CETRA 2018
5th International Conference on Road and Rail Infrastructure
17 - 19 May 2018, Zadar, Croatia

The use of logit model for modal split estimation: a case study

Davor Krasić
Institute for Tourism, Croatia

Abstract

One of the possible approaches to classifying the transport demand models is the division into aggregated and disaggregated models. Aggregated models deal with transport demand at the zone level while disaggregated models try to describe the behaviour of a single user of the transport system. The basis of this second approach is the term "utility maximization" which is based on the assumption that each transport system user has enough information that can make a rational choice how to travel. In the paper, a case study of a smaller satellite city has been carried out by using the logit model as well as stated preference survey in order to estimate the share of common use of a car (car – pooling) for travel to a large city.

Keywords: transport demand model, modal split, logit model, stated preference survey, car - pooling

1 Introduction

Different quantitative methods from which a model approach to planning has been developed are used in modern transport planning. Today it prevails in almost all transport studies of national, regional or urban level. Models are a simplified representation of the real state of system. Their task is to simulate the change of user behaviour if new transport investments or transport policy measures would be carried out. The main feature of such model is that the user decisions are simulated through several successive phases (steps), with the results of the previous phase being the input data for the next phase. The first use of the conventional 4 - step model dates back to the sixties of the last century. Because of its logic, it has become the model that is most commonly used in developing transport plans and studies so far.

The four – step model of transport demand consists of:
• trip generation model (production / attraction of trips)
• trip distribution model (spatial distribution of trips)
• modal split model (distribution of trips by means of transport)
• trip assignment (allocation of transport flows on network).

This paper deals with the development of modal split model with the help of the stated preference survey. The stated preference survey is a method of finding out about the attitudes of a transport system users in case where a new alternative that users have not yet had the opportunity to try is offered [1]. The combination of stated preference survey and logit model can give an answer to the question how much the offered transport alternatives will in the future be chosen by users.

2 Theory

Models of transport demand, so modal split model, can be aggregated or disaggregated. Aggregated models deal with transport demand at the transport zone level (all users) while disaggregated models describe the behaviour of a single user of the transport system. The basis of the second approach is the term "utility maximization" which is based on the assumption that each transport system user has sufficient information about the transport system that can make a rational decision on how to travel [2]. When making a decision, the user evaluates the offered alternatives. In addition to the rational (measurable) parameters, there are parameters that are the result of the passenger's personal preferences. Therefore the utility function U (i, k) has two components, one measurable V (i, k) and the other random E (i, k) [3]:

$$U(i, k) = V(i, k) + E(i, k) \quad (1)$$

where U (i, k) denotes the total utility of the alternative "i" for the person (passenger) "k".

The measurable part of the utility V (i, k) consists of the characteristics of alternative "i" represented by the variables "x" and the characteristics of the passenger "k" represented by the variables "y":

$$V(i,k) = a_1x_1 + a_2x_2 + \ldots a_nx_n + b_1y_1 + b_2y_2 + \ldots + b_ny_n. \quad (2)$$

Given the existence of the random component E (i, k) in utility function it is not possible to determine with certainty which alternative will be chosen. Instead, it has been used the concept of probability "P" that the alternative will be chosen. If we have two alternatives "i" and "j" then the likelihood that the user "k" will choose the alternative "i" can be displayed as follows:

$$P(i, k) = P(U(i, k) > U(j, k)). \quad (3)$$

By incorporating the values for V (measurable part) and E (random part) into the above formula we obtain:

$$P(i, k) = P(E(j, k) - E(i, k) < V(i, k) - V(j, k)). \quad (4)$$

If the distribution of random variable E (j, k)-E (i, k) is known then we can calculate the likelihood of choosing the alternative "i". In the logit model most commonly used in the disaggregated approach, the random variable E behaves in accordance with Gumbel's distribution, which has the shape of asymmetric normal distribution.

In the logit model there are only two possible outcomes:
• "1" meaning that the user has chosen a particular alternative and,
• "0" meaning that the user did not choose a particular alternative.

If there are two alternatives ("i" and "j"), the likelihood that the user chooses the alternative "i" is:

$$P(i) = e^{V_{i,k}}/(1+e^{V_{j,k}}). \quad (5)$$

The utility function V may consist of several variables such as: travel time, travel cost, waiting time, parking price, monthly income of passenger, etc. Coefficients a1, a2,...an, b1, b2,...bn in utility function can be calculated using logistic regression which unlike ordinary regression has a dependent variable with only two possible values (1 or 0).

In order to gain a logistic regression function, it is necessary to have a number of observations that show how transport system users have chosen between alternatives. If it is about the existing alternatives, then the revealed preference is determined through the survey, and if it is an alternative that will emerge in the future then a stated preference survey should be carried out.

3 Case study

Samobor (about 38,000 inhabitants) is a satellite city 20 km far from the city of Zagreb which has about 800,000 inhabitants. More than 50% of daily trips made by the residents of Samobor are towards Zagreb and back. Given the high cost of public transport (buses) between Samobor and Zagreb citizens of Samobor for as much as 72% of the trips use a private cars. They heavily burden roads between Samobor and Zagreb and the streets in the western part of Zagreb.

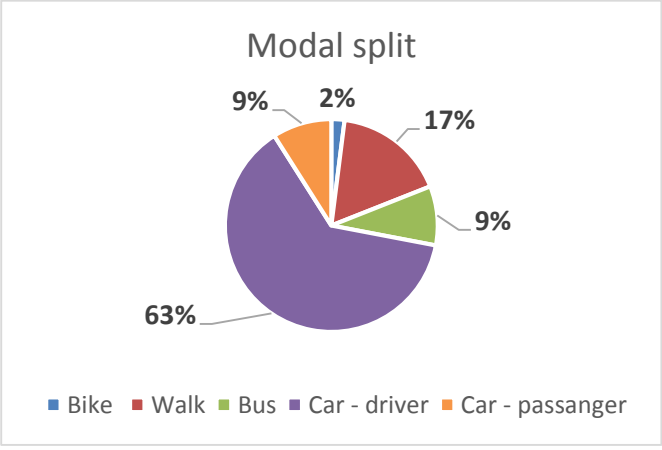

Figure 1. Modal split of daily trips

In order to explore the potential of alternative transport mode an option for carpooling has been considered. Car - pooling is conceived as a transport option that citizens can use with the help of smartphone user applications. The intention is to use private cars more rationally, i.e. to increase the occupancy of the cars. The existing average occupancy of private cars for car trips between Samobor and Zagreb is 1.3 people per vehicle.

Within the implementation of the household survey, a stated preference survey was conducted since car – pooling option has not be a transport offer in the past [4]. Each driver from randomly selected household sample was offered different combinations of travel time and travel costs. For each travel combination he/she should have chosen one of the offered transport alternatives: car pooling or private car. A few days before the interviewer's arrival in their household, a questionnaire was sent to them to have enough time to consider the answers. In the presence of the interviewer, with additional explanations and videos showing the way in which car pooling works, survey questionnaires were completed. In total the survey included 94 drivers, which resulted in 564 their responses on the choice of transport alternative since each questionnaire contained 6 different combinations of travel time and travel cost for both car – pooling and private car options.

Offered travel times and travel costs were set to be close to realistic values so that the choice of means of transport have been facilitated to surveyed drivers. The travel time was in the range of 30 to 50 minutes and the travel cost (out of pocket) ranged from 10 to 30 Kuna. After processing the questionnaires, the regression logit model was established, which has the following general form:

$$L = a1\ (\text{Time cp} - \text{Time pa}) + a2\ (\text{Cost cp} - \text{Cost pa}) + c \qquad (6)$$

where „L“ denotes the dependent variable, „cp“ denotes the car - pooling alternative, „pa“ denotes the private car alternative, „a1“ and „a2“ are regression coefficients, and "c" is the constant in the model. Logit regression was obtained using the MedCalc software tool [5] and the results are shown in table 1.

In the design of the logit model, a stepwise method was used which eliminates the variable in the regression process if it is not statistically significant. From the results it can be seen that both independent variables (travel time and travel cost) are statistically significant.

Pseudo R^2 shows the predictive power of the logit regression model, i.e. how much a model is better if it contains one or more independent variables (full model) with respect to a model that only contains a constant (null model) [6] . Pseudo R^2 values (Cox-Smell or Nagelkerk) may not be interpreted as R^2 values from the ordinary linear regression model. These pseudo R^2 values are useful for comparing different specifications of the same model (within a single set of data), i.e. they give an answer to the question of what model specification is better, whereas it is more difficult to say how much a particular model is good. For Cox-Smell pseudo R^2 the upper limit is not necessarily 1.0 as in standard R^2 from linear regression. The Nagelkerk coefficient was generated by the transformation of the Cox-Smell coefficient and its maximum possible value is 1.0. Typical values of pseudo R^2 are much lower than ordinary R^2. According to Muijs [7], the predictive model quality can be estimated using the Nagelkerk coefficient.

The ranges of the model improvement are as follows:

<0.1 = slight improvement,
0.1 - 0.3 = modest improvement,
0.3 - 0.5 = medium improvement,
> 0.5 = great improvement.

Table 1. Logistic regression parameters

Item	Value
Dependent variable L	1= Car - pooling 0= Private car
Method of analysis	Stepwise
Enter variable if P<	0.05
Remove variable if P>	0.1
Sample size	564
Positive cases (1)	329
Negative cases (0)	235
Chi - squared	151.2
DF	2
Model significance level	P<0.0001
Cox & Snell R^2	0.235
Nagelkerke R^2	0.316
Coeff. of independent variable Time	-0.15486
Coeff. of independent variable Cost	-0.24483
Constant	-0.82805
Wald (of independent variable Time)	65.159
Wald (of independent variable Cost)	36.416
Hosmer & Lemeshov test chi - squared	11.411
Hosmer & Lemeshov test DF	2
Hosmer & Lemeshov significance level	P=0.0033
Area under ROC curve	0.766

The Wald indicator shows whether coefficients of independent variables in the logit model are significant or not, i.e. whether a variable should be extracted from the model. It is obtained as the quadratic value of the quotient of the regression coefficient of an independent variable and its standard error. The higher the Wald

indicator, the magnitude of the regression coefficient significance is higher. In the obtained model, both regression coefficients are statistically significant.

The Hosmer – Lemeshov test [8] serves to evaluate how good the model is, i.e. how well it represents the actual data (goodness of fit). It is based on the grouping of logistic model results (usually in 10 groups). Then it is compared (for each group) the number of actual (observed) outcomes versus the model obtained the number of outcomes. This test uses a formula similar to the conventional chi-square test, and the model is better if the value is smaller. This test is not applicable if the total sample size is less than 400 because there is likely that the sample is too small in some groups. Since the sample in this model was 564, it is possible to apply the Hosmer - Lemeshov test.

The surface analysis under the ROC curve shows how well the model differentiates positive and negative outcomes, i.e. whether and to what extent negative outcomes are classified as positive and vice versa. If this value is 0.5 it means the model does not differentiate the outcomes anything better than it would be accidental, while the value of 1.0 indicates that the area under the ROC curve is 100%, i.e. the model perfectly distinguishes positive from the negative outcomes. Since the resulting model has a surface area below the ROC curve of 76.6% it can be said that it distinguishes the outcomes well.

The final form of the logistic regression function is:

$$L = -0.155*(\text{Time cp} - \text{Time pa}) - 0.245 *(\text{Cost cp} - \text{Cost pa}) - 0.828. \quad (7)$$

In order to obtain a probability function the logit regression function should be transformed as follows:

$$P(i) = e^{L}/(e^{L} + 1) = 1/(1+1/e^{(-0.155*(Tcp-Tpa)-0.245*(Ccp-Cpa)-0.828)}). \quad (8)$$

The following 3D graphs show the probability of choosing car - pooling relative to the values of independent variables travel time (i.e. difference in travel time) and travel cost (i.e. difference in travel cost). In graphs axis „z“ is dependent variable P(i) and axis „x“ and „y“ are independent variables that represent both travel time and travel cost.

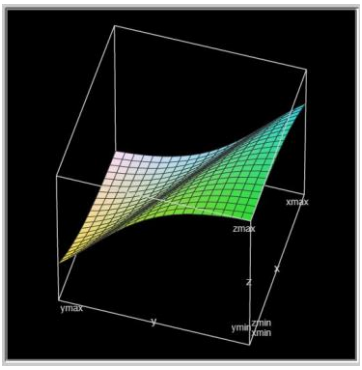

Figure 2. Probability of choosing car - pooling depending on travel time and cost

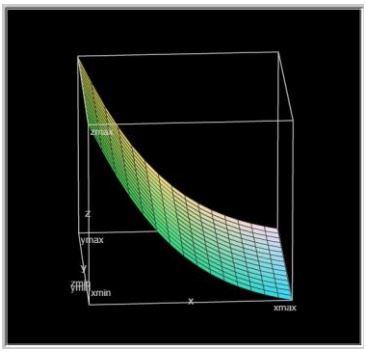

Figure 3. Probability of choosing car - pooling depending only on travel time

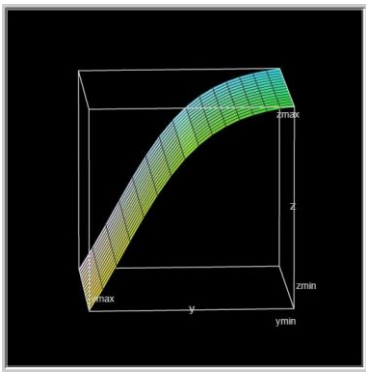

Figure 4. Probability of choosing car - pooling depending only on travel cost

4 Conclusion

Regarding the value of regression coefficients, in this case study the noticeably higher weight has a coefficient associated with travel costs in comparison with the coefficient associated with travel time. This is a logical consequence of the fact that daily travel by car (between Samobor and Zagreb) is long lasting (mostly in the range of 30 to 35 minutes), so any small differences in travel time between private cars and car - pooling do not have so much importance on the choice of alternatives such as the difference in travel cost. Since citizens of Samobor spend a lot of money for daily travel (to go to Zagreb and return to Samobor they need to drive 45 - 50 km) it is understandable that the influence of cost on the choice of alternative car - pooling in such circumstances is considerably higher.

The value of time calculated from the logit model shows that the model is well – grounded. By comparing the average monthly net earnings obtained from the

model coefficients and the actual data on average net salaries of Samobor citizens, one can get almost identical values of 6840 Kuna.

As can be seen from this case study, applying a logit model to estimating modal split is the appropriate tool if it is based on a well-conducted stated preference survey. The approach presented in this paper can serve as an example of exploring the potential of introducing new modes of transport that users have not had experience so far.

References

[1] Kroes, E. P., Sheldon, R. J.: Stated preference methods, Journal of transport economics and policy, 22, pp. 11-25, 1988.

[2] Meyer, M. D., Miller E. J.: Urban Transportation Planning, second edition, McGraw-Hill, Boston, 2001.

[3] Ben-Akiva, M., Bierlaire, M.: Discrete choice methods and their application to short term travel decision (Chapter), Handbook of Transportation Science, Kluwer Academic Publisher, Norwell, 1999.

[4] Novačko, L., Krasić, D., Pilko, H., Fosin, J., Babojelić, K.: Analiza postojećeg stanja i prikupljanje ulaznih podataka za izradu prometnog modela na relaciji Samobor – Zagreb u svrhu EU projekta SocialCar (Study), Fakultet prometnih znanosti, Zagreb, 2017.

[5] Medcalc software, www.medcalc.org.

[6] What are pseudo R-squared? UCLA: Statistical Consulting Group, https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/ (accessed August 22, 2016)

[7] Muijs, D.: Doing Quantitative Research in Education with SPSS, 2nd edition, SAGE Publications, London, 2011.

[8] Bewick, V., Cheek, L., Ball, J.: Logistic regression, Critical Care, 9, pp. 112-118, 2005.