Journal of Business Economics and Management
2007, Vol VIII, No 4, 253–258
www.jbem.lt

ISSN 1611-1699

APPLICATION OF LOGIT REGRESSION MODELS FOR THE IDENTIFICATION OF MARKET SEGMENTS

Marija Burinskienė[1] and Vitalija Rudzkienė[2]

[1]Urban Engineering Department, Vilnius Gediminas Technical University, Saulėtekio al. 11, 10223 Vilnius, Lithuania, e-mail: marbur@ap.vgtu.lt (corresponding author)
[2]Department of Informatics and Statistics, Mykolas Romeris University, Ateities g. 20, 11104 Vilnius, Lithuania, e-mail: vital@mruni.lt

Received 28 January 2007; accepted 3 March 2007

Abstract. A success of the currently implemented projects and measures is determined not only by the urgency and soundness of idea and the size of the budget, but also by the direction of resources to those users and organizations, from which the largest return could be expected, by the public opinion about the current business and a success of project presentation in various forms of mass media. For identification of target market, to which the business strategy will be directed, one should know the features, needs and opinions of the users in order to define the coherent homogenous groups. Various mathematical models are applied to define these groups. When developing the empirical topology the factor and cluster analysis methods are mostly used. Logit regression may be used to analyse and forecast relations of the dependent dichotomic variable and independent variables measured at any scale. Above-given algorithms of the quality data analysis are illustrated by the case when the dependent variable is of dichotomic nature. Drafting general plans of Akmenė region a questionnaire survey of inhabitants of the region and towns was carried out. Application of quality analysis methods is a valuable measure enabling specialists and planners to apply the proposed solutions by taking account of their specific features and peculiarities.

Keywords: social research, segmentation, qualitative analysis methods, logit regression model.

1. Introduction

When taking strategic decisions, implementing new projects and developing business strategies the specific market research more and more frequently becomes the basis for the specialists. *The idea that social phenomena may be measured and researched is rather new. Lately, social researches have been actively developed in the field of marketing (Crouch, Housden, 2003; Gummesson, 2000; Gilbert, Churchill, 1999).*

When developing the empirical topology the factor and cluster analysis methods are mostly used. For this purpose the interrelated variables are grouped into different categories (quality of life, stability, health protection, legal base), which can be used to make the groups of objects with similar characteristics, to define and compare their relationships. The exploratory factor analysis is most widely applied to define and describe the relationships between the variables and patterns. Where it is suspected that the data collected (reporting data, answers to a questionnaire, personality characteristics or political beliefs) can be interrelated in complex relations, the factor analysis methods untangle these relationships, separate them and help to define the variation patterns. Each pattern appears as a factor delineating a distinct cluster of interrelated data.

The cluster analysis methods are successfully applied for the analysis of socio-economic problems, identification of territorial differences. In recent years cluster analysis became a constituent part of data acquisition technologies and, together with the other methods, is used to define the groups with similar patterns. If the number of variables is high the cluster and the factor analysis methods are combined.

In recent years a segmentation problem gets more and more attention, new segmentation methods are being developed, the currently used methods are being improved. Applying the probability methods the assignment of points to clusters can be performed by calculating the posterior probability of a point belonging to a cluster. Another class of segmentation methods is based on eigenvectors of the affinity matrix with different definitions of affinity. These algorithms use dominant eigenvector of matrices to perform segmentation (Shi, Malik, 1997; Perona, Freeman, 1998).

Social and market researches have a characteristic feature of the prevailing qualitative information, which does not satisfy major assumptions used in the multiple statistical methods. One of the frequently solved tasks is to forecast the values of category variables and to be able to distinguish two or more market segments. The article analyzes forecasting of dichotomic category variables and creation of two segments according to the weights of independent variables by applying a logit regression model. This model is suitable to be used under rather general assumptions: normal distribution of variables and residuals is not required. As the other regression models, this model can also include the pseudo-variables.

2. Qualitative data and analysis methods

Surveys use not only quantitative variables with a specifically defined scale (investment, income, prices, inhabitants, migration, unemployment, crimes, etc.) but also variables that in essence are of qualitative nature (wishes, intentions, social, economic or political changes, gender, race, religion, etc.) (Cannon, 1992; Chetty, 1996; Morgan, Smircich, 1980).

Qualitative variables usually show availability or non-availability of a certain feature. Such feature may be the following: male or female, Christian or non-Christian, black or white, etc. One of the ways helping to define such variables is conclusion of artificial variables that are able of gaining on such values as 1 or 0. For example: 1 may mean that a person is male, 0 that a person is female; 1 may mean that a person is with higher education, 0 may mean that a person is without higher education, etc. Variables with the values of 0 or 1 are also called *pseudo-variables.* They are also called dichotomic variables, binary variables and indicated variables.

Although pseudo-variables usually have the values of 0 and 1, it is not a mandatory condition. The pair (0.1) may be transformed into the pair of another shape in the linear function ($Y = a + bK$ ($b \neq 0$) when a and b are constants, and K may have the value of 1 or 0. When $K = 1$, then $Y = a + b$; and when $K = 0$, then $Y = a$. So the pair (0,1) becomes the pair (a; $a + b$). For example, when $a = 1$ and $b = 2$, then the pseudo-variable may be (1, 3). Thus, quantitative variables do not have a natural measurement scale (Juškevičius, Burinskienė, 2004). The analysis of qualitative data is the most complicated stage of the research. In order to provide results of the research that would reflect the transformation of the initial data into the new information, a surveyor of qualitative research has to be actively engaged in analytical work and implement all stages of the research. The stages of qualitative analysis must be understood not only to be able to carry out qualitative researches but also to be able to read, assess and interpret them (Wolcott, 1990; Strauss, 1987).

The qualitative analysis consists of different methods intended for singling out, analysis and comparison of meaningful patterns or themes and for determination and interpretation of contrast. Meaningfulness is defined by goals and tasks of the project research. The same data may be analysed and synthesised in different profiles depending on the tasks of the research. Different research methods – methods of constant comparative, phenomenological, ethnographic or descriptive and textual analysis – are chosen taking account of the data type, research goals and philosophical standpoint. However, there are several general features that make a difference between qualitative and quantitative research (Curran, Blackburn, 2001; Gill, Johnson, 1997; Harvey, Pettigrew, Ferlie, 2002). The data of quantitative data analysis usually have a digital form and are obtained for measuring the features of the researched object applying different measurement scales. Meanwhile, words serve as the material of the qualitative data analysis. Carrying out the qualitative analysis, the set of rules and procedures is much less than the sets of statistical analysis methods applied for quantitative analysis.

The following parts of the qualitative analysis could be singled out:

- absorption into a researched phenomenon;
- synthesis of the image of the researched phenomenon, taking account of relations and interaction with its aspects;
- theoretical reasoning on relations and their influence;

Great diversity is characteristic of the qualitative data: their database may consist of open and closed questions, half-structural interviews, focus group surveys, results of project or associative researches, etc. There are no restrictions with regard to the data of qualitative databases and they gain increasingly more variable forms, namely: audio and video entries, photographs, stenographs of political sessions, etc. Qualitative data may include anything that is presented in non-quantitative form or is not entered into a quantitative form. The major difference between quantitative and qualitative data mainly lies in the principles of assumptions and analysis. Surveyors of qualitative researches assume in advance that the goal of the science is to reveal the

objective laws existing in the world by using scientific methods, which will help to get a fuller conception of reality (Dzemydienė, Rudzkienė, 2003; Alvesson, Sköldberg, 2000; Storey, 1997). Qualitative surveyors, although their initial positions are similar, acknowledge that the reality being surveyed is related to human experience and this experience is subjective and gained in a certain social context and historical time. Thus, qualitative surveyors are more interested in revealing knowledge on how people think, also on their feelings and estimation of the existing situation and not in substantiating feelings and thoughts of people.

During the whole period of qualitative analysis, the surveyor must answer the following questions:

- What patterns and general themes are characteristic of specific questions? What is the meaning of these patterns (or absence of these patterns) for wider comprehension of the question?
- Are there any deviations from these patterns? If yes, which factors could serve for explaining such non-typical answers?
- What interesting plots arise from these answers? How could plots help to wider comprehend the question?
- Do the received patterns and unexpected discoveries mean that additional data should be collected? Should some questions be analysed anew?
- How do the received patterns fit to the results of previous researches? If not, how could these differences be explained?

3. Application of regression models for the analysis of qualitative data

Material collected during interviews, deep and half-structural interviews, or through project tests or associative methods becomes useful only after it is properly described, analysed and interpreted. To that end, qualitative and quantitative analysis methods intended for determination of the features of goods and services that are most important for the users are applied. During analysis, a relative importance and weight of every feature is defined. Designing and production of a successful product or service depend on the achieved results.

Different algorithms and computer programmes of analysis use qualitative and quantitative analysis methods and different measurement systems. Such use reflects potential methods and algorithms of data measurement and collection. Models of categorical measurements, ANOVA models, regression methods and linear programming methods are worth noting.

OLS (Ordinary Least Squares) regression models are most simple and frequent. With the help of a traditional least square method the same results are received as in case of other possible methods, but its application and interpretation are easier. OLS models may also use pseudo-variables. Dependent variable of this model estimates opinions of respondents by profiles described by independent variables.

In regressive models pseudo-variables are used in the same way as qualitative variables. Really, the regression model may have several auxiliary variables that in essence are qualitative. Here, a question should arise: would not it be better to analyse the cases of a qualitative variable separately? For example, if we analyse the dependence of the salary on the work experience and gender, we should separately analyse the dependence of the salary on the work experience of men and women. In this case the following rule should be followed: if the dependence of the men's salary on the work experience is described by the same function as the dependence of women's salary on the work experience with only an addition described by the feature "male" of the variable "gender" (regression lines are parallel), then it is advisable to use pseudo-variables. When the dependence of variables is described by different functions (with different coefficients; regression lines are not parallel), such cases should be analysed separately.

In the cases when a direct relation between one dichotomic variable and one or more independent variables has to be determined, the methods of logit regression are applied. Logit regression is used when the dependent variable may gain two values. For example, a voter may vote for a certain political party and may refuse to vote for it; a student may pass an exam and may fail; a patient may recover and may fail to recover; a competition may be won and may be lost, etc.

Why in order to calculate regression coefficients standard multiple regression models do not suit? Multiple regression "does not know" that the nature of the dependent variable is dichotomic, thus the concluded model may serve to forecast values higher than 1 and lower than 0. Therefore, standard models of multiple regression would ignore restrictions (dichotomic nature) applied to the dependent variables.

Regression models may be concluded in the way that instead of forecasting a binary variable a continuous variable will be forecasted which will naturally fluctuate between 0 and 1. Usually the following two models are used: *logit regression* and *probit regression*.

4. Logit regression model

Logit regression may be used to analyse and forecast relations of the dependent dichotomic variable and independent variables measured at any scale (independent variables and residuals may be without normal distribution). Nominal scales present substantial problems from the standpoint of conceptualization, as well as complexities (Borooah, 2002; Chetty, 1996; Christensen, 1997). For classifying individuals into several distinct categories the approach of discriminant analysis is useful. An alternative approach whenever there are multiple categories for a dependent variable is to dichotomize successively, each time comparing one of categories with all of the others, or perhaps a subset of them. Applying logit regression model the forecasted values will never exceed (and be equal to) 1 and will never be below (and equal to) 0. That is achieved by applying logit regression model:

$$y = \frac{\exp(a + b_1x_1 + b_2x_2 + ... + b_nx_n)}{1 + \exp(a + b_1x_1 + b_2x_2 + ... + b_nx_n)}. \quad (1)$$

It is obvious that whatever x values are, the forecasted y values of the model will always be between 0 and 1. This model is called logit due to the fact that it is easily linearised applying logit transformation. As the dependent variable y fluctuates from 0 to 1, it may be analysed as a continuous probability p that fluctuates between 0 and 1. This probability may be transformed as follows:

$$p' = \ln\left(\frac{p}{1-p}\right). \quad (2)$$

This transformation is called logit transformation. As p' may theoretically gain any value from minus to plus infinity, thus these transformed values may be used in a usual linear regression model:

$$p' = a + b_1x_1 + b_2x_2 + ... + b_nx_n + \text{e}. \quad (3)$$

When the dependent variable is dichotomic, then the assumption of the regression analysis that the dispersion of all residuals is the same is not satisfied, thus the maximum probability model is used to assess the parameters instead of using the least squares.

There are certain complications in working with log-linear models. One is that the assumption of homoscedasticity of the error term may be violated, especially if the proportions in the total sample are close to either 0 or 1.

5. Application of the logit regression model in the survey of public opinion

Above-given algorithms of the quality data analysis are illustrated by the case when the dependent variable is of dichotomic nature.

Drafting general plans of Akmenė region and Naujoji Akmenė, Akmenė and Venta towns, at the stage of the assessment of the existing state a questionnaire survey of inhabitants of the region and towns was carried out; it was arranged by the employees of Akmenė region municipality bearing direct responsibility for the arrangement of general planning of the above-mentioned territories. The survey was carried out using cultural objects located in the area, such as libraries, schools, community centres, etc. The total number of questionnaires was 600; after rejecting questionnaires with mistakes only 521 questionnaires remained.

One of the goals of the research is to find out businesses to be developed in Akmenė region in the opinion of the inhabitants of the region. To this end, inhabitants were given a closed question: "What kinds of businesses in your opinion should be developed in Akmenė region?" and the following possible answers were offered: tourism; rural tourism; agriculture; cafés and restaurants; construction; minor services; no opinion. Analysing features of inhabitants in whose opinion *tourism* and *rural tourism business* should be developed, the following 5 features out of 185 turned out to be significant (verifying significance by t criterion): 1) assessment of work of municipality's administration (x_1); 2) an offer to develop the residential location of Papilė (x_2); 3) age factor (x_3); 4) education (x_4); 5) watching television LNK (x_5). Having calculated estimators of logit regression parameters, the following equation was received:

$$\hat{p}' = 1.648 - 0.299x_1 + 0.175x_2 - 0.031x_3 - 0.288x_4 + 0.599x_5. \quad (4)$$

Equation (4) parameters are related to $ln(p/(1-p))$ but not directly to the opinion of inhabitants. In order to find out probability of inhabitants' opinion formula (2) is applied in every specific case. Inhabitants considering that tourism should be developed gave a very good estimation to the work of municipality's administration; they offer to develop Papilė, are young, have higher or college education and usually watch television LNK. The probability that an inhabitant with these features will offer to develop tourism or rural tourism is at least 0.85.

Features of inhabitants who offer development of *agriculture* are different. In this case, the following out of

185 features are statistically significant: going to trade and recreation centre by bus (x_1); surroundings of the house/apartment would improve after building of the street (x_2); an offer to develop the residential area of Akmenė (x_3); a favourite regional newspaper *Vienybė* (x_4); favourite televisions LNK and TV3 (x_6). In this case, the logit regression equation gains the following shape (5):

$$\hat{p}' = -2.18 + 0.285x_1 + 0.315x_2 6 + +0.301x_3 + 0.644x_4 + 0.440x_5 - 0.331x_6. \quad (5)$$

After transformation of equation (5) into the initial shape it would result in the fact that inhabitants offering development of agriculture, have different features from those who offer development of tourism. This group of inhabitants who gave low assessment for getting to trade and recreation centres by bus think that their residential surroundings would be improved by a built street, they offer to develop residential area of Akmenė and usually read the newspaper *Vienybė*, watch LNK and do not watch TV3. The probability that a persons with such features will be for development of agriculture is at least 0.9.

The data collected during such surveys allow to estimate specific features of inhabitants of towns and regions of Lithuania and to define their scale of values and preferences. At first glance, people living in small administration units of Lithuania seem to be very different and have different plans for their future activities. After defining these characteristic features of local inhabitants, perspectives of development and priority given to the way of development could be foreseen: will inhabitants want to concentrate at perspective settlements or will they remain supporters of special structure? Possibilities of choice and stereotypes of thinking will exert influence on perspective given in the planning documents for the period of 20 and 10 years when calculating and itemising solutions. During the whole process of planning, a public dialogue should exist between inhabitants, whose interests are represented by local politicians and plan drafters, i.e. specialists proposing perspective solutions; this dialogue should be based on mutual understanding and trust, as without consensus no rational solution could be achieved, which would result in poorer living quality of inhabitants in future or achievement of quality would be more complicated and will need more time. Therefore, application of quality analysis methods is a valuable measure enabling specialists and planners to better know local inhabitant and to apply the proposed solutions by taking account of their specific features and peculiarities.

6. Conclusions

1. Marketing researches usually become the reason for innovations or help to develop new strategies that professionals sometimes fail to see without market researches. A research-based strategic plan of companies may earn trust of external estimators and substantiate applications for financing or other resources.
2. The major difference between quantitative and qualitative data mainly lies in the principles of assumptions and analysis. The analysis of qualitative data is the most complicated stage of the research. In order to provide results of the research that would reflect the transformation of the initial data into the new information, a surveyor of qualitative research has to be actively engaged in analytical work and implement all stages of the research.
3. In the cases when a direct relation between one qualitative variable and one or more independent variables has to be determined, the non-standard methods of multiple regression are applied. With qualitative variables the standard regression models would ignore restrictions (e.g., dichotomic nature) applied to the dependent variables.
4. Analysing what businesses should be developed in the opinion of Akmenė inhabitants, the obtained results helped to reveal features characteristic of groups of inhabitants offering development of different types of business.
5. Once the segments of inhabitants are defined it is possible to strategically more successfully select the regional development strategies, to direct funds for business development. The efficiency of the planned projects is significantly higher when knowing the segments of inhabitants and their features. Application of quality analysis methods is a valuable measure enabling specialists and planners to better know local inhabitant and to apply the proposed solutions by taking account of their specific features and peculiarities.

References

ALVESSON, M.; SKÖLDBERG, K. (2000) *Reflexive methodology: new vistas for qualitative research*. London: Sage Publications.

BOROOAH, V. K. (2002) *Logit and probit.* Thousand Oaks, CA: Sage Publications.

CANNON, T. (1992) *Basic marketing: principles and practice.* London: Cassel.

CHRISTENSEN, R. (1997) *Log-linear models and logistic regression*. NY: Springer-Verlag.

CHETTY, S. (1996) The case study method for research in small- and medium-sized firms. *International Small Business Journal*, 15 (1), p. 73–85.

CROUCH, S.; HOUSDEN, M. (2003) *Marketing research for managers*. 3rd edition, Butterworth-Heinemann, Oxford.

CURRAN, J.; BLACKBURN, R. A. (2001) *Researching the small enterprise*. London: Sage Publications.

DZEMYDIENĖ, D.; RUDZKIENĖ, V. (2003) Data analysis strategy for revealing multivariate structures in social-economic data warehouses. *Informatica,* 14 (4), p. 471–486. ISSN 0868–4952.

GILL, J.; JOHNSON, P. (1997) *Research methods for managers*. 2nd edition. London: Paul Chapman Publishing Limited.

GILBERT, V.; CHURCHILL, A. (1999) *Marketing research: methodological foundations.* Fort Worth: The Dryden Press.

GUMMESSON, E. (2000) *Qualitative methods in management research*. Thousand Oaks, CA: Sage.

HARVEY, J.; PETTIGREW, A.; FERLIE, E. (2002) The determinants of research group performance towards mode. *Journal of Management Studies*, 39 (6), p. 747–73.

JUŠKEVIČIUS, P.; BURINSKIENĖ, M. (2004) *Factors of residential environment in the urban planning*. In 2nd WHO International Housing & Health Symposium. September 29 – October 1, 2004, Vilnius, Lithuania, p. 69.

MORGAN, G.; SMIRCICH, L. (1980) The case for qualitative research. *Academy of Management Review*, 5 (4), p. 491–500.

PERONA, P.; FREEMAN, W. T. (1998) A factorization approach to grouping. In Burkardt and B. Neumann, editors. *Proc ECCV*, p. 655–670.

SHI, J.; MALIK, J. (1997) Normalized cuts and image segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, p. 731–737.

STOREY, D. (1997) *Understanding the small business sector*. London: International Thomson Business Press.

STRAUSS, A. L. (1987) *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.

WOLCOTT, H. F. (1990) *Writing up qualitative research, qualitative research methods*. Series 20. Thousand Oaks, CA: Sage.