Hailey Irwin
October 15th, 2023
AH2170 HT23

A Multinomial Logit Model Statistical Analysis of Transportation Mode Preference in Sweden

I. Introduction

Today, users can choose from many modes of transportation to determine how they get from point A to point B. The ever-evolving dynamic of these multiple transportation modes raises the question of how individuals choose their preferred modes of transportation. Various factors include cost, time, and individual preferences in play. In order to deepen the understanding of travel choice decisions, a travel survey was conducted in Sweden to gather data. The comprehensive dataset comprises 4,000 observations of individuals and notes their chosen travel mode out of those available. The six modes that comprise the data are car driver, passenger, bus, train, walk, and bike.

Assuming travelers will choose among available travel options, we hypothesize that travelers seek to minimize two variables: total travel time and cost. Travels will make trade-offs between the two based on individual preferences and situational needs. For longer-duration trips, travelers prefer lower-cost options (measured as cost per hour). In rank order are train, bus, car passenger, and car diver. For free travel modes, the choice of travel seems less related to the duration of travel and may depend on outside factors. In a-priori hypotheses format, the following are our hypotheses.

- Hypothesis 0: Travelers will choose among available options.
- Hypothesis 1: Travelers aim to reduce both their total travel time and cost.
- Hypothesis 2: Individual preferences and situational play a decisive role in the trade-offs travelers make between travel time and cost.
- Hypothesis 3: As the duration of a trip increases, travelers show an increased preference for low-cost per hour travel options. Travelers rank their preference for low-cost options in the following order: train, bus, car passenger, and car diver.
- Hypothesis 4: For free travel modes, the decision on the mode of transportation is less influenced by the trip duration and external factors play a significant role in the choice of travel mode.

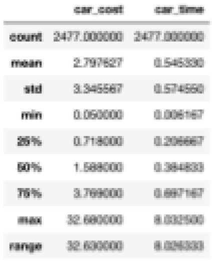
Note that the variables that will be used to answer the questions of how individuals make their choices regarding their preferred modes of transportation are mode, car cost, car time, pass cost, pass time, bus cost, bus total time, train cost, train total time, walk time, and bike time.

Through the wide variety of variables from the dataset used in the analysis, the overarching goal of the study is to highlight the nuanced decisions that individuals make in choosing their travel mode, to understand the trade-offs, and to provide insights into the transportation industry in Sweden.

II. *Method*

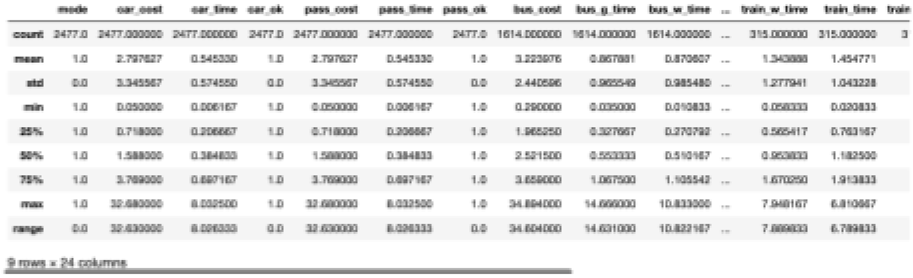
In order to understand transportation preferences in Sweden and through data analytics and a systematic approach, the dataset provided insight into 4000 individuals' transportation choices among six available transportation options. The dataset consisted of 4,000 observations of an individual's chosen mode of transport among the available options of car driver, car passenger, bus, train, walk, and bike. Along with the specific chosen mode, the data also included information associated with each available transportation mode of the 4,000 individual observations. These attributes were cost, travel time, and availability for the respective mode. The overarching methodology used within the case study of travel preference was the multinomial logit (MNL) model, explained later.

A comprehensive dataset exploration was undertaken in the initial data analysis phase to formulate the a-priori hypotheses based on assumed principles rather than actual observation. Through the use of descriptive statistics such as averages, standard deviations, and value ranges, the numbers enabled an understanding of the underlying distribution and trends within the dataset. Note that the dataset was manipulated before the descriptive statistics analysis. This manipulation consisted of dividing cost (SEK) by 10 to equal a dollar, travel time from minutes to hour, and calculating the total time taken for bus and train by adding walking time to stop waiting and travel time. In approaching descriptive statistics, the breaking of the data up into when mode x was chosen, and the other available options better provide the sense of the statistical analysis rather than looking at the dataset as one whole. This is because, in some cases, ie. A train or bus option is only sometimes available when people choose to walk. Tables one and two below provide an example for the descriptive statistics analysis when mode one car driver is chosen. The data analytical process was repeated six times for each mode of travel.



	car_cost	car_time
count	2477.000000	2477.000000
mean	2.797627	0.545330
std	3.345567	0.574550
min	0.050000	0.006167
25%	0.718000	0.206667
50%	1.588000	0.384833
75%	3.769000	0.697167
max	32.680000	8.032500
range	32.630000	8.026333

Table 1: descriptive statistics for car_cost and car_time when mode one, car driver, is chosen.



	mode	car_cost	car_time	car_ok	pass_cost	pass_time	pass_ok	bus_cost	bus_g_time	bus_w_time	...	train_w_time	train_time	train
count	2477.0	2477.000000	2477.000000	2477.0	2477.000000	2477.000000	2477.0	1614.000000	1614.000000	1614.000000	...	315.000000	315.000000	3
mean	1.0	2.797627	0.545330	1.0	2.797627	0.545330	1.0	3.223976	0.867881	0.870607	...	1.343888	1.454771	
std	0.0	3.345567	0.574550	0.0	3.345567	0.574550	0.0	2.440596	0.965549	0.985480	...	1.277941	1.043228	
min	1.0	0.050000	0.006167	1.0	0.050000	0.006167	1.0	0.290000	0.035000	0.010833	...	0.058333	0.020833	
25%	1.0	0.718000	0.206667	1.0	0.718000	0.206667	1.0	1.965250	0.327667	0.270792	...	0.565417	0.763167	
50%	1.0	1.588000	0.384833	1.0	1.588000	0.384833	1.0	2.521500	0.553333	0.510167	...	0.953833	1.182500	
75%	1.0	3.769000	0.697167	1.0	3.769000	0.697167	1.0	3.659000	1.067500	1.105542	...	1.670250	1.913833	
max	1.0	32.680000	8.032500	1.0	32.680000	8.032500	1.0	34.894000	14.666000	10.833000	...	7.948167	6.810667	
range	0.0	32.630000	8.026333	0.0	32.630000	8.026333	0.0	34.604000	14.631000	10.822167	...	7.889833	6.789833	

9 rows × 24 columns

Table 2: descriptive statistics for all variables when mode one, car driver, is chosen.

In continuing the exploratory data analysis, histograms, a type of graph, and correlation coefficients were used to provide insight into the strength and direction of the relationships among the key variables. Like with the statistical analysis, breaking it down by when the mode is chosen highlights the distributive nature of the chosen mode about cost and time. An example histogram of bus total time and cost is shown below in Figure 1. Along with plotting histograms, we analyzed correlation coefficients. Correlation coefficients measure the strength and direction of the relationship between two variables. In Figure 2, such a correlation matrix is presented. The correlation coefficients matrix was also applied to data sets for when each mode was selected for additional data analysis.

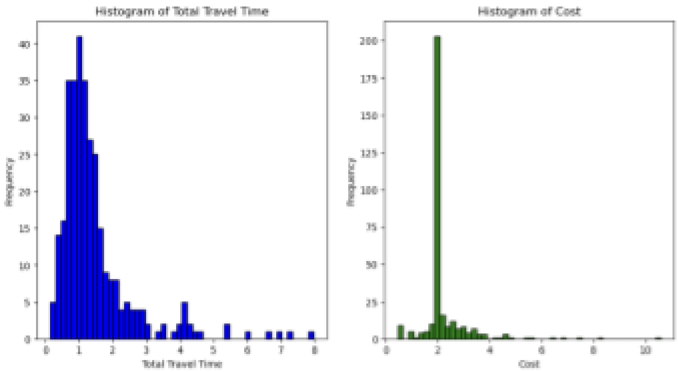


Figure 1: Histogram of bus total time and cost when mode 3, bus, was chosen.

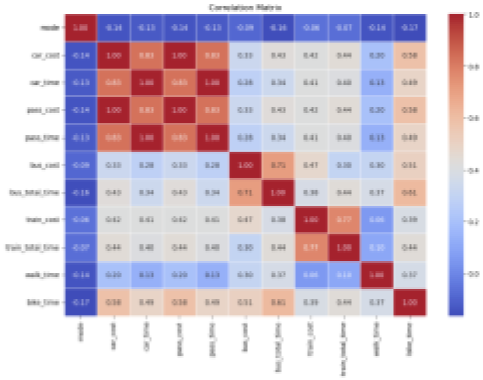


Figure 2: Correlation coefficients heat matrix for the key variables in the study.

Multinomial logit (MNL) model, was the critical method used within the study to understand and predict the mode preferences of the respondents. The multinomial logit model is a statistical method used to model choice outcomes when there are more than two possible alternatives. In the MNL model, it denotes individual making choice, i.e. like the mode of preference choice. Furthermore, each alternative has attributes which influence its utility. Utility is used to quantify the preference of an individual for alternatives. Note the key challenges with MNL models are that individuals are individuals and thus traveler's choice may be perceived differently by different users, i.e. some prefer things more then others (Sharmeen, 2023). Because MNL models are able to handle categorical dependent variables with more then two possible alternatives, MNL models are able to provide model estimation and specification. The MNL model returns the log likelihood and as an input takes the list of coefficients to be estimated, i.e. the data. Moreover, for different models, different predictors are used and the number of coefficients differs. First, a base model is run with all coefficients equaling zero. Then, to optimize the log likelihood function, the use of the optimize.minimize code is applied. The code determines the best-fitting parameters/ coefficients for the MNL model. In model one, all parameters determined critical were used, while in model two, only the statistically significant were used. Note that the ASC, alternative specific constant, presents were in both models, even if not statistically significant. This is because the ASC are there to represent “the net influence of all unobserved characteristics of the individuals and the option in its utility function” (Sharmeen, 2023). The two models which were estimated with the MNL model are in figure three and four below.

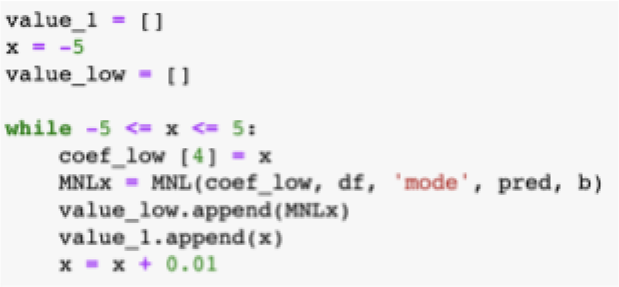
$$
\begin{aligned}
V^i_{j,\text{car}} &= \beta_{\text{car}} t^i_{j,\text{car}} + \gamma_{\text{car}} c^i_{j,\text{car}} \\
V^i_{j,\text{pass}} &= \alpha_{\text{pass}} + \beta_{\text{pass}} t^i_{j,\text{pass}} + \gamma_{\text{pass}} c^i_{j,\text{pass}} \\
V^i_{j,\text{bus}} &= \alpha_{\text{bus}} + \beta_{\text{bus}} t^i_{j,\text{bus}} + \gamma_{\text{bus}} c^i_{j,\text{bus}} \\
V^i_{j,\text{train}} &= \alpha_{\text{train}} + \beta_{\text{train}} t^i_{j,\text{train}} + \gamma_{\text{train}} c^i_{j,\text{train}} \\
V^i_{j,\text{walk}} &= \alpha_{\text{walk}} + \beta_{\text{walk}} t^i_{j,\text{walk}} \\
V^i_{j,\text{bike}} &= \alpha_{\text{bike}} + \beta_{\text{bike}} t^i_{j,\text{bike}}
\end{aligned}
$$

Figure 3: Model one equation with 15 parameters.

$$
\begin{aligned}
V^i_{j,\text{car}} &= \beta_{\text{car}} t^i_{j,\text{car}} + \gamma_{\text{car}} c^i_{j,\text{car}} \\
V^i_{j,\text{pass}} &= \alpha_{\text{pass}} + \gamma_{\text{pass}} c^i_{j,\text{pass}} \\
V^i_{j,\text{bus}} &= \alpha_{\text{bus}} + \beta_{\text{bus}} t^i_{j,\text{bus}} + \gamma_{\text{bus}} c^i_{j,\text{bus}} \\
V^i_{j,\text{train}} &= \alpha_{\text{train}} + \beta_{\text{train}} t^i_{j,\text{train}} + \gamma_{\text{train}} c^i_{j,\text{train}} \\
V^i_{j,\text{walk}} &= \alpha_{\text{walk}} + \beta_{\text{walk}} t^i_{j,\text{walk}} \\
V^i_{j,\text{bike}} &= \alpha_{\text{bike}} + \beta_{\text{bike}} t^i_{j,\text{bike}}
\end{aligned}
$$

Figure 4: Model two equation with 14 parameters.

Interpretation of the results and tests from our MNL models followed, with paying close attention to the statistically significant parameters and the log of likelihood output. What followed was a deeper insight into our models and the analysis of the estimation results. Tasked with selecting two parameters for our key variables, one with a high t-value and another with a low t-value, we examined how the log-likelihood changes when the coefficient of that parameter varies from negative to 0 to positive. We were tasked with writing the code, as shown in Figure 5 below. The analysis helped us understand the influence of the parameters on the model's likelihood. Furthermore, in comparison, we looked at the t-test results of the log-likelihood ratio test for the two parameters compared to our picked model from the MNL estimation and then when we removed the two parameters individually. Thus, we compared the best overall MNL estimation model with a model with the low t-statistic parameter removed and vice versa. This was done to see how a low and a high t-statistic parameter affects the overall log-likelihood. Further, the aggregate sample demand for the alternative modes with the estimated parameters was made to calculate the probability of each individual choosing a particular mode of transportation given the attributes and predicting the expected number of individuals who would choose each mode. The aggregate sample demand was visualized when respective attributes were altered to understand the sensitivity and flexibility of mode choice to specific attributes. Throughout the methodology, the goal was to understand the attributes and their influence on an individual's mode choice.



```
value_1 = []
x = -5
value_low = []

while -5 <= x <= 5:
    coef_low [4] = x
    MNLx = MNL(coef_low, df, 'mode', pred, b)
    value_low.append(MNLx)
    value_1.append(x)
    x = x + 0.01
```

Figure 5: Code for having one parameter change and how it affects the log- likelihood. Used in plot show in result section.

III. *Results*

Through the manipulation and exploration of Sweden's 4,000 observational modes of transportation preference data, the results are divided into descriptive statistics, model estimation and specification, and analytics of the estimation result.

The descriptive statistics average values, standard deviations, and ranges of values for the variable formulated the hypothesis. The tool enables you to break down the statistics value into a comparison ratio of cost per hour. Using the 50%, the variable less influenced by outliers, unline mean, which could be skewed, ratios were done for each mode. Note that for the free transportation modes, only time was looked at. The ratios are as follows: car driver: 4.126465245963834, car passenger: 2.778317193747558, bus: 1.7437436683673604, train: 1.0334759885692986, walk: 0.65225, and bike: 0.442833. The ratios highlight that if cost were the critical concern, trains would cost the least money to take you the longest time. Bus, car passenger, and car driver follow this. In terms of the free transportation modes, the time which is chosen is short time lengths. The histograms highlight the lack of frequency with which trains and buses were taken compared to cars. This result suggests an individual preference for convenience, no matter the cost factor. Also, it could indicate the need for more public transportation options for routes. Figures six, seven, and eight are histograms supporting the claim that individual preferences play a substantial role in transportation methods.

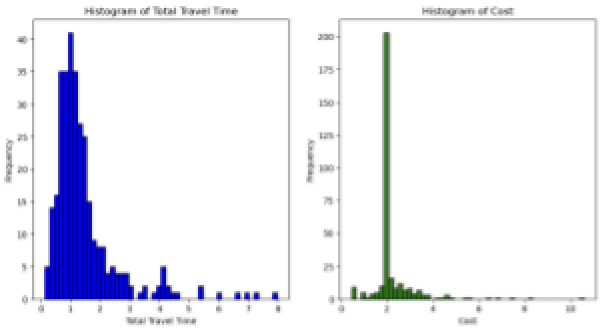


Figure 6: Cost and total time histograms when mode 3, bus, was chosen.

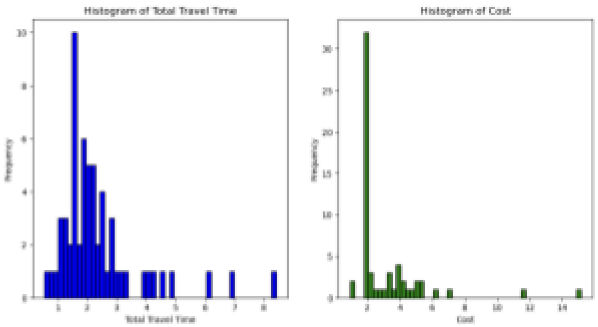


Figure 7: Cost and total time histograms when mode 4, train, was chosen.

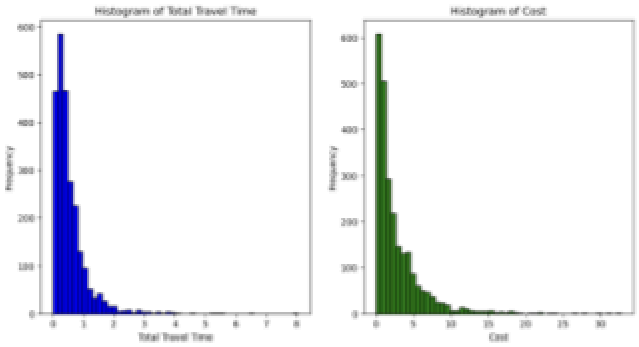
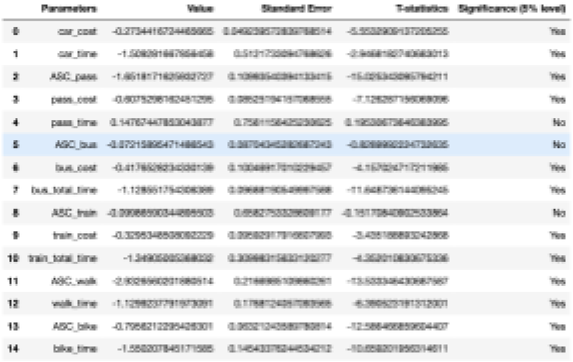


Figure 8: Cost and total time histograms when mode 1, car driver, was chosen.

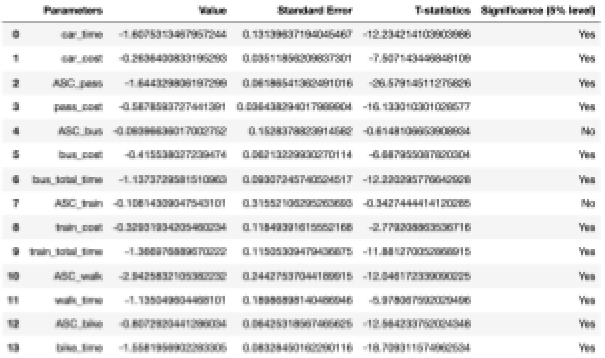
The correlation heat matrix is a representation between multiple variables of how closely related different variables are, with 1 - red being perfect and 0 - blue being none. When examining the matrix in figure two above, one notices a highly positive relationship between car users, cost and time, and passenger users. Of the other correlations, all maintain similar around 0.4 to 0.5, indicating a moderate positive relationship, while a relationship exits the result indicates that other factors and preferences likely play a role. Note that the walking time relationship often hovers around 0.1 to 0.2 compared to the others. Thus, having a low positive correlation but a fair assumption could indicate that walk time would impact individual preferences to choose other modes of transportation. The highly correlated attributes are train time and cost, indicating that preference for trains is closely related and not as influenced by cost/time factors.

The two model specifications were estimated using the MNL model, the second results section group within the lab. The MNL model produced two table datasets highlighting parameters' value, standard error, t-statistics, and, if statistically significant, at a 5% level. The first model was run with all 15 deemed key parameters, as determined in the a-priori hypotheses, while the second was run with all key parameters minus any key parameters that were not deemed statistically significant. Within the text, this concluded to exclude pass_time as a parameter. Model one variables were [['car_time,' 'car_cost'], ['pass_time', 'pass_cost'], ['bus_cost,' 'bus_total_time'], ['train_cost,' 'train_total_time'], ['walk_time'], ['bike_time']] and model two variables were [['car_time,' 'car_cost'], ['pass_cost'], ['bus_cost,' 'bus_total_time'], ['train_cost,' 'train_total_time'], ['walk_time'], ['bike_time']]. The two tables are below in tables three and four. Furthermore, through optimization of the log-likelihood function with 15 parameters, the log-likelihood output was -3848.0403935477702, while in model two, an optimization of the log-likelihood function with 14 parameters, the log-likelihood output was -3848.239141451594. The log-likelihood helps contextualize how well a statistical model explains observed data. Thus, model two does not fit the data and model one. We are looking for the lower log-likelihood because we are using the optimization.minimize code.



	Parameters	Value	Standard Error	T-statistics	Significance (5% level)
0	car_cost	-0.2734416724465665	0.049239572839788514	-5.5532909137205255	Yes
1	car_time	-1.508091667656458	0.5121723094768626	-2.9468192740683013	Yes
2	ASC_pass	-1.6518171625932737	0.1098354039413315	-15.025343095784211	Yes
3	pass_cost	-0.6075298162451295	0.08525194157066556	-7.126267156068096	Yes
4	pass_time	0.14767447853043877	0.7561156425230925	0.19530673646383995	No
5	ASC_bus	-0.07215895471488543	0.08704345282687243	-0.8289892234732635	No
6	bus_cost	-0.4176528234030139	0.10048917010229457	-4.157024717211966	Yes
7	bus_total_time	-1.128551754308389	0.09688190549667568	-11.648736144095245	Yes
8	ASC_train	-0.09986500344866503	0.6582753328628177	-0.15170840802533864	No
9	train_cost	-0.3295348508092229	0.0958291791607958	-3.435188893242868	Yes
10	train_total_time	-1.349050052668032	0.30998315831202277	-4.352010830675206	Yes
11	ASC_walk	-2.9326560201880514	0.21698851098662261	-13.533346430687587	Yes
12	walk_time	-1.1298237791973091	0.17681242672630566	-6.390523191312001	Yes
13	ASC_bike	-0.7956212295426301	0.06321242588780814	-12.586468859624407	Yes
14	bike_time	-1.550207845171585	0.14543378244634212	-10.659201956314611	Yes

Table 3: Model One, 15 parameter, dataset.



	Parameters	Value	Standard Error	T-statistics	Significance (5% level)
0	car_time	-1.6075313467957244	0.13139637194045467	-12.234214103903866	Yes
1	car_cost	-0.26364008303195263	0.03511856209837301	-7.507143446848109	Yes
2	ASC_pass	-1.644329806197299	0.06186541362491016	-26.57914511275826	Yes
3	pass_cost	-0.5878593727441391	0.036438294017989904	-16.133010301026577	Yes
4	ASC_bus	-0.09399638017002752	0.1528378823914582	-0.6148109653908934	No
5	bus_cost	-0.4155380272239474	0.06213229930270114	-6.687955087820304	Yes
6	bus_total_time	-1.1373729581510963	0.09307245740524517	-12.220295776642926	Yes
7	ASC_train	-0.10814309047543101	0.31552106295263693	-0.3427444414120265	No
8	train_cost	-0.32931934205460234	0.11849391615552168	-2.779208863536716	Yes
9	train_total_time	-1.366975889670222	0.11505339479436875	-11.881270052866915	Yes
10	ASC_walk	-2.9425832105382232	0.24427537044189915	-12.046172339090025	Yes
11	walk_time	-1.1350496034468101	0.18986898140486946	-5.978087592029496	Yes
12	ASC_bike	-0.8072920441286034	0.06425318567465625	-12.564233762024346	Yes
13	bike_time	-1.5581956902283305	0.08328450162290116	-18.709311574862534	Yes

Table 4: Model Two, 14 parameter, dataset.

The MNL results affected the input for the analytics of the estimation result. The estimation MNL showed that model one with 15 parameters and the log-likelihood output was -3848.0403935477702 better fit the model. Using model one, we were tasked with picking a parameter with a low and high t-statistic and testing the log-likelihood against variations in this parameter revealed. The parameter with a high t-statistic is bus total time with a statistic of -11.648736144095245, and the parameter with a low t-statistic is pass time with a statistic of 0.19530673646383995. In Figures nine and 10, the log-likelihood against variations of the parameter coefficient is shown. Note that the curves of each parameter are below and that the better the log-likelihood is, the better the selected parameters make a better-predicting model. Thus, when comparing the two curves, one can see that the change in the pass time has a larger impact on the log-likelihood, whereas the bus total time does not affect the log-likelihood.

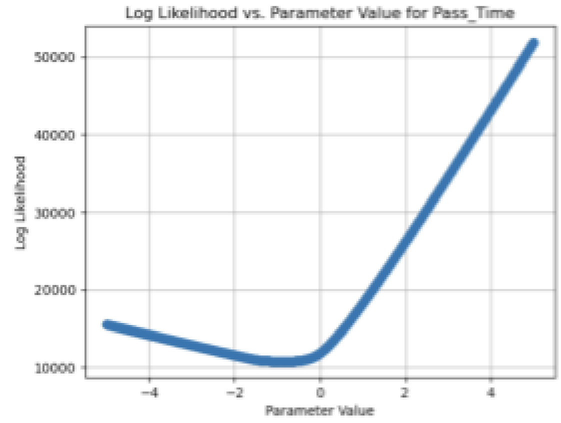


Figure 9: Log Likelihood vs. Parameter Value for Pass Time.

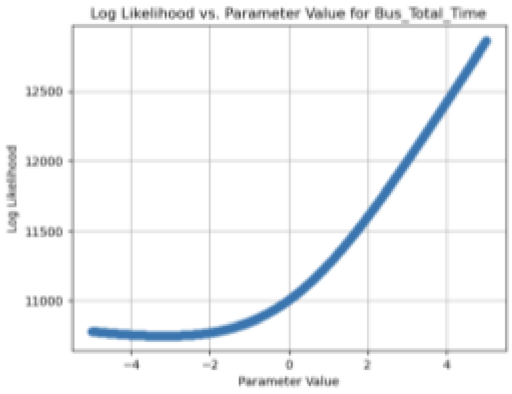


Figure 10: Log Likelihood vs. Parameter Value for Pass Time.

Moreover, the two parameters were compared using a t-test and log-likelihood ratio test. First, for the t-statistic comparison, we pulled the t-statistic and standard deviation from model one for pass time and bus total time. Using the two numbers, the degree of freedom and p-value (0.8451627458611308 and 0.0), respectively, were calculated in order to determine if they were statistically significant. Note that the pass value is not statistically significant, and the bus total time is statistically significant. Furthermore, when comparing the log-likelihood ratio test, we compared model one and the 15 parameters and two different models, model two and three, each with 14 parameters, as we removed pass time (model two) and bus total time (model three). The log-likelihood ratio test for pass time resulted in a degree of freedom 1, a log-likelihood ratio of 0.39749580764782877, and a critical value of 3.841458820694124. For the log-likelihood ratio, a larger ratio represents a larger difference between the two models because we are optimize.minimize, the smaller log-likelihood ratio represents how well the model fits the data. The critical value is compared to the log-likelihood ratio to see if it is statistically significant, with log-likelihood ratio > critical value meaning it is. For the bus total time, the degree of freedom is 1, a log-likelihood ratio of 203.0706325743713, and a critical value of 3.841458820694124. This aligns with the previous data, indicating that bus total time is statistically significant and model three does not fit the data nor pass value, the non-statistically significant.

The aggregate sample demand for the alternative modes is the predicted number of individuals choosing each mode. The expected values are: driver: 720.754239, car passenger: 1,874.59102, bus: 117.847755, train: 1.81549618, walk: 284.62399, and bike: 1000.36750. Note the low expected value in trains and the trend of preference for car drivers and passengers. Furthermore, we used the plot to show the impact on aggregate sample demand as different attributes change in value. Figure eleven below shows six different graphs of the aggregate sample demand determined by attributes change in value. As shown in the graphs, the car driver time, bus total time, train total time and pass cost have a low effect on the aggregate sample demand. The car driver cost and walk time both have a negative effect on the aggregate sample demand, indicating that as car driver cost/ walk time increases, the demand for such a mode decreases. These are just a sample of the aggregate sample demand curve.

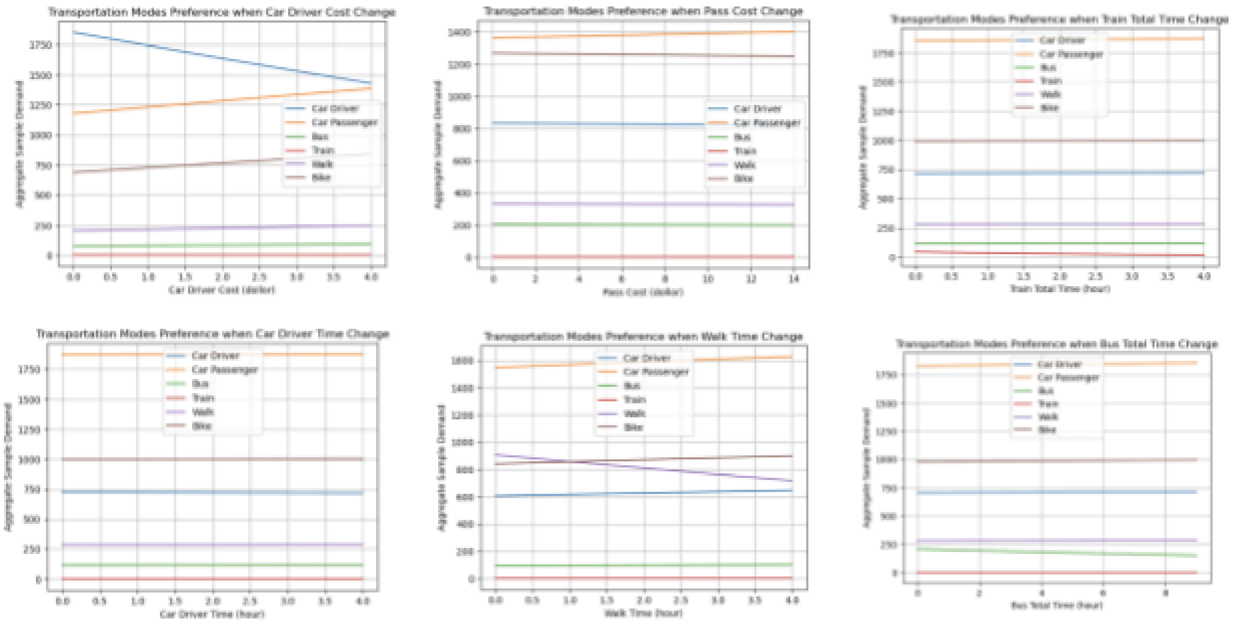


Figure 11: Example Aggregate Sample Demand Change when Attributes Change.

IV. *Discussion*

The multinomial logit model was a favorable tool for analyzing multivariable datasets, like transportation mode choice. Using parameters to determine the utility and log-like home enabled the capability to understand and uncover patterns and relationships within the data deeply. However, the major limitation within the dataset was the data provided. Only availability, cost, and time were provided. Information that would make the model study more comprehensive would be distance, demographics, and locations. More details would have provided insight into the use of travel and patterns within geographic areas. Moreover, it is important to highlight that a deeper analysis is always available. While, in the case study, we have only touched a surface level, diving deeper into the dataset could provide a more comprehensive account of travel habits in Sweden. In conclusion, the study provided valuable insights into applying the multinomial logit model and log-likelihood. The findings in Python, while challenging, provided practical practice and applications. It showed the criticalness of statistical analysis, the importance of data-driven research, and how important proper surveying is to create a comprehensive and extensive image.

V. *References*

AH2170-AG2301 Transportation Data Collection and Analysis. (2020, September). *Mode Data*.

Sharmeen, F. (2023). *Lecture 8: Choice Modelling, contd.*