,

1 is the higher parameterized model, and are degrees of freedom for model 1 and model 2. If the comparison statistic is not significant, then the nested model is not significantly worse than the saturated model.

In order to find the best model from a set of possible models, additional measurements should be considered. The Akaike information criterion [Akaike 1973] refers to the information contained in a statistical model according to equation:

$$AIC = G^2 - 2df \, . \tag{4}$$

Another information measurement is the Bayesian information criterion [Raftery 1986]:

$$BIC = G^2 - df \cdot \ln n \, , \tag{5}$$

where: n – total sample size. The model that minimizes AIC and BIC will be chosen.

3. Latent class analysis

When a researcher wishes to obtain a better understanding of the relationship between two observed dichotomous variables (e.g. variables A and B) it is advisable to consider another variable (dichotomous or polytomous, let us say, variable X) that can explain this relationship. We say that variable X explains the relationship between variables A and B if this relationship disappears when the level of variable X is held constant.

The methodology and basic notions of latent class analysis were developed by Paul Lazarsfeld during the early 1950s [Lazarsfeld 1950a; 1950b; Lazarsfeld and Henry 1968]. The credit for feasible and flexible algorithms for testing the validity of a wide variety of latent class models and estimating their parameters is due to Goodman [1974a; 1974b] and Haberman [1979]. An introduction into these more recent developments are provided by Clogg [1981], Formann [1985], Hagenaars [1990] and McCutcheon [1987].

A standard latent class analysis examines the cross-classification of several observed categorical variables with one or more latent categorical variable(s). There are two kinds of variables in latent class models. First, there are the directly observed manifest variables. These act as an indicator for the second kind, which are not directly observed latent variables. In the standard latent class model, both manifest and latent variables are treated as categorical nominal-level variables. Furthermore, it is assumed that the scores on the manifest variables do not influence each other directly and that the outside influence and measurement errors affecting the score on a particular manifest item are independent of the outside influence affecting other manifest items. All the items have in common is their being an indicator of the same latent variable. The manifest variables are correlated with each other, but this correlation disappears when the latent variable is held constant. This is the assumption of local independence (axiom of local independence), an assumption fundamental to all latent structure models.

Let $\mathbf{x}^{`} = \left[ x_1, x_2, ..., x_p \right]$ represent the manifest variables, and $\mathbf{y}^{`} = \left[y_1, y_2, ..., y_m \right]$ the latent variables. The number of latent variables, m, is usually much less than the number of manifest variables, p, since a latent variable model can be considered in some sense a data reduction method which reduces the full set of manifest observations to a few latent values. In essence all latent variable models assume that $x_1, x_2, ..., x_p$ have a joint probability distribution conditional on $y_1, y_2, ..., y_m$; this we shall denote by $\varphi(\mathbf{x}|\mathbf{y})$. If the manifest variables are all continuous, ϕ is a density function, but if they are discrete it is a set of probabilities. If the density function of y is $h(y)$ then the unconditional density of x is defined as:

$$f(x) = \int \phi(\mathbf{x}|\mathbf{y}) h(\mathbf{y}) d\mathbf{y}.$$ (6)

In general terms it is the density functions, ϕ and h, that we would like to infer from the known or assumed density of x, in order to discover how the manifest variables depend upon the latent ones. However, it is clearly impossible to infer ϕ and h uniquely from f unless some assumptions are made about their form.

The crucial assumption of latent variable models is that of conditional independence, which states that given the values of the latent variables, the manifest variables are independent of one another. This can be expressed as follows:

$$\phi(\mathbf{x}|\mathbf{y}) = \phi(x_1|\mathbf{y}) \phi(x_2|\mathbf{y}) ... \phi(x_p|\mathbf{y}).$$ (7)

The assumption of conditional independence implies that it is the latent variables which produce the observed relationships amongst the manifest variables. The observed interdependence amongst the manifest variables is due to their common dependence on the latent variables and that once these have been determined, the behavior of the manifest variables is essentially random.

In latent class analysis the latent structure is usually unknown and it is the analyst's goal to estimate the parameters of the latent class model of interest. There are usually several competing models of interest that must be compared in terms of whether or not the given model fits the observed data within acceptable limits.

In this paper we present the most popular and known methods that can be used in this step of latent class analysis. Based on a total of n observations, the expected frequency for the s-th response vector is defined as:

$$\hat{n}_s = n \cdot \hat{P}(\mathbf{y}_s),$$ (8)

where $\hat{n}_s$ is the corresponding observed frequency for the response vector.

The goodness of fit of a given latent class model to the observed data can be assessed using a chi-square test based on observed and expected frequencies. We can use Pearson's chi-square statistic based on the differences between the observed and the expected frequencies:

$$\chi^2 = \sum_{s=1}^{2^v} \frac{(n_s - \hat{n}_s)}{\hat{n}_s}.$$ (9)

Another well known measurement for testing goodness-of-fit for latent class model is the likelihood-ratio statistic:

$$G^2 = 2\sum_{s=1}^{2^v} n_s \cdot \log_e\left(\frac{n_s}{\hat{n}_s}\right).$$ (10)

Another group of criteria used for model selection are information criteria. The Akaike information criterion (*AIC*) [Akaike 1973] defined as:

$$\log_e(L_h) = \sum_{s=1}^{2^v} n_s \cdot \log_e\left[P_h(\mathbf{Y}_s)\right],$$ (11)

where: $P_h(\mathbf{Y}_s)$ is the probability for an observed response vector based on maximum-likelihood estimates for the *h*-th model. Another Akaike information criterion is defined as:

$$AIC_h = -2\log_e(L_h) + 2m_h,$$ (12)

where m_h is the number of independent parameters that are estimated when fitting the *h*-th model to the data. An alternative, simpler approach is based on the likelihood-ratio chi-square statistics. For *H* different latent class models fitted to the same data, let G^2 be the chi-square value, and v_h be corresponding degrees of freedom. An alternative Akaike information criterion for the *h*-th model is:

$$AIC_h^* = G_h^2 - 2v_h.$$ (13)

AIC_h and AIC_h^* differ only by a constant amount that involved the sample size *n*, and the number of observed response vectors (2^v).

For model selection we can also apply the Bayesian information criterion (*BIC*) [Schwartz 1978]:

$$BIC_h = -2\log_e(L_h) + \log_e(N) \cdot m_h,$$ (14)

or the modification BIC:

$$BIC_h^* = G_h^2 - \log_e(N) \cdot m_h.$$ (15)

The minimum value of information criterion will be chosen indicating the model that fits best.

4. Application and item analysis in R

In this paper we apply item analysis for the preferences and attitude measurement based on the author`s own survey study. All the calculations presented in this paper are conducted in **R**, which can be downloaded from www.r-project.org. **R** Development Core Team has become a successful language for statistical computing and graphics. Today, it is often considered as the *lingua franca* of all computational statistical tools. This means it can be used, distributed, copied and shared without restrictions. It enables and encourages learning, collaboration, adaptation and inspection of the code base and allows reproducible statistical analyses. This software is widely used in the academic environment providing full statistical analysis.

Our own survey in 2018, was conducted using a questionnaire based on 26 questions on nominal scale (answers: Yes or No). There were 242 respondents who were students at the University of Economics in Katowice. 74.38% of respondents were female, and 25.62% were male at average age 21-24 (91.90%).

The correlation coefficients between all the possible pairs of questions were computed, and the questions with the highest correlation values were chosen for the analysis:

- *do you want to be successful in life* (X_{19}) and *is the quality of your future life important for you* (X_{21}) with $r = 0.628$,
- *do you want to study in future* (X_{23}) and *do you want to improve your professional skills* (X_{24}) with $r = 0.778$.

These items were the main constructs in the measurement of attitudes towards ability in their professional future.

Firstly, we apply log-linear analysis to select the categorical model presenting the relation between the categories of variables that are analyzed. For the building process we use symbols in bracket notation instead of variable names. We use [G] for gender, [S] for success in life, [Q] quality of life, [C] for upgrading competences and future studies. We build models using the hierarchy principle and we present the goodness of fit for models using (3), (4) and (5) coefficients with corresponding p-value and degrees of freedom (*df*).

From the analysis of log-linear models we should choose a model that is not too complex, and at the same time has a satisfactory level of fit. The model that contains few variables and has sufficient information criteria is model [QSC] three-way having interactions. In this model we have also all lower-order pairs of interactions and single variables. This model is difficult in its interpretation, also the interpretation of estimates parameters of higher-order will be difficult. That is why it is always advisable to consider models easy to interprete and we conduct a latent class analysis.

Now we conduct a latent class analysis. In the first stem we test the goodness- of-fit for models with 2, 3 and 4 latent classes. We apply several coefficients for model selection and the corresponding log likelihood value. The results of the analysis are presented in Table 2.

Table 1. Log-linear models and goodness-of-fit coefficients for some models for a four-way table

Model	df	G^2	p	AIC	BIC
[GQSC]	0	0	1	0	0
[GQC] [GQS] [QSC]	8	0	1	142.8203	234.3135
[GQC] [GQS]	16	0.6415	0.9999	128.4618	204.0432
[GQC] [QSC]	8	0	1	142.8203	234.3135
[GQC]	36	455.0089	0	541.8292	577.6309
[QSC]	27	100.1566	0.00011	204.9769	258.6795
[GQ] [GC] [QC]	40	4535.010	0	533.8306	561.6764
[GQ] [GC]	44	474.5241	0	533.8306	561.6764
[GQ] [QC]	40	455.0103	0	545.3444	565.2342
[GC] [QC]	40	455.0103	0	533.8306	561.6764
[GQ] [G]	44	474.5241	0	545.3444	565.2342

Source: own calculations in R.

Table 2. Log-linear models and goodness-of-fit coefficients

Latent clases	AIC	BIC	G^2	χ^2	df
2	644.9676	704.2796	57.2200	209.2463	63
3	638.8415	729.5539	33.0939	76.9919	54
4	656.0123	778.1251	32.2647	74.0101	45

Source: own calculations in R.

As latent class models with lower AIC and BIC values are preferred to those with higher values for this criteria, we consider a model with 3 latent classes as the latent class model fitting best. For this model also other criteria based on chi-square coefficient are the minimum.

Now we present graphically the probability of 3 latent class membership. Figure 1 presents the probability of 4 latent class membership and probabilities for variables that were examined in each latent class.

Population share for class 1 is 0.05, for class 2 it is 0.888 and for class 3, 0.062. In Figure 1 each group of red bars represents the conditional probabilities, by latent class, of being rated by respondents of the three answers: 1: Yes, 2: No. 3: 'I don`t know', with taller bars corresponding to the conditional probabilities closer to 1 of a positive rating.

We can see that in class 1 the highest bars can be seen for X_{19} and X_{21} were obtained mostly for answer 1: 'Yes', whereas for X_{23} and X_{24} were obtained for answer 2: 'No'. Total class population share for class 1 is 0.05. This means that in class 1 there are respondents who want to be successful in life and for whom the future is important, but they are not interested in future studies nor in improving their competences in the future.
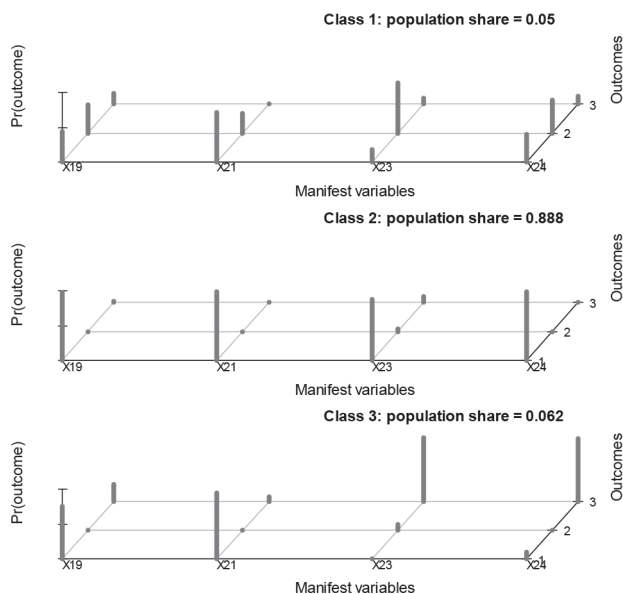
Fig. 1. Probability of 3 latent classes

Source: own calculations.

In class 2 we can see that most of respondents answered mostly 'Yes' to all of the questions. This refers to the high red bars for all four variables: X_{19}, X_{21}, X_{23} and X_{24}. This means that this group is future and success oriented. The total class population share for class 2 is 0.888.

In class 3 we observe most of respondents answered Yes for questions X_{19} and X_{21}, whereas the 'don`t know' answer was for X_{23} and X_{24}. This group is strongly oriented towards success and future but doesn not know about their preferences about upgrading competences and future studies.

5. Conclusions

In this paper we presented the results from the author`s study on the attitude towards ability in their professional future among young Polish students. From the primary set of variables only a few were chosen for the analysis. Log-linear analysis and latent class analysis were applied. As log-linear models are difficult to interpret when they include several variables, we apply latent class analysis for further studies. Several latent models were tested to check the goodness of fit using information criteria (AIC, BIC) and measurements based on the chi-square coefficient (χ^2, G^2). The 4 latent class model was chosen and presented using conditional probabilities for each latent class, as well as providing a graphical presentation of membership to latent classes. From the analysis, three latent classes were separated with latent

class membership probability. The first class were a group of respondents who are interested in the future and success but not in their studies and further upgrading their competences. The second group with the highest probability membership are people strongly oriented in upgrading their competences and further studies.

The last group are respondents interested in achieving life success in future but not sure about future upgrading their professional competences. This study proved that item analysis is a powerful and useful tool in the analysis of preferences and attitudes in questionnaire research. R software provides `poLCA` package for the analysis of latent class which should encourage researchers to conduct their own studies using this software.

Bibliography

Agresti A., 2002, *Categorical Data Analysis*, Wiley & Sons, Hoboken, New Jersey.

Akaike H., 1973, *Information Theory and an Extension of the Maximum Likelihood Principle*, Proceedings of the 2nd International Symposium on Information, Petrow B.N., Czaki F., Akademiai Kiado, Budapest.

Binet A., Simon T., 1905, *Méthodes nouvelles pour le diagnostique du niveau intellectuel des anormaux*, Annee Psychol., 11, pp. 245-366.

Christensen R., 1997, *Log-linear Models And Logistic Regression*, Springer-Verlag, New York.

Clogg C.C., 1981, *Latent structures models for mobility*, American Journal of Sociology, 86, pp. 863-868.

Deming W., Stephan F., 1940, *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*, Annals of Mathematical Statistics, 11, 4, pp. 427-444.

Formann A.K., 1985, *Constrained latent class analysis: Theory and applications*, British Journal of Mathematical and Statistical Psychology, 38, pp. 89-111.

Goodman L.A., 1974a, *Exploratory latent structure analysis using both identifiable and unidentifiable models*, Biometrica, 61, pp. 215-231.

Goodman L.A., 1974b, *The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach*, American Journal of Sociology, 79, pp. 1179-1259.

Guilford J.P., 1936, *Psychometric Methods*, McGraw-Hill, New York.

Gulliksen H., 1950, *Theory of Mental Tests*, Erlbaum, Hillsdale, NJ.

Haberman S.J., 1974, *The Analysis of Frequency Data*, University of Chicago Press, Chicago.

Haberman S.J., 1979, *Analysis of Qualitative Data, Vol. 2, New Developments*, Academic Press, New York.

Hagenaars J.A., 1990, *Categorical Longitudinal Data – Loglinear Analysis of Panel, Trend and Cohort Data*, Sage, Newbury Park.

Knoke D., Burke P.J., 1980, *Log-linear models*, Sage University Paper Series on Quantitative Applications in the Social Science, series no. 07-020, Sage, Beverly Hills and London.

Lazarsfeld P.F., 1950a, *The interpretation and mathematical foundation of latent class structure analysis*, Souffer S. (ed.), *Measurement and Prediction*, Princeton University Press, Princeton, NJ.

Lazarsfeld P.F., 1950b, *The logical and mathematical foundation of latent structure analysis*, Souffer S. (ed.), *Measurement and Prediction*, Princeton University Press, Princeton, NJ.

Lazarsfeld P.F., Henry N.W., 1968, *Latent Structure Analysis*, Houghton Mifflin, Boston.

Loevinger J., 1948, *The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis*, Psychological Bulletin, 45, pp. 507-530.

Lord F.M., 1952, *The relation of the reliability of multiple-choice tests to the distribution of item difficulties*, Psychometrika, 17(2), pp. 181-193.

Lord F.M., 1953, *The relation of test score to the trait underlying the test*, Educational and Psychological Measurement, 13, pp. 517-548.

McCutcheon A.L., 1987, *Latent Class Analysis*, Sage University Paper. Sage Publications, Newbury Park.

Raftery A.E., 1986, *Choosing models for cross-classification*, Amer. Sociol. Rev., 51, pp. 145-146.

Rasch G., 1960, *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen.

Schwartz G., 1978, *Estimating the dimensions of a model*, Annals of Statistics 6, pp. 461-464.

Spearman C., 1904, *The proof and measurement of association between two things*, The American Journal of Psychology, 15(1), pp. 72-101.

Thurstone L.L., 1927, *A law of comparative judgment*, Psychological Review 34, pp. 273-286.

Walker D.A., 1931, *Answer pattern and score scatter in tests and examinations*, British Journal of Psychology, 22, pp. 73-86.

ANALIZA KLAS UKRYTYCH W OCENIE POZYCJI TESTOWYCH W BADANIACH SONDAŻOWYCH

Streszczenie: Modele oparte na zmiennych ukrytych znajdują szerokie zastosowanie w przypadku powtarzających się obserwacji, danych panelowych czy też modeli wielopoziomowych. Jako pierwszą zastosowano analizę log-liniową, która jest trudna do interpretacji w przypadku modeli z wieloma zmiennymi. W kolejnej części pracy zastosowano analizę klas ukrytych, która zawiera zmienne ukryte. Analiza klas ukrytych oparta jest na dwóch założeniach. Pierwsze mówi o tym, że populacja składa się z rozłącznych (*mutually exclusive*) i spójnych (*exhaustive*) jednorodnych podpopulacji, które łącznie tworzą klasę ukrytą; drugie natomiast nazywane jest warunkiem lub aksjomatem lokalnej niezależności (*local independence assumption*), zgodnie z którym związek między zmiennymi obserwowalnymi zależy od relacji pomiędzy zmiennymi obserwowalnymi a zmiennymi ukrytymi. W artykule zaprezentowano autorskie wyniki badań nt. postaw studentów stojących u progu kariery zawodowej. Ponadto przedstawiono zarówno teoretyczne, jak i praktyczne aspekty analizy klas ukrytych.

Słowa kluczowe: analiza klas ukrytych, analiza logarytmiczno-liniowa, pozycje testowe, badania sondażowe.