DISCRETE CHOICE MODELLING USING MULTINOMIAL LOGIT MODEL

Student Name: Jiya Verma
Roll No. 220480
Department: Statistics and Data Science
Project Mentor: Prof. Shankar Prawesh

Date: 8th July,2024

PROJECT REPORT

ABSTRACT

This project focuses on discrete choice modelling using the Swissmetro dataset with Biogeme and extends to an itinerary choice dataset using the Larch library. The primary objective is to several factors on travel mode choices. The methodology involves understand and predict travellers' mode choices based on various attributes such as time periods, carriers, equipment types, fares, elapsed times, and the number of connections for itinerary choice model and travel time, cost, and alternative availability for Swissmetro Model. The model is specified, estimated, and compared to understand the influence of key steps such as data loading and pre-processing, model specification, model building and estimation, and evaluation and prediction. The model's performance is evaluated, and predictions are made to assess its accuracy.

In parallel, the itinerary choice dataset is utilized to build and estimate an MNL model using Larch. Thisinvolvesdefiningautilityfunctionwithvariablesinfluencingitinerary choices, constructing the model, and estimating its parameters by maximizing the log likelihood function. The findings from both datasets highlight significant variables affecting mode and itinerary choices, demonstrating the application of advanced choice modelling techniques using Biogeme and Larch. This project provides valuable insights for transportation planners and policymakers, helping to enhance travel demand models and improve transportation services.

CONTENTS

1 INTRODUCTION

1.1 BACKGROUND

In the realm of transportation planning and decision-making, understanding traveller behaviour and preferences plays a pivotal role. The ability to accurately model and predict choices made by travellers is crucial for optimizing transportation systems and infrastructure development. Previous studies have shown that discrete choice models, such as the multinomial logit model (MNL), Mixed Logit Model, Nested Logit Model, etc. provide effective frameworks for analysing and forecasting travel behaviour based on observed choices. However, applying these models to datasets like the Swissmetro dataset and itinerary choice datasets requires robust methodologies and appropriate tools to derive meaningful insights.

1.2 GOAL

This project aims to employ multinomial logit (MNL) modelling techniques to analyze and predict traveller behaviour in two distinct contexts:

- **Swissmetro Dataset Analysis**: Utilizing Biogeme for comprehensive analysis of mode choice behaviour among Swissmetro travellers, identifying key factors influencing transportation mode preferences.
- **Itinerary Choice Modelling**: Using Larch to explore traveller' preferences in itinerary planning scenarios, investigating the factors that drive itinerary selection and decision-making.

By applying MNL models to these datasets, the project seeks to enhance understanding of traveller decision processes and provide insights valuable for optimizing transportation systems and informing policy decisions.

1.3 SETUP AND SCOPE

This project focuses on utilizing multinomial logit (MNL) modelling techniques to analyze traveller behaviour using two distinct datasets: the Swissmetro dataset and an itinerary choice dataset. The Swissmetro dataset will be analyzed using Biogeme to understand mode choice behaviour among travellers, while the itinerary choice dataset will be explored using Larch to investigate factors influencing itinerary selection. The scope includes applying MNL models to these datasets to uncover insights into traveller decision-making processes. The project aims to contribute to the optimization of transportation systems and inform policy decisions by providing valuable insights into traveller preferences and behaviour. Transportation systems and inform policy decisions by providing valuable insights into traveller preferences and behaviour.

2 Theoretical Background

2.1 Discrete Choice Model

Discrete Choice Modelling (DCM) is structured around four fundamental components: the decision-maker, alternatives, attributes, and decision rules. Each decision-maker $i \in I$ faces a choice set A_i composed of J_i alternatives, where $j \in \{1, \ldots, J_i\}$ indexes each alternative. Simplifying without loss of generality, we denote the number of alternatives as J.

The decision-maker assigns a utility U_{ij} to each alternative j and selects alternative $\hat{j}$ if and only if

$$U_{i,\hat{j}} \geq U_{ij} \quad \text{for all } j \in A_i \tag{1}$$

While the utility function U_{ij} is unobservable, it can be related to observed attributes x_{ij} and personal characteristics S_i through a function $V(\cdot)$, where

$$V_{ij} = V(X_{ij}) \tag{2}$$

Typically, V_{ij} is a linear combination of attributes, such as

$$V_{ij} = a \cdot \text{price}_{ij} + b \cdot \text{tripDuration}_{ij} \tag{3}$$

with a and b as parameters (commonly denoted as β) to be estimated. Since U_{ij} incorporates unobservable components, it is modeled as

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{4}$$

where ϵ_{ij} captures random factors affecting U_{ij}.

Thus, the decision rule is expressed as the probability that decision-maker i selects alternative k from A_i, considering

$$P(k|A_i) = P(U_{ik} \geq U_{ij}; \forall j \in A_i) \tag{5}$$

Different assumptions about ϵ_{ij} and V_{ij} yield specific models tailored to various decision-making scenarios.

2.2 MULTINOMIAL LOGIT MODEL

2.2.1 Model Description

The project employs the Multinomial Logit (MNL) model to analyze decision-making among multiple alternatives present in our dataset. Each alternative in the dataset represents a distinct choice available to individuals, such as various travel itineraries or product options. The model aims to understand how individuals make choices based on attributes associated with each alternative. These attributes typically include factors like cost, duration, departure times, and other relevant features that influence decision-making.

2.2.2 Mathematical Formulation

The Multinomial Logit (MNL) model provides a structured framework to analyze and predict choice behavior:

- Utility Calculation: For each decision-maker facing a choice set, the utility U_{ij} of selecting alternative j is decomposed into two components:
  - Systematic Utility V_{ij}: This component reflects the perceived value or attractiveness of alternative j, based on observed attributes X_{ij}. It is modeled as a linear function:
    $$V_{ij} = \beta' X_{ij} \tag{6}$$
    where β is a vector of parameters to be estimated.
  - Random Utility Component ϵ_{ij}: This captures unobserved factors influencing choice, such as individual preferences or other stochastic elements.
- Choice Probability: The probability that a decision-maker i selects alternative j from the choice set A_i is given by the multinomial logit probability formula:
  $$P_{ij} = \frac{\exp(V_{ij})}{\sum_{k \in A_i} \exp(V_{ik})} \tag{7}$$
  This formula normalizes the exponentiated utilities across all alternatives in the choice set, providing the probability distribution over possible choices.

The MNL model is widely used in empirical research to analyze discrete choice behavior across various domains, including transportation planning, consumer behavior, and market research. It provides insights into how individuals evaluate and choose among multiple options based on observable attributes and underlying preferences.

2.3 Likelihood function and Estimation

The likelihood function measures the probability of the observed choices given a set of parameters. The objective is to find the parameter values that maximize this probability, which is known as Maximum Likelihood Estimation (MLE).

For a multinomial logit (MNL) model, the probability P_{ni} that individual n chooses alternative i from a set of J alternatives is given by:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j \in J} \exp(V_{nj})}$$

where V_{ni} is the systematic utility component for individual n and alternative i.

The likelihood function $L(\theta)$ for a sample of N individuals is the product of the individual probabilities:

$$L(\theta) = \prod_{n=1}^{N} \prod_{i=1}^{J} P_{ni}^{y_{ni}}$$

where y_{ni} is a binary variable that equals 1 if individual n chooses alternative i, and 0 otherwise.

The log-likelihood function $\mathcal{L}(\theta)$ is often used instead of the likelihood function because it is easier to work with, especially for optimization purposes. The log-likelihood function is given by:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \sum_{i=1}^{J} y_{ni} \log(P_{ni})$$

The parameters θ are estimated by maximizing the log-likelihood function. This process involves numerical optimization techniques, as the log-likelihood function is generally non-linear and cannot be solved analytically.

The quality of the estimated model can be evaluated using various metrics, such as the log-likelihood value at convergence, Likelihood Ratio Test, and the Bayesian Information Criterion (BIC). These metrics help compare different models and assess their goodness of fit.

2.4 Numerical Optimization

Numerical optimization techniques are essential for estimating the parameters of complex models. Two commonly used methods that can be used in our model are Newton-Raphson method and Gradient Descent.

2.4.1 Newton-Raphson Method

The Newton-Raphson method is an iterative technique for finding successively better approximations to the roots (or zeroes) of a real-valued function. When applied to optimization, it aims to find the maximum (or minimum) of a function by solving the equation where the gradient (first derivative) is zero. The update rule for the parameter vector $\boldsymbol{\theta}$ is given by:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathsf{H}^{-1}(\boldsymbol{\theta}_k) \nabla L(\boldsymbol{\theta}_k)$$

where $\mathsf{H}(\boldsymbol{\theta}_k)$ is the Hessian matrix of second derivatives and $\nabla L(\boldsymbol{\theta}_k)$ is the gradient vector of first derivatives of the log-likelihood function L.

The steps involved are:

1. Compute the gradient $\nabla L(\boldsymbol{\theta}_k)$ and the Hessian matrix $\mathsf{H}(\boldsymbol{\theta}_k)$ of the log-likelihood function L at the current parameter estimate $\boldsymbol{\theta}_k$.

2. Update the parameter vector using the update rule:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathsf{H}^{-1}(\boldsymbol{\theta}_k)\nabla L(\boldsymbol{\theta}_k)$$

3. Repeat steps 1 and 2 until convergence, i.e., until the change in the log-likelihood or the parameter estimates is smaller than a predefined threshold.

The Newton-Raphson method is known for its fast convergence properties, especially near the optimum. However, it requires the computation and inversion of the Hessian matrix, which can be computationally intensive for high-dimensional parameter spaces.

2.4.2 Gradient Descent

Gradient Descent is another iterative optimization algorithm used to find the minimum (or maximum) of a function. Unlike the Newton-Raphson method, it only requires the gradient of the function. The update rule for the parameter vector $\boldsymbol{\theta}$ is given by:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha\nabla L(\boldsymbol{\theta}_k)$$

where α is the learning rate, a positive scalar that controls the step size of each iteration, and $\nabla L(\boldsymbol{\theta}_k)$ is the gradient of the log-likelihood function L at the current parameter estimate $\boldsymbol{\theta}_k$.

The steps involved are:

1. Compute the gradient $\nabla L(\boldsymbol{\theta}_k)$ at the current parameter estimate $\boldsymbol{\theta}_k$.

2. Update the parameter vector using the update rule:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha\nabla L(\boldsymbol{\theta}_k)$$

3. Repeat steps 1 and 2 until convergence.

Gradient Descent is simpler to implement and can handle large-scale optimization problems as it does not require the computation of the Hessian matrix. However, its convergence rate can be slow, and the choice of learning rate α is crucial for the algorithm's performance. A learning rate that is too large can cause the algorithm to diverge, while a learning rate that is too small can result in very slow convergence.

3 METHODOLOGY

3.1 Data Sources and Description

3.1.1 SwissMetro Dataset

The Swissmetro dataset, gathered between 2009 and 2010 in collaboration with CarPostal, Switzerland's public transport arm, forms the basis of this study on mode choice behavior.

In total, the dataset comprises 1124 completed surveys, with a focus on 1906 trip sequences linked to psychometric and socio-economic attributes of the respondents. Each recorded sequence begins and ends at the respondent's residence, providing insights into their daily commuting patterns and transport preferences. To ensure a representative sample, observations are weighted across six socio-economic dimensions: possession of a driving license, gender, education level, number of cars in the household, age, and household size. These weightings enhance the model's accuracy in predicting outcomes such as cost and time elasticities, thereby aiming to better reflect the broader Swiss population.

This dataset serves as a valuable resource for analyzing the determinants of travel behavior, particularly in low-density areas where CarPostal operates, and provides a robust foundation for exploring the market potential of integrated mobility solutions within Swiss agglomerations.

3.1.2 Itinerary Choice Dataset

The itinerary choice dataset used in this model is based on data sourced from a ticketing database provided by the Airlines Reporting Corporation (ARC). The dataset covers ten origin-destination pairs within U.S. continental markets, specifically from May 2013. To comply with nondisclosure agreements, certain itinerary characteristics have been anonymized: airlines are denoted as generic entities like "carrier X," and departure times have been aggregated into categories. Additionally, a random error has been intentionally added to each fare, rendering the fare data slightly inaccurate.

These modifications were implemented to ensure the dataset's suitability for educational and demonstration purposes while maintaining confidentiality. While the dataset offers insights into real-world itinerary choice patterns from a behavioral perspective, its accuracy is compromised and therefore caution should be exercised in applying its findings to behavioral studies.

This dataset serves as a representative example commonly used in practical applications, offering intuitive results that align with behavioral expectations within the travel industry.

3.2 Libraries Used

The project leverages several libraries to facilitate data analysis, modeling, and visualization:

- Biogeme: Python package used for the estimation of discrete choice models. Biogeme provides tools for model specification, estimation, and evaluation, particularly suited for transportation studies and choice modeling.
- Larch: Python library focused on choice modeling and econometrics, providing efficient tools for estimating and simulating discrete choice models. Larch supports various model specifications and estimation techniques, including multinomial logit models and advanced choice models.
- Pandas: A versatile data analysis and manipulation tool for Python, widely used for handling structured data. Pandas facilitates data preprocessing, transformation, and integration, crucial for preparing datasets before modeling.
- NumPy: A fundamental package for scientific computing in Python, providing support for large, multi-dimensional arrays and matrices. NumPy is essential for mathematical operations and statistical computations within the project.
- SciPy: A Python library for scientific and technical computing, offering modules for optimization, integration, interpolation, and more. SciPy is employed for advanced mathematical functions and statistical testing within the project.

These libraries collectively support the project's objectives in modeling discrete choice behavior, analyzing travel preferences, and evaluating model performance using real-world datasets.

4 RESULTS AND ANALYSIS

4.1 Findings from the Swissmetro Model

In this section, we present the findings from the multinomial logit (MNL) model applied to the Swissmetro dataset. The analysis includes parameter estimates and interpretation as well as model performance metrics.

4.1.1 Parameter Estimates and Interpretation

The estimated parameters from the MNL model are summarized in Table ??. Each parameter's estimate, standard error, t-statistic, and p-value are provided.

Table 1: Parameter Estimates from the Swissmetro MNL Model

Parameter	Estimate	Std. Error	t-statistic	p-value
ASC_CAR	-0.155	0.0432	-3.58	0.000348
ASC_TRAIN	-0.701	0.0549	-12.8	0.000000
B_COST	-1.080	0.0518	-20.9	0.000000
B_TIME	-1.280	0.0569	-22.5	0.000000

The interpretation of these estimates is as follows:

- ASC_CAR: The negative alternative specific constant (ASC) for the car indicates that, all else being equal, the car mode is less preferred compared to the reference mode (Swissmetro).
- ASC_TRAIN: The negative ASC for the train suggests that, all else being equal, the train mode is less preferred compared to the reference mode (Swissmetro).
- B_COST: The negative coefficient for cost indicates that higher travel costs reduce the likelihood of choosing a mode, as expected.
- B_TIME: The negative coefficient for travel time indicates that longer travel times reduce the likelihood of choosing a mode, as expected.

4.1.2 Model Performance Metrics

The performance of the MNL model is evaluated using various metrics, including the log-likelihood, likelihood ratio test, rho-squared values, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). These metrics are summarized in Table ??.

The metrics indicate the following:

- Log-Likelihood: The final log-likelihood value of -5331.252 is slightly higher than the initial log-likelihood value, indicating that the model has improved but only marginally.

Table 2: Model Performance Metrics for the Swissmetro MNL Model

Metric	Value
Number of Parameters	4
Sample Size	6768
Initial Log Likelihood	-5332.093
Final Log Likelihood	-5331.252
Likelihood Ratio Test (Init)	1.68217
Rho Square (Init)	0.000158
Rho Bar Square (Init)	-0.000592
Akaike Information Criterion (AIC)	10670.5
Bayesian Information Criterion (BIC)	10697.78
Final Gradient Norm	0.0001231191

- Rho-Squared and Rho-Bar-Squared: Both rho-squared (0.000158) and rho-bar-squared (-0.000592) values are very close to zero, suggesting that the model explains very little of the variation in the data.
- AIC and BIC: The Akaike Information Criterion (10670.5) and Bayesian Information Criterion (10697.78) provide measures for model comparison, with lower values generally indicating a better fit.

Overall, while the estimated parameters align with theoretical expectations, the performance metrics suggest that the model has limited explanatory power. Further model refinement or additional data may be necessary to improve model performance.

4.1.3 Predicting Values and Calculating Accuracy

- Data Preparation and Model Application

  After preparing the Swissmetro dataset and splitting it into training and testing subsets, we applied the Multinomial Logit (MNL) model using Biogeme. The model was trained on the training set to estimate parameters that govern transportation mode choice based on variables such as travel time and cost.

- Prediction of Choices

  Using the trained MNL model, we simulated choice probabilities (`Prob1`, `Prob2`, `Prob3`) for each observation in the test dataset. These probabilities represent the likelihood of choosing each transportation mode (Train, Swissmetro, Car) based on the modeled utility functions and availability conditions.

- Accuracy Assessment

  To evaluate the model's performance, we compared the predicted choices (`simulated_choices`) against the actual choices (`actual_choices`) from the test dataset. The accuracy of the model was calculated as the percentage of correctly predicted choices.

- Results

  The Multinomial Logit (MNL) model demonstrated strong predictive performance, achieving an accuracy of 68.69% on the test dataset. This indicates that the model effectively predicts transportation mode choices based on the specified variables and conditions, providing valuable insights into mode choice behavior within the Swissmetro context.

4.2 Findings from the Itinerary Choice Model

In this section, we present the findings from the estimation of the itinerary choice model.

4.2.1 Parameter Estimates and Interpretation

The estimated parameters from the model are presented in Table Table ?? presents the parameter estimates from the model.

Variable	Value	Std Err	t Stat	Signif	Null Value
carrier==2	0.117	0.00869	13.49	***	0.00
carrier==3	0.639	0.00813	78.55	***	0.00
carrier==4	0.565	0.0176	32.18	***	0.00
carrier==5	-0.624	0.0130	-48.17	***	0.00
elapsed_time	-0.00609	0.000111	-54.63	***	0.00
equipment==2	0.466	0.00931	50.10	***	0.00
fare_hy	-0.00118	2.83e-05	-41.54	***	0.00
fare_ly	-0.00118	8.51e-05	-13.83	***	0.00
nb_cnxs	-2.95	0.0254	-115.82	***	0.00
timeperiod==2	0.0959	0.00948	10.12	***	0.00
timeperiod==3	0.127	0.00953	13.28	***	0.00
timeperiod==4	0.0606	0.00978	6.19	***	0.00
timeperiod==5	0.141	0.00973	14.49	***	0.00
timeperiod==6	0.238	0.00973	24.49	***	0.00
timeperiod==7	0.351	0.00996	35.26	***	0.00
timeperiod==8	0.353	0.0105	33.79	***	0.00
timeperiod==9	-0.0103	0.0110	-0.94		0.00

Table 3: Parameter Estimates from the Model

The interpretation of these estimates is critical to understanding their impact on the model's predictions.

4.2.2 Optimization Results

The optimization results are summarized as follows:

Metric	Value
Best Log Likelihood	-777770.0688722525
Number of Cases	105
Aggregate Log Likelihood	-7407.33
Log Loss	7407.33

Table 4: Optimization Results

These metrics provide insights into the quality of the estimated model and the optimization process.

5 Discussion

5.1 Practical Implications and Applications

The findings from the model have several practical implications and applications:

- Transport Planning: The estimated coefficients provide insights into passenger preferences across different carriers, time periods, and fare structures. This information can be utilized by transportation planners to optimize service offerings and pricing strategies.
- Policy Making: Understanding the impact of variables such as travel time, connection numbers, and equipment type on passenger choices can assist policymakers in formulating effective transportation policies that cater to passenger needs and preferences.
- Marketing and Promotion: Airlines and transportation providers can leverage these insights to tailor marketing campaigns and promotional offers that resonate with their target audience segments, thereby enhancing customer acquisition and retention strategies.

5.2 Limitations and Challenges

Despite the valuable insights provided by the model, there are several limitations and challenges to consider:

- Implementation Issues: During the implementation of the Larch model, it was observed that the outputs in the data summaries for IDCA data currently ignore the alternative availability criteria. This limitation can affect the accuracy of predicting values, especially when alternative availability plays a significant role in decision-making processes.
- Data Limitations: The accuracy and reliability of the model heavily depend on the quality and representativeness of the input data. Inaccuracies or biases in the dataset can impact the robustness of the model's predictions.
- Assumptions of the Model: The multinomial logit model assumes independence of irrelevant alternatives (IIA), which may not always hold true in real-world scenarios where alternatives are not independent of each other.
- Generalizability: The findings of the model are specific to the dataset and context in which it was developed. Extrapolating these results to different geographical regions or time periods may require additional validation and adjustments.
- Complexity of Decision Making: Passenger decision-making processes are complex and influenced by numerous factors beyond those included in the model. Factors such as personal preferences, loyalty programs, and customer service experiences may not be fully captured by the model.

6 Conclusion

6.1 Summary of Findings

The multinomial logit (MNL) model applied to the itinerary choice dataset using the Larch library has provided valuable insights into passenger preferences and decision-making processes. Key findings from the model include:

- Parameter Estimates: The estimated coefficients reveal significant influences of variables such as carrier choice, elapsed time, fare structures, and time periods on passenger decisions.

- Model Performance: The model demonstrates robust performance metrics, including a high log likelihood and significant t-statistics for most variables, indicating a good fit to the data.
- Practical Implications: The findings have practical implications for transport planning, policy-making, and marketing strategies within the airline industry. They provide actionable insights into optimizing service offerings, pricing strategies, and promotional activities to better meet passenger preferences and enhance overall customer satisfaction.

Overall, the MNL model has proven effective in analyzing and predicting itinerary choices based on the dataset provided. However, it is important to consider the limitations and challenges identified during the implementation phase, such as data limitations and model assumptions. Future research could focus on refining the model with additional variables or exploring alternative choice modeling techniques to further enhance predictive accuracy and applicability.

7 References

1. Swissmetro dataset. Available at: `https://biogeme.epfl.ch/#data`.
2. Larch library itinerary choice example. Available at: `https://larch.newman.me/v5.7.0/example/itinerary.html#`.
3. "Understanding Customer Choices to Improve Recommendations in the Air Travel Industry." Research paper.
4. Train, K., & McFadden, D. (2009). Discrete Choice Methods with Simulation (Second Edition). Cambridge University Press. Available at: `https://eml.berkeley.edu/books/choice2.html`.