

# A Mathematical Framework for Logit Model in Transportation Mode Choice Analysis

**Ahmad Altaweel**

*Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan*

**Kazuhito Mine**

*School of Computer Science and Systems Engineering, Kyushu Institute of Technology  
680-4 Kawazu, Iizuka-shi, Fukuoka, 820-8502, Japan*

**Bo-Young Lee**

*Logistics Revolution Korea Co., Ltd.  
5th Fl. 47, Gangnam-daero 101-gil, Seocho-gu, Seoul 06034, Republic of Korea*

**Jang-Sok Yoon**

*Logistics Revolution Korea Co., Ltd.  
5th Fl. 47, Gangnam-daero 101-gil, Seocho-gu, Seoul 06034, Republic of Korea*

**Hiroaki Wagatsuma**

*Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan  
Email: altaweel.ahmad770@mail.kyutech.jp, mine.kazuhito139@mail.kyutech.jp, {bylee, kjsyoon}@logisroad.com,  
waga@brain.kyutech.ac.jp*

## Abstract

Traditional transportation demand forecasting has relied on massive zone-specific aggregations, which assume a linear demand increase. Such models may lack the flexibility needed to perform dynamic and context-sensitive analyses. Recently, disaggregated behavioral models have gained prominence for requiring less data and enabling sensitivity analyses in policy decisions. This study explores the feasibility of a model focusing on the mathematical formulation and validation of the transportation mode choice model. The study uses the logit model with a non-linear probability distribution function represented by a logistic curve and incorporates a linear combination of independent predictor variables. The mathematical model is examined for its ability to estimate choice probabilities. The methodology is formulated to be adaptable to diverse contexts that provide an analytical framework for transportation systems independently of geographic or demographic considerations.

*Keywords:* Transportation Mode Choice, Logit model, maximum likelihood estimation, Gumbel distribution.

## 1. Introduction

In microeconomics, disaggregated behavioral models such as the Logit model have been used to explain qualitative choice in specific phenomena. The logit model uses a logistic function to fit a probability distribution as a nonlinear S-shaped sigmoid curve. This allows for predicting discrete transitions between different modes. The sigmoid function is frequently utilized in soft computing for binary transitions or classifications. Consistently, the Logit model simulates individual decision-making by assuming rational behavior aimed at maximizing personal utility. Utility is a quantifiable measure of an individual's satisfaction from a specific behavioral change.

This paper is intended to enhance the understanding of model structure and estimation of the Logit model with a focus on the data structure and estimation procedure. Through the incorporation of attributes such as distance, time, and cost, the model evaluates their impact on decision-making processes and the resulting probabilities. Recognizing the challenges associated with collecting

data to validate such models, this study adopts a structured data generation approach, enabling control over the attributes and the diversity of decision-making scenarios. This method allows for evaluation of the logit model precision under predetermined conditions which support critical alternatives selection. Section 2 explores the assumptions and formulation of the logit model. In addition, the design of the structured dataset. We then discuss the dataset and its estimation in the results and discussion section and finally the conclusion.

## 2. Methodology

### 2.1. Logit Model Assumptions and General Framework

The logit model is a choice model that uses probabilistic approaches to evaluate individual preferences for different modes of transport. It formulates mathematically the relation between attributes of the alternative and characteristics of the decision maker into

a utility function. The concept of utility allows one to rank a series of alternatives. The utility maximization rule states that an individual will select the alternative that maximizes his utility [1]. Despite this theoretical framework, there are three primary sources of error in using deterministic utility functions. First: the individual has incomplete or incorrect information or misperceptions about the attributes of alternatives. Second, the observer has different or incomplete information about the attribute. Third: when analysts do not fully understand the decision process.

The general framework for using the logit model in travel mode can be summaries in Fig. 1 in the following steps: (1) Define problem context, (2) Gather mode attributes and observe choices (3) Define utility function and Identify predictor vars (4) Estimate model coefficients (5) Evaluate the model by comparing predicted choices and observed choices (6) Validate the model by performing sensitivity analysis of coefficient values (7) Analyze results to provide policy insights.

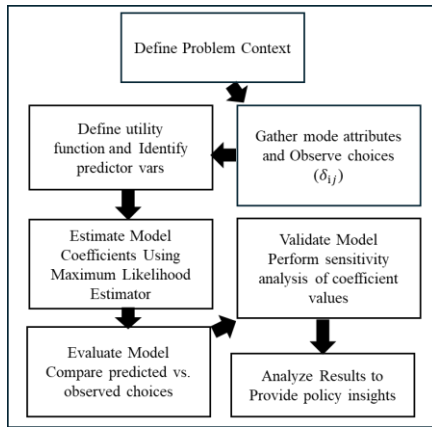


Fig. 1 The general framework for using the logit model

## 2.2. Logit Model Formulation

The logit model in this study is estimated in three steps:

1. Determine the shape of the utility function.
2. Use the least squares method to approximate the coefficient values.
3. Find a solution that maximizes the log-likelihood function.

Assume that the utility  $U_{ij}$  of individual  $j$  when choosing option  $i$  is expressed as:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \quad (1)$$

In the model,  $V_{ij}$  in Eq. (1) represents the deterministic component of the utility, while  $\varepsilon_{ij}$  denotes the stochastic component which accounts for random error in the true utility. Assuming  $i$  represents a binary choice of either 0 or 1, then the deterministic component

$V_{ij}$  is expressed as a linear function of time  $t_{ij}$ , cost  $c_{ij}$  and distance  $d_{ij}$  as:

$$\begin{aligned} V_{0j} &= \alpha_0 d_{0j} + \alpha_1 c_{0j} + \alpha_2 t_{0j} \\ V_{1j} &= \alpha_0 d_{1j} + \alpha_1 c_{1j} + \alpha_2 t_{1j} \end{aligned} \quad (2)$$

In the Eq. (2),  $\alpha_k$  represents an unknown coefficient. Once  $\alpha_k$  is known, the deterministic term  $V_{ij}$  can be calculated by using time, cost, and distance. If the stochastic component  $\varepsilon_{ij}$  follows normal distribution, it is possible to use the least squares method to find the coefficient  $\alpha_{ij}$ . On the other hand, if  $\varepsilon_{ij}$  follows Gumbel distribution, the maximum likelihood estimation method is more flexible and can handle non-normal error terms. Since an analytical solution is generally unattainable, iterative methods like Newton's are employed to find a numerical solution.

The probability  $P_i$  of choosing option  $i$  is expressed as:

$$P_i = \frac{e^{V_i}}{e^{V_0} + e^{V_1}} \quad (3)$$

Then the binary probability is:

$$P_0 = \frac{1}{1 + e^{V_1 - V_0}}, P_1 = \frac{1}{1 + e^{V_0 - V_1}} \quad (4)$$

By taking the logarithm of the ratio of probabilities  $P_0$  and  $P_1$ , we derive the log-odds ratio or logit:

$$\log \frac{P_1}{P_0} = \log \frac{e^{V_1}}{e^{V_0}} = V_1 - V_0 \quad (5)$$

$$\begin{aligned} \log \frac{P_1}{P_0} &= \alpha_0(d_1 - d_0) + \alpha_1(t_1 - t_0) \\ &\quad + \alpha_2(c_1 - c_0) \end{aligned} \quad (6)$$

Thus, the data for  $J$  individuals represent matrix:

$$\begin{aligned} p &= \left[ \log \frac{P_{10}}{P_{00}}, \dots, \log \frac{P_{1J}}{P_{0J}} \right]^T \\ \alpha &= [\alpha_0, \alpha_1, \alpha_2]^T \end{aligned} \quad (7)$$

$$X = \begin{bmatrix} d_{11} - d_{01} & t_{11} - t_{01} & c_{11} - c_{01} \\ \vdots & \vdots & \vdots \\ d_{1J} - d_{0J} & t_{1J} - t_{0J} & c_{1J} - c_{0J} \end{bmatrix}$$

As for the normal equations:

$$\begin{aligned} X\alpha &\approx p \\ \alpha &\approx [X^T X]^{-1} X^T p \end{aligned} \quad (8)$$

An estimated value of  $\alpha$  can be obtained via the least squares method, denoted as  $\hat{\alpha}'$ .

Taking the logarithm of the probability in Eq. (4) gives:

$$\log P_{0j} = -\log(1 + e^{V_{1j}-V_{0j}})$$

$$\log P_{1j} = -\log(1 + e^{V_{0j}-V_{1j}}) \quad (9)$$

In this case, the log-likelihood function  $L(\beta)$  over  $J$  observations can be written as:

$$L(\alpha) = \sum_{j=1}^J (\delta_{0j} \log P_{0j} + \delta_{1j} \log P_{1j}) \quad (10)$$

The term  $\delta_{ij}$  denotes the Kronecker Delta function, which return 1 when individual  $j$  chooses option  $i$  and 0 otherwise.

### 2.3. Structured Data Set Design

The dataset design shows intercity structure with nodes representing stops accessible by bus or car, and edges representing roads or pathways. The structure is a square grid with "n" nodes per side. Path selection between nodes has three criteria: travel distance, time, and cost. The bus route follows a fixed, clockwise loop along the grid's outer contour excluding interior nodes (Fig. 2). This predefined loop imitates real bus route and to enhance realism, individuals can walk to the nearest node on the bus route if their starting point is not directly on it. Individuals total travel then is walking to the nearest bus stop, taking the bus along its predefined route and walking to their destination. In contrast, Fig. 3. shows in this structure cars have bi-directional connectivity between all adjacent nodes (two ways paths). Individuals using cars directly access any node bypassing the restrictions in the bus case.

To compute the shortest paths between all pairs of nodes, the Floyd-Warshall algorithm [2] was used. For the bus route, it is assumed to operate continuously moving along the route. As a result, distances between non-adjacent nodes on the bus route are often longer due to directional restrictions. Conversely, car distances are symmetric and consistent regardless of the direction of travel. The distance between adjacent nodes is assumed to be uniform at one distance unit of 1 km.

Assuming the individuals can walk to the nearest node on the bus route and then use the bus to travel to other nodes, a mixed distance, time, and cost were made for bus mode. The design involves the process: (1) identifying nodes that have no adjacent nodes in the adjacency matrix (2) determining the nearest node prioritizing nodes located on the bus route (3) updating the distance matrix by adding the distance between the node and the nearest node.

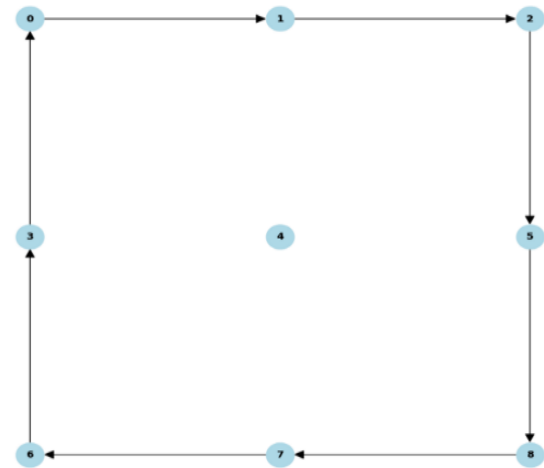


Fig. 2. Grid 3 × 3 represents the bus route

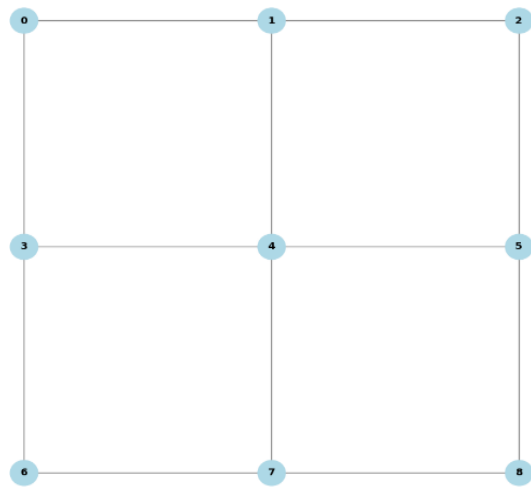


Fig. 3. Grid 3 × 3 for the car route

To calculate travel time, we assumed a bus speed of 30 km/h and a walking speed of 5 km per hour as fixed values across all distances. This assumption allows us to derive the travel time for bus and walking segments by dividing the respective distances by their speeds. For a car trip we assumed a minimum speed of 5 km/h and a maximum speed of 50 km/h with a random distribution on the trips in the dataset. Walking is assumed to have no cost, while the bus cost is 150 units, and the car is 170 units. For the car mode, the cost is then influenced by multiple factors, including driver behavior factor, fuel rate factor, maintenance rate, and vehicle type factor. Similarly, the cost matrix for the mixed-mode bus trips assumes a walking cost of zero for movement between any pair of nodes. Under this framework, for nodes without adjacent connections, the cost of movement is assigned as the bus travel cost from the nearest node that has access to the bus route.

### 3. Results and Discussion

#### 3.1. Dataset Formation

It is important to consider that decision-makers assign different levels of importance to travel distance, time, and cost. Developing choice models at the decision-maker level allows us to incorporate variables that capture these differences. We assume travelers are business-oriented and prioritize time over cost. For that, time is more critical than cost, which is reflected in the sensitivity relationship:  $\beta_{time,work} = r_{work} \cdot \beta_{cost,work}$ . By defining base sensitivities as  $\beta_{cost,work} = -0.9$ ,  $\beta_{time,work} = -3.6$ , and  $\beta_{distance,work} = -0.5$  that emphasize the importance of time to the cost for these individuals. Fig. 4 and Fig. 5 illustrate the distribution of the generated utility of buses and cars for time and cost.

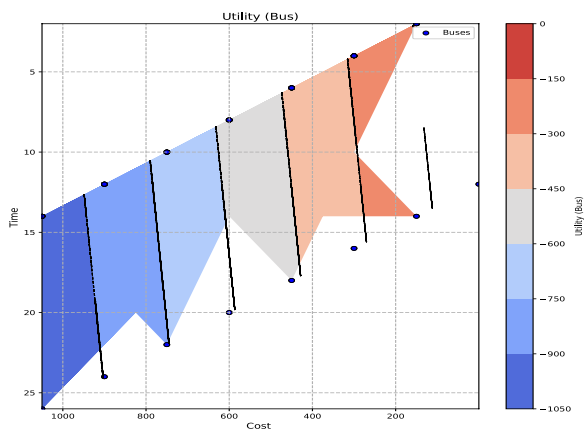


Fig. 4. The generated utility associated with the bus for all combinations of cost and time.

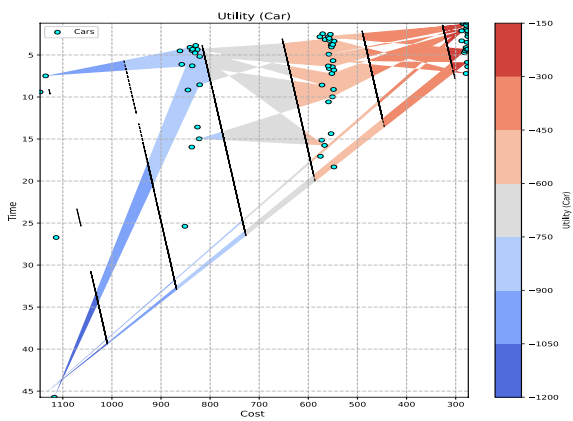


Fig. 5. The generated utility associated with the car for all combinations of cost and time.

Using the logit model, we determined the observed probabilities by maximizing the utility captured by the logit model. Fig. 6 shows the probabilities observed for the bus and Fig. 7 shows the observed probabilities for the car. According to the observed probabilities, the share of buses is 0.5937, while the share of cars is 0.4063.

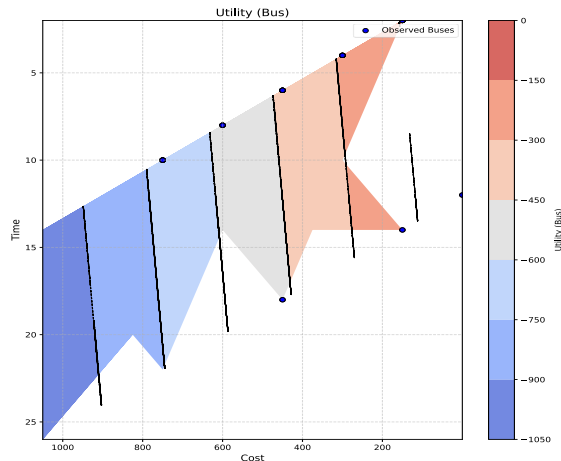


Fig. 3. The probabilities observed for the bus

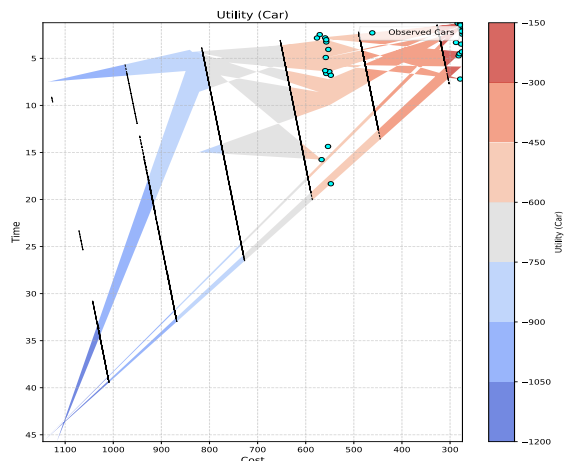


Fig. 7. The observed probabilities for the car

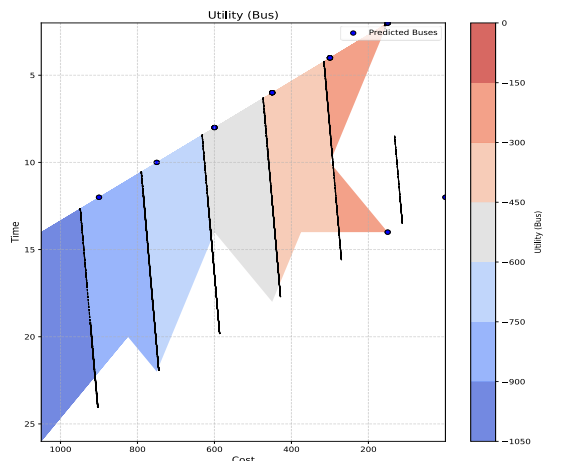


Fig. 8. The estimated probabilities for the bus.

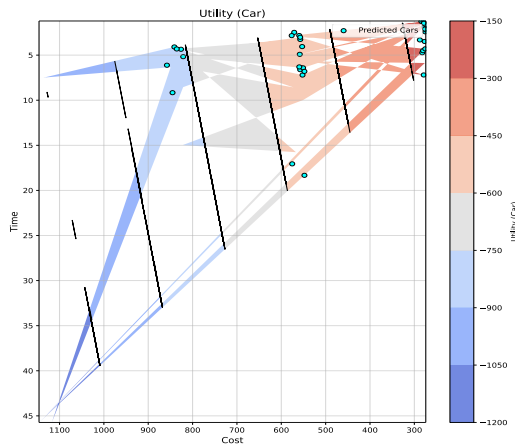


Fig. 9. The estimated probabilities for the car.

To estimate the parameters of the logit model ( $\beta$ ) based on distance, time, and cost, we initiated with an estimation obtained using the least squares error method. These  $\beta$  values were then refined through maximum likelihood estimation and Newton's method. The mean squared error (MSE) between the observed probabilities of the bus and the predicted probabilities was calculated to be 0.139. Fig. 8 and Fig. 9 show the estimated probabilities for the bus and the car. After estimation, the estimated share of the bus is 0.5139, and the estimated share of cars is 0.4861.

#### 4. Conclusion

This study evaluates a structured dataset to assess the logit model's accuracy in predicting predesigned observed probabilities. To conclude, relying solely on distance, time, and cost restricts the model's versatility as these attributes make utility and probability differences overly predictable. The simplicity limits the evaluation of the model's ability in this straightforward scenario. Stressing the need for additional attributes to enhance competition between modes and less diverted probabilities estimated by the mode.

#### Acknowledgments

This work was supported in part by the cooperative research project on digital logistics between the Kyushu Institute of Technology and Logistics Revolution Korea Co., Ltd.

#### References

1. F. S. Koppelman and C. Bhat, *A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models*, prepared for U.S. Department of Transportation, Federal Transit Administration, Jan. 31, 2006.
2. R. W. Floyd, "Algorithm 97: Shortest Path," *Communications of the ACM*, vol. 5, no. 6, p. 345, June 1962.

---

### Authors Introduction

#### Mr. Ahmad Altaweel



He received his Bachelor's degree in Electronics and Communication Engineering, ALBAATH University, Syria. He is currently a Ph.D. student at Kyushu Institute of Technology, Japan.

#### Mr. Kazuhito Mine



He is a bachelor's student in the School of Computer Science and Systems Engineering, Kyushu Institute of Technology, Japan

#### Ms. BoYoung Lee



She received her bachelor's degree in human relations from Keio University and M.S. in logistics from Inha University in Korea and is currently a Ph.D. student. She is a logistics consultant in the Logistics Revolution Korea Co., Ltd.

#### Dr. JangSok Yoon



He received his M.S. and Ph.D. degrees from Kyunghee University, in South Korea, in 2004 and 2008. He is the founder, CEO, and Consultant over 23 years of Logistics Revolution Korea Co., Ltd.

#### Dr. Hiroaki Wagatsuma



He received his M.S., and Ph.D. degrees from Tokyo Denki University, Japan, in 1997 and 2005, respectively. He is currently an Associate Professor at Kyushu Institute of Technology.

---