

Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology, Kharagpur

Lecture - 37
Mode Choice Model

(Refer Slide Time: 00:26)

CONCEPTS COVERED

- Mode choice model using binary logistic regression in SPSS
- Mode choice model using multinomial logistic regression

In this lecture of mode choice modelling, the different concepts covered are, mode choice model using binary logistic regression in SPSS; mode choice model using multinomial logistic regression.

(Refer Slide Time: 00:45)

Mode choice model using binary logistic regression in SPSS

Variables	Datatypes
Total expected time of journey	Scale
Age	Scale
Gender	Nominal
Two-wheeler ownership	Nominal
Car ownership	Nominal
Income category	Nominal
Bus fare	Scale
Relative occupancy in bus (crowding)	Scale
Bus Headway	Scale
Expected delay in journey by bus.	Scale
Auto fare.	Scale
Waiting time for auto-rickshaw	Scale
Bus Reliability	Derived from
Safety in bus	Satisfaction data
Auto-rickshaw Reliability	(Ordinal)
Safety in Auto-rickshaw	
Chosen mode	Nominal(classes)

- Logistic regression is essentially a tool for classification based on nominal dependent variable.
- The data used given in the demonstration belongs to a mode choice survey for bus and auto-rickshaw.
- Both Revealed preference (RP) and Stated preference (SP) data are present in the dataset.
- The demonstration has separate models based on RP data and SP data.

NPTEL Online Certification Courses
IIT Kharagpur

Mode choice model using binary logistic regression in SPSS:

The demonstrated model is a part of a study undertaken in Kolkata, India, along one of the corridors in the city namely, Eastern Metropolitan bypass. The study involved a part which required the determination of mode choice model between auto-rickshaw and bus. Other modes can also be included, but they have been kept out of this particular example. These two modes have been chosen as these two modes were found to be competing with each other for ridership and there is cross elasticity between these two modes.

In the entire corridor only a few stretches has got auto-rickshaw routes as well as bus routes overlapping each other. In most of the stretches, there were no auto-rickshaw routes along the corridor. This leaves a scope of introducing such modes/routes after prior evaluation. Thus, new scenarios can be analyzed to see if the total pollution decreases or there is an increase in the efficiency of service along the corridor through introduction of these new modes/routes. Accordingly, both revealed preference (RP) survey as well as stated preference (SP) surveys were conducted. In the revealed preference survey, the existing choice of people for a particular mode along that corridor, the reasons for choosing that particular mode, and the individual characteristics were enquired. The number of auto-rickshaw users were found to be pretty less. So, along with RP surveys, SP surveys were also conducted, in which respondents were subjected to hypothetical scenarios where auto-rickshaws were assumed to be introduced in that particular corridor. In order to understand the choice behaviour of the respondents, these scenarios were designed in such a way that the respondent could carry out a mental evaluation to estimate the utility of either modes and choose one of the two modes for a given hypothetical scenario. Eventually these two datasets, RP data and SP data, were used to estimate separate mode choice models.

The different variables that have been used in the example to understand the travel behaviour as well as to create the scenarios are, total expected time of journey; age; gender of the individual; two-wheeler ownership; car ownership; income category; bus fare; relative occupancy in bus; bus headway; expected delay in journey time by bus; auto-rickshaw fare; waiting time for auto-rickshaw. Response on many parameters which talked about satisfaction from different attributes for the two modes were also considered. All of these were reduced to four factors related to safety and reliability of bus and auto-rickshaw. The methods of reducing variables to factors is covered in the next lecture.

(Refer Slide Time: 04:36)

Survey Questionnaire

State the existing situation for the below listed attributes for buses and auto-rickshaws:

Mode	Attribute	Value
Bus	Boarding and alighting time	seconds
	Time between two consecutive buses on your route	minutes
	Constant waiting time for an auto-rickshaw	minutes
Auto-rickshaw	Average number of passengers travelling in an auto-rickshaw	number

For each item identified below, rate your level of satisfaction for the listed attributes:

Indicator	Rating Scale				
	Very Poor (1)	Poor (2)	Average (3)	Good (4)	Very Good (5)
Bus Service					
Waiting time at bus stop					
On-time performance of bus services					
Safety from road accidents while travelling in a bus					
Boarding and alighting time					
Constant headway of bus					
Quality of bus driving practice					
Quality of customer service (availability of helpline number, assistance provided by staff, etc.)					
Route network information					
Real time information on arrival and departure of buses					
Real time information on disruption of services					
Safety inside buses during early morning and late night hours					
Safety inside bus from assault and harassment					
Safety from theft and robbery while travelling in a bus					
Safety from accidents while travelling in a bus					
Auto-rickshaw Service					
Waiting time for auto-rickshaw					
On-time performance of auto-rickshaws					
Safety from road accidents while travelling in an auto-rickshaw					
Quality of auto-rickshaw driving practice					
Quality of customer service (availability of helpline number, assistance provided by staff, etc.)					
Safety inside auto-rickshaws during early morning and late night hours					
Safety inside auto-rickshaws from assault and harassment					
Safety from theft and robbery while travelling in an auto-rickshaw					
Safety from accidents while travelling in an auto-rickshaw					

Kindly state your current trip details:

Origin (Location)	Origin to Stop 1			Stop 1 to Transfer Station			Transfer Station (Location)	Stop 1 to Transfer Station			Transfer Station (Location)	PT Cost
	Step 1 Location	Access Mode	Access Distance	Access Cost	PT Mode	PT Distance		PT Cost	PT Mode	PT Distance		

Kindly state your alternate modes of transport:

Step 1 Location	Transfer Station to Stop 1			Stop 1 to Destination			Destination (Location)
	PT Mode	PT Distance	PT Cost	Egress Mode	Egress Distance	Egress Cost	

Trip Purpose: Work / Education / Health / Shopping (in terms of grocery shopping to the market) / Leisure / Home / Others

Alternate Modes: Auto-rickshaw / Private Vehicle / Taxi / BRTS / Metro / Others

Survey Questionnaire:

The figure shows parts of the actual questionnaire used in the study. The questionnaire is segregated into various sections. The first section takes information about the user profile; **boarding-alighting stops**, **age**, **gender**, if he/she **regularly travels** in this route, **frequency** of using bus service, possession of **driving license**, private **vehicle ownership**, **income** category.

The next section takes information about the characteristics of the trips; **origin**, **destination**, bus stop **location**, **boarding** bus stop, **alighting** bus stop, **mode to the stop**, **access distance**, **access cost**, **chosen mode** is auto-rickshaw or bus, **travel distance**, **travel cost**, **transfer station(s)**, and egress details as well. **Egress** here refers to the part of travel after alighting the transit mode and reaching the destination. Details regarding egress include; mode and time. If the respondents cannot avail their usual mode, the alternate modes available to them were also asked. The **purpose of the trip** was also asked in order to be able to segregate work trips, shopping trips, leisure trips, etc. The **party size**, or the size of the group, in case the respondent travels along with other people in a group, was also asked. Some information pertinent to bus and auto-rickshaw were collected like; boarding-alighting time and headway of bus service for bus; and waiting time and the average number of people travelling in auto-rickshaw.

The next section had questions regarding the qualitative parameters, which the respondents were asked to rate using a 5-point satisfaction scale for each parameter and each mode. The parameters

were mainly based on operation perspective and safety perspective. **Waiting time** at bus stop; **on-time performance** of bus service; **delay** in total journey time; **boarding and alighting time**, etc. were based on the operation of service. Safety during **early morning and late nights**; safety from **assault**, safety from **theft and robbery**; safety from **accidents**, etc. were also considered. Some parameters specific to bus such as **information disbursement in real time** regarding route network, disruption, and arrival-departure times were also included.

(Refer Slide Time: 07:47)

SP Card

Pre-processing of data

No missing values for any variable.
Outlier detection and removal from the dataset.
(plotting dependent vs. independent variables)
In case scripts have to be written to perform regression, the data needs to be in a particular format (csv, dat, json, etc.).

	BUS FARE	WAITING TIME AT THE BUS STOP	DELAY IN JOURNEY TIME BY BUS	SEAT AVAILABILITY IN BUS	AUTO-RICKSHAW FARE	Your Choice
Current Trip Details	Rs. _____	_____ Min	_____ Min		Rs. _____	<input type="checkbox"/> Bus <input type="checkbox"/> Auto rickshaw
Scenario 1	75% more than current ticket price	5 minutes	Same as current	All seats full + few passengers standing	50% more than current fare	<input type="checkbox"/> Bus <input type="checkbox"/> Cannot Choose <input type="checkbox"/> Auto rickshaw
Scenario 2	Same as current ticket price	5 minutes	5 minutes	All seats + few seats empty	75% more than current fare	<input type="checkbox"/> Bus <input type="checkbox"/> Cannot Choose <input type="checkbox"/> Auto rickshaw
Scenario 3	50% more than current ticket price	10 minutes	20 minutes	Same as current	50% more than current fare	<input type="checkbox"/> Bus <input type="checkbox"/> Cannot Choose <input type="checkbox"/> Auto rickshaw
Scenario 4	50% more than current ticket price	Same as current	Same as current	All seats full + 0 passengers standing	75% more than current fare	<input type="checkbox"/> Bus <input type="checkbox"/> Cannot Choose <input type="checkbox"/> Auto rickshaw

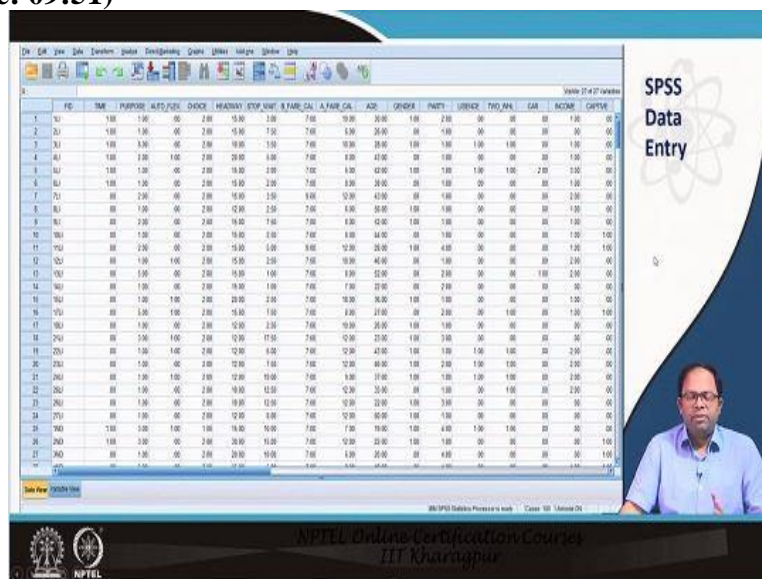
NPTEL Online Certification Course
IIT Kharagpur

The last section of the questionnaire consisted of the SP cards. This card has 5 rows, the first one being the current scenario and the rest four being hypothetical scenarios. These were hypothetical scenarios designed by varying certain variables like; **bus fare**, **auto-rickshaw fare**, **waiting time at bus stop**, **delay in journey time by bus**, **seat availability**. For each scenario, respondents were asked to select a mode (bus or auto-rickshaw) assuming auto-rickshaw waiting time does not exceed 3 minutes, and all other factors/ variables being same as existing. There were cases where auto-rickshaw service did not exist, for such instances, the respondents were asked to consider the bus fare as the base auto-rickshaw fare. These scenarios have been designed using fractional factorial design, that has been covered in an earlier lecture. Each of the variables used to design the scenarios had different levels. For example, bus fare varied between same as current, 50% more, and 75% more; waiting time at bus stop varied between same as current, 5 minutes, 10 minutes; seat availability varied between ‘all seated and few empty seats’, ‘all seated and no empty seats’, ‘all seated and some standing’, ‘over crowded’; etc.

Pre-processing of data:

Once data collection is done, the database needs to be pre-processed before it is subjected to analysis. One of the first thing that is checked is ‘**missing values**’. Often due to the shortage of time, or the lack of information with the respondent, or due to any human error by the surveyor, few information/fields in the questionnaire form is not recorded. Such questionnaire forms/data should be removed from the database altogether if the total number of questionnaires with missing values is less than 10% of the whole sample size. If they are more than 10%, specialized statistical techniques needs to be undertaken to fill in the missing values. **Outliers** are another thing that needs to be removed. The data needs to be saved in particular format that is compatible with the statistical analysis tools/software. ‘*.dat’, ‘*.csv’, ‘*.json’ few of the popular formats.

(Refer Slide Time: 09:51)



The image shows a screenshot of the SPSS Data Entry window. The window displays a data grid with columns for variables and rows for data entries. The variables listed in the columns include FE, BM, PMPROB, ALCOHOL, DRUGS, HEARTWAY, STP, VOT, B, PAR, CA, A, PAR, CA, AGE, GENDER, PAPER, URBAN, THO, AN, CAR, NCOAR, and GPTSA. The data grid shows numerical values for each variable across multiple rows. On the right side of the window, there is a logo for 'SPSS Data Entry' and a video feed of a presenter. At the bottom of the window, there is a footer for 'NPTEL Online Certification Courses' and 'IIT Khariapar'.

SPSS data entry:

To import the dataset in SPSS, the sequence of steps are *File* → *Open* → *Data* → *select dataset file*. When the file gets imported, a window appears with two view tabs i.e. Data view, and Variable view. The Variable view displays the different variables used in the present study. The attributes of the variable are variable name (Gender, Age), data type (numeric or string), column width, value (label for each value of variables), role (input, target, or both), and measurement level (nominal, ordinal, or scale). Also, these attributes can be modified as per requirement.

(Refer Slide Time: 10:22)

Model with SP data

The variables present in the SP choice cards and some socio-economic variables were considered while estimating the model.

Null hypothesis: Constant (or intercept) and coefficient of all the variables are 0.
A significant chi-square value means we reject the null hypothesis and we can say that intercept and the coefficients are non-zero.


Likelihood ratio and pseudo r-square:
Likelihood ratio can be a good measure to compare between multiple models using the same data. Higher the value, better the model.

Logistic regression does not have R-square like linear regression. Cox & Snell R² and Nagelkerke R² are few of the pseudo-r square estimates for logistic regression.

		Chi-square	df	Sig.
Step 1	Step	428.323	7	.000
	Block	428.323	7	.000
	Model	428.323	7	.000

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	626.395 ^a	.428	.573

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.



NPTEL Online Certification Course
IIT Kharagpur

Results of the model:

The steps to run the model is same as described in residential location choice and is also described below. The results shown is for a model developed with the SP data only which has variables mentioned in the SP cards, and some socio-economic variables. In the model results, the first thing is the **omnibus test** of model coefficient, where the model chi-square statistic is given in which ‘**df**’ signifies degree of freedom, or the number of variables included in the model. Null hypothesis is a model in which the constant or the intercept is present and the coefficients of all variables are 0. A significant chi square value implies the rejection of the null hypothesis and it can be claimed that the intercept and the coefficients are non-zero.

The next table is of **model summary**, which gives the likelihood ratio of initial and final log likelihood; the Coz and Snell R-square; and the Nagelkerke R-square. Both the R-squares are pseudo R-square values for logistic regression since, a conventional R-square cannot be estimated for logistic regression. Likelihood ratio can be a good measure to compare between multiple models using the same data. Higher the value better is the model. For example, comparing a constant only model and a model with parameters, the model for which the likelihood ratio value is more, is a better model. Other models can also be developed using the same dataset by choosing different variables based on apriori knowledge. All these models can be compared using the likelihood ratio test for a better fit.

(Refer Slide Time: 12:56)

The classification table shows the accuracy of the model in correctly predicting the classes for the dataset itself which is 80%.


Since the dataset has both **RP and SP data**, only the SP data was **SELECTED** and the RP ones were **UNSELECTED**.

Classification Table^a

Observed	SPSS_choice	Predicted						
		Selected Cases ^b			Unselected Cases ^c			
		SPSS_choice	Percentage Correct	SPSS_choice	Percentage Correct	SPSS_choice	Percentage Correct	
00	1.00	353	71	83.3	101	159	38.8	
1.00	1.00	79	264	77.0	1	29	96.7	
Overall Percentage							80.0	44.8

a. The cut value is .500
 b. Selected cases SP EQ 1
 c. Unselected cases SP NE 1

True Positive True Negative Accuracy



The **classification table** shows the prediction accuracy of the model. The model is developed using a dataset, and this table shows how many observations of the same dataset have been classified correctly after being subjected to the model. The model shown exhibits 80% accuracy. Since the data selected were only the SP responses, the classification of ‘selected cases’ shows the results for the SP data. The classification result under ‘unselected cases’ is for the RP data, which is not of concern as of now.

(Refer Slide Time: 13:24)

The model here predicts the probability of choosing BUS.

Table of model estimates: Coefficient of variables, level of significance (p-value) and Exp(B) or Odds ratio.


Variables in the Equation

Step 1 ^a	Variable	B	S.E.	Wald	df	Sig.	Exp(B)
	JOURNEY_TIME	.153	.014	128.563	1	.000	1.165
	MALE(1)	.562	.216	6.738	1	.009	1.754
	BUS_FARE_CALC	-.414	.042	97.897	1	.000	.661
	BUS_CROWDING	-.867	.379	5.229	1	.022	.420
	BUS_DELAY	-.079	.016	24.699	1	.000	.924
	BUS_RELIABILITY	1.087	.256	18.010	1	.000	2.966
	AUTO_RELIABILITY	-.582	.154	14.215	1	.000	.559
	Constant	1.775	.869	4.167	1	.041	5.899

a. Variable(s) entered on step 1: JOURNEY_TIME, MALE, BUS_FARE_CALC, BUS_CROWDING, BUS_DELAY, BUS_RELIABILITY, AUTO_RELIABILITY.

Significance (p-value) more than 0.05 is also included

- As the journey time increases i.e. long distance journey, people are willing to opt for bus.
- As bus fare, reliability of auto-rickshaw and expected delay by bus (from past experience) increases, the probability of selection of bus decreases.
- Similarly other variables can be interpreted.



The table named ‘**Variables in the Equation**’ shows the regression coefficients, level of significance (**p-value**), and the odds ratio (Exp B or **e^B**) for the variables included in the model. The given model predicts the probability of selection of bus given the values of the variables in the model. All the variables selected in the model are significant at 95% confidence interval ($p \leq$

0.05). Other variables in the dataset were not included either due to the lack of statistical significance, due to lack of theoretical evidence, or due to counter intuitive estimates.

As per the model, as journey time increases i.e. longer the distance of journey, people are willing to select bus, which makes sense. As bus fare; reliability of auto rickshaw; and expected delay by bus from past experience increases, the probability of selection of bus decreases.

(Refer Slide Time: 14:40)

Correlation Matrix

	Constant	JOURNEY_TIME	MALE()	BUS_FARE_CALC	BUS_CROWDING	BUS_DELAY	BUS_RELIABILITY	AUTO_RELIABILITY
Step 1 Constant	1.000	.017	-.165	-.381	-.564	-.020	-.494	-.329
JOURNEY_TIME	-.017	1.000	.043	-.311	-.109	-.152	.025	-.082
MALE()	-.165	.043	1.000	-.153	.015	-.054	.152	-.083
BUS_FARE_CALC	-.381	-.311	-.153	1.000	.119	.063	-.169	.120
BUS_CROWDING	-.564	-.109	.015	.119	1.000	-.037	.091	-.011
BUS_DELAY	-.020	-.152	-.054	.063	-.037	1.000	-.096	-.009
BUS_RELIABILITY	-.494	.025	.152	-.169	.091	-.096	1.000	-.326
AUTO_RELIABILITY	-.329	-.082	-.083	.120	-.011	-.009	-.326	1.000

The correlation matrix shows the amount of correlation between each pair of variables.

Two variables can be considered to be correlated if the correlation coefficient is greater than or equal to 0.5 (exact critical value is debatable and depends on the context)

According to the given data, none of the variables in the model are highly correlated with each other.

Pairwise correlation between variables can also be estimated which can be seen in the **correlation matrix**. This can be used to evaluate if some variables are highly correlated with another variable. Different people have used different values as the limit to determine correlation. In this particular case, value of 0.5 has been taken as the limit. According to the given data, none of the variables in the model have been found to be highly correlated with each other.

(Video Starts: 15:22)

Steps to estimate a mode choice model using SPSS:

The following steps have been followed to estimate the model shown above. Similarly, other variations of the same model can be estimated, and a different model with different dataset can also be estimated using the same steps.

- Analyze → Regression → Binary logistic → Select independent variables
- Select dependent variable i.e. CHOICE variable
- Select **Selection variable** (if any) and define the selection rule
- From the categorical button, select the reference category for categorical variables (if any)
- From the options button, select the estimates that are required in the output → Click OK

From the **analyze** option regression needs to be selected. Among the various **regression** options available, **binary logistic** needs to be selected. in the Logistic regression tab, there are various sections like; **dependent variable**, **independent variable**, **selection variable**. From the list of variables available, desired independent variables need to be included in the independent variable section. Similarly, appropriate variables need to be selected for dependent variable and selection variable (if any). In this case, since the model is only based on SP data, nominal variable SP is selected to differentiate between the SP data and RP data. The selection rule needs to be defined, which is, in this case **SP= 1**. In the **categorical** button, reference group all the categorical variables (if any) needs to be selected. in the **options** button, the estimates which are required in the output needs to be selected. in order to carry out the estimation, click **OK**.

(Video Ends: 16:44)

(Refer Slide Time: 16:47)

Model with RP data

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a AGE	.042	.017	5.838	1	.016	1.043
BUS_HEADWAY	-.207	.058	12.607	1	.000	.813
Constant	-.669	1.025	.426	1	.514	.512

a. Variable(s) entered on step 1: AGE, BUS_HEADWAY.

Correlation Matrix

	Constant	AGE	BUS_HEADWAY
Step 1 Constant	1.000	-.661	-.756
AGE	-.661	1.000	.051
BUS_HEADWAY	-.756	.051	1.000

Model estimated using RP data:
 Passenger age and bus headway were found to be significant.

NPTEL Online Certification Courses
 IIT Kharagpur

Mode choice model with RP data:

Another model was estimated using the RP data, which shows the choice behaviour under real life scenarios. In the study corridor, auto rickshaw and bus routes both existed only in a few stretches. In addition, it is understood that there are less variations in RP data. For example, in the market, all the modes competing with each other have more or less similar fare structure and service quality. Thus the variation in the data is very limited. From the estimation point of view, the more the variation in a data set, easier is the estimation process. So, models based on RP data are not that good. A poorly estimated model will result in poor probability estimation. In that sense, SP data is better than RP data as it allows to introduce the variability through different levels of service, that could be provided in future. This also implies that SP data is hypothetical and hence combining the two datasets for estimation is a better approach, which has been covered in another section.

In the RP model shown, there are only two variables; age and bus headway. As per the RP data, only these two variables seem to be impacting the mode choice behaviour of people travelling along the network corridor under consideration. No other bus related variables, or any of the variables related to auto-rickshaw were found to be significant. This RP based model does not allow the testing of different policies except for may be changing of bus headway. Thus, SP based model are usually better for predicting the outcome of any policy i.e., any kind of measures that administration may take to improve the service; or to introduce a new mode in that particular corridor.

(Refer Slide Time: 19:18)

Classification Table^{a,b}

Observed	Predicted						
	Selected Cases ^a			Unselected Cases ^c			
	SPSS_choice	Percentage Correct	SPSS_choice	Percentage Correct	SPSS_choice	Percentage Correct	
Step 0	00	200	0	100.0	424	0	100.0
	1.00	30	0	0	343	0	0
Overall Percentage				89.7			85.3

Classification table of INTERCEPT model. It is shown as BLOCK 0 in SPSS output.

Classification Table^a

Observed	Predicted						
	Selected Cases ^b			Unselected Cases ^c			
	SPSS_choice	Percentage Correct	SPSS_choice	Percentage Correct	SPSS_choice	Percentage Correct	
Step 1	00	260	0	100.0	406	18	95.8
	1.00	28	2	6.7	321	22	6.4
Overall Percentage				90.3			85.8

Model Summary

Step	-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
1	164.410 ^a	.094	.193

^a Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification table of main model. It is shown as BLOCK 1 in SPSS output.
The pseudo R-square (Nagelkerke) was found to be 0.193, which is very low.
The classification table shows not much increase in accuracy.

Ben-akiva et. al. proposed combining RP and SP to get a model that has the practicality of RP and prediction power of SP.
A scale parameter is introduced to account for the different variance of the error term in RP and SP data.

a. The cutvalue is .500
b. Selected cases SP EO 0
c. Unselected cases SP NE 0

NPTEL Online Certification Course
IIT Kharagpur

There are two classification tables shown for the RP based model; one is for the intercept model, and another for the estimated model. The intercept model, which has only the constant term, has an accuracy of 89.7%, whereas the estimated model has an accuracy of 90.3%. So, the change in accuracy is not much profound and so was in the change in the log likelihood ratio. The pseudo R-square is also very less, around 0.193.

In order to overcome such problems, Ben-akiva et.al. proposed combining RP and SP to estimate a model. Such an approach has the capability to exploit the practicality of RP and the prediction power of SP. For these kind of models, scale parameters are introduced to account for the difference in variance of the error term in RP and SP datasets. In other words, due to the different variance in error for both the datasets, they cannot be combined directly. A scale parameter is estimated which could be multiplied with the utility equation of SP based model, and a joint RP-SP model can be developed. A more detailed discussion on this is done in later sections.

(Refer Slide Time: 20:57)

Mode choice model using multinomial logistic regression

Surveys:
 User and household characteristics
 Trip characteristics (Purpose, time of day, frequency of travel, origin(O), and destination(D))
 Mode chosen
 User perception of travel service(sometimes perception data is used).

Mode availability and level of service data is obtained from secondary sources.
Modal data (Cost, Time, Waiting time, Availability) is prepared from available schedules, fare charts, observed volumes, travel time and toll information either link wise or trip OD wise .

- Network analysis is done to determine in-vehicle travel times for transit and personal vehicles.
- Out-of-vehicle time for personal mode is assigned a small value (walk access/egress to/from the car)
- Transit out-of-vehicle time (transit schedules, transfers, and location of bus stops from O-D locations).
- Parking costs are obtained from parking rates at the destination and duration.
- Non-motorized mode (cycle and walk) travel time (zone-to-zone distance and walk/cycling speed)

Travel time and cost between zones based on the origin and destination zones of the trip.

(Source: Koppelman and Bhat, 2006)

NPTEL Online Certification Course
 IIT Kharagpur

Mode choice model using multinomial logistic regression:

SPSS is not used to develop mode choice models with multiple modes because it is difficult to set multiple mode choice utility equations in this software. LIMDEP is a popular software that is used to develop such models. Python programming can also be used to develop such kind of models.

The steps to estimate mode choice model using multinomial logistic regression are discussed in this section. The **first step** is the survey, where data on user and household characteristics; trip characteristics like, purpose, time of day, frequency of travel, origin, and destination; chosen mode; user perception of travel service, are collected.

In the **second step**, mode availability and level of service data is obtained from secondary sources. For mode availability data, the urban area is surveyed to understand which modes are available in which locations. Help from the local authorities also needs to be taken in order to obtain the secondary data like; bus routes, fare charts, toll information, observed volumes, schedules, etc. Link wise or OD pair wise data related to travel cost, travel time, waiting time, and availability are compiled from available schedule, fare charts, observed volume, travel time and toll information. Like the skim tables, matrices are developed for each of these variables. These matrices are then referred for the values of these variables for each observation (individual). These data are collected from secondary sources because different people may perceive the same thing differently. For example, a person might say a certain time taken for travel between one point to another, whereas

another person might say that the travel time between the same points are not same as mentioned by the previous person.

Some of the data obtained are link-wise, while others being available OD pair-wise. For example, cost is not set link-wise, but OD pair-wise; travel times are estimated link-wise, for different times of the day and needs to be added up to obtain the OD pair-wise travel time data; waiting time may be available be as per the origin-destination for each stop and as per mode; and availability of mode varies as per origin or as per the destination. Often such an extensive database is not accessible from secondary sources. In such cases, averages of user specified values for particular links can be taken. For example, travel times reported by ten users traveling from 'i' to 'j', using a particular mode(bus), can be averaged to get the bus travel time between origin 'i' and destination 'j'.

Network analysis is done to determine in vehicle travel times for transit and personal vehicles. Out-of-vehicle time (walk access, egress to/ from the car) for personal mode is assigned small values, so much so that often these are ignored. Transit out-of-vehicle travel time is based on transit schedules, transfers, location of bus stops, and OD location. Parking costs are obtained from parking rates at the destination and duration of parking. For non-motorized mode (bicycle and walk) travel times, instead of recording the travel times perceived by the users, it is calculated by dividing the zone to zone link distance by the average walking and cycling speed. So, this is how the database or matrices are created for using in the mode choice model data set.

(Refer Slide Time: 25:36)

Case	Trip Number	Income	Alternative 1		Alternative 2		Alternative 3		Alternative Chosen
			Time	Cost	Time	Cost	Time	Cost	
	1	30000	30	150	40	100	20	200	1
	2	30000	25	125	35	100	0	0	2
	3	40000	40	125	50	75	30	175	3
	4	50000	15	225	20	150	10	250	3

Case-Alternative	Trip Number	Alternative Number	Number of Alternatives	Income	Time	Cost	Alternative Chosen
1	1	1	3	30000	30	150	1
1	1	2	3	30000	40	100	0
1	1	3	3	30000	20	200	0
2	2	1	2	30000	25	125	0
2	2	2	2	30000	35	100	1
3	3	1	3	40000	40	125	0
3	3	2	3	40000	50	75	0
3	3	3	3	40000	30	175	1
4	4	1	3	50000	15	225	0
4	4	2	3	50000	20	150	0
4	4	3	3	50000	10	250	1

(Source: Koppelman and Bhat, 2006)

In the third step, the database is restructured to reflect the selection of a particular alternative among other alternatives. For example, as shown in the ‘case’ table, four cases or observations are recorded. In the first case, the person has income of 30,000; travel times for alternative 1, 2, and 3 are 30, 40, and 20 respectively; travel cost for alternative 1, 2, and 3 are 150, 100, and 200 respectively; the chosen alternative is ‘alternative 1’. In the second case, the person has income of 30,000; travel times for alternative 1, 2, and 3 are 25, 35, and 0 respectively; travel cost for alternative 1, 2, and 3 are 125, 100, and 0 respectively; the chosen alternative is ‘alternative 2’. In the third case, the person has income of 40,000; travel times for alternative 1, 2, and 3 are 40, 50, and 30 respectively; travel cost for alternative 1, 2, and 3 are 125, 75, and 175 respectively; the chosen alternative is ‘alternative 3’. Similarly, in the fourth case also the person is getting to choose an alternative out of the three available alternatives. If for any person, any particular mode is not available, the values are put as zero, like in case 2.

In order to enable the software to estimate the utilities for each of the alternative, the data needs to be restructured such that each alternative has a separate row. In other words, each case-alternative pair is represented as a separate observation. For example, in the table ‘case-alternative’, the first case has been translated to three separate cases, each for an alternative. The first row in this table is for alternative 1 in which the income of the owner is 30,000; the travel time for this mode is 30; and the travel cost for this mode is 150. In order to represent the selection of mode (or from analytical perspective) among the available modes for a person of a given user profile, a separate column ‘Alternative chosen’ is added that takes only 0 and 1; 0 representing ‘not selected’, and

1 representing ‘selected’ in order to represent the higher utility of an alternative. For alternative 1, the value in column ‘**Alternative chosen**’ is 1. The second row in the same table is for alternative 2 in which the income of the owner is 30,000; the travel time for this mode is 40; and the travel cost for this mode is 100, and ‘alternative chosen’ is 0. The third row in the same table is for alternative 3 in which the income of the owner is 30,000; the travel time for this mode is 20; and the travel cost for this mode is 200, and ‘alternative chosen’ is also 0. Similarly, the other cases are translated into individual case-alternative scenarios. So, basically, for each person, as many alternatives are available, equal number of rows are created with the socio-economic information being constant and the alternative attributes varying based on the given data. This approach is similar to that of location choice model where the chosen alternative was taken along with other randomly selected not-chosen alternatives in order to create the choice-set. The random selection is done in location choice because, unlike mode choice, location choice has many alternatives which cannot be estimated at the same time for a particular observation.

(Refer Slide Time: 28:17)

Example :Work Mode Choice in the San Francisco Bay Area (Source: Koppelman and Bhat, 2006)

$V_{DA} = \beta_1 X TT_{DA} + \beta_2 X TC_{DA}$	DA (drive alone),
$V_{SR2} = \beta_{SR2} + \beta_1 X TT_{SR2} + \beta_2 X TC_{SR2} + Y_{SR2} X Inc$	SR2 (shared ride with 2 people),
$V_{SR3+} = \beta_{SR3+} + \beta_1 X TT_{SR3+} + \beta_2 X TC_{SR3+} + Y_{SR3+} X Inc$	SR3+ (shared ride with 3 or more people),
$V_{TR} = \beta_{TR} + \beta_1 X TT_{TR} + \beta_2 X TC_{TR} + Y_{TR} X Inc$	TR(transit),
$V_{BK} = \beta_{BK} + \beta_1 X TT_{BK} + \beta_2 X TC_{BK} + Y_{BK} X Inc$	BK (bike) and
$V_{WK} = \beta_{WK} + \beta_1 X TT_{WK} + \beta_2 X TC_{WK} + Y_{WK} X Inc$	WK (walk)

Travel time (TT) and travel cost (TC) (generic: same impact on modal utility for modes)
Income (Inc) Alternative-specific variable.
Drive alone (base alternative for household income and the modal constants)

Example: Work mode choice in San Francisco Bay Area:

This example has been taken from the paper by Koppelman and Bhat, which can be referred for further reading. In this example, the work mode choice for San Francisco Bay has been modelled. The various modes in the model are; drive alone (DA); shared ride with 2 people (SR2); shared ride with 3 or more people (SR3); transit mode (TR); bike (BK); and walk (WK). The utility equations for each of these modes are given as follows:

$$V_{DA} = \beta_1 \times TT_{DA} + \beta_2 \times TC_{DA} \qquad \text{DA (Drive alone)}$$

$$V_{SR2} = \beta_{SR2} + \beta_1 \times TT_{SR2} + \beta_2 \times TC_{SR2} + \gamma_{SR2} \times Income$$

SR2 (Shared ride with 2 people)

$$V_{SR3+} = \beta_{SR3+} + \beta_1 \times TT_{SR3+} + \beta_2 \times TC_{SR3+} + \gamma_{SR3+} \times Income$$

SR3+ (Shared ride with 3 or more people)

$$V_{TR} = \beta_{TR} + \beta_1 \times TT_{TR} + \beta_2 \times TC_{TR} + \gamma_{TR} \times Income$$

TR (Transit)

$$V_{BK} = \beta_{BK} + \beta_1 \times TT_{BK} + \beta_2 \times TC_{BK} + \gamma_{BK} \times Income$$

BK (Bike)

$$V_{WK} = \beta_{WK} + \beta_1 \times TT_{WK} + \beta_2 \times TC_{WK} + \gamma_{WK} \times Income$$

WK (Walk)


Among the independent variables, the ones that have been used are travel time, travel cost, and income. DA is reference alternative in the models and hence the utility equation of DA does not have any alternative specific coefficient. β_{SR2} , β_{SR3+} , β_{TR} , β_{BK} , β_{WK} are the alternative specific coefficients in the utility equation of shared ride with 2 people, shared ride with 3 or more people, transit, bike, and walk, respectively. Travel time and travel cost are generic variable, which means their impact on the utility of each alternative will be same, and so these two variables have a common estimated coefficients β_1 and β_2 respectively. Income in this case is an alternative specific variable. It is expected to have a high positive impact on DA and less impact for other modes. Since, DA is the reference alternative, and coefficient of income is not estimated for DA, the impact of income will be slightly negative for shared ride modes, and highly negative for transit and non-motorized modes. γ_{SR2} , γ_{SR3+} , γ_{TR} , γ_{BK} , γ_{WK} are the alternative specific coefficients for income in the utility equation of shared ride with 2 people, shared ride with 3 or more people, transit, bike, and walk, respectively.

(Refer Slide Time: 30:44)

Example :Work Mode Choice in the San Francisco Bay Area (Source: Koppelman and Bhat, 2006)

Variables	Zero Coefficients Model	Constants Only Model	Base Model	
Travel Cost (1990 cents)			-0.0049	(-20.6)
Total Travel Time (minutes)			-0.0513	(-16.6)
Income (1,000's of 1990 dollars)				
Drive Alone (Base)			0	
Shared Ride 2			-0.0022	(-1.4)
Shared Ride 3+			0.0004	(0.1)
Transit			-0.0053	(-2.9)
Bike			-0.0128	(-2.4)
Walk			-0.0097	(-3.2)
Mode Constants				
Drive Alone (Base)		0	0	
Shared Ride 2		-2.137 (-44.1)	-2.178	(-20.8)
Shared Ride 3+		-3.303 (-40.6)	-3.725	(-21.0)
Transit		-1.950 (-38.5)	-0.6709	(-5.1)
Bike		-3.334 (-23.1)	-2.376	(-7.8)
Walk		-2.040 (-23.9)	-0.2068	(-1.1)
Log-likelihood at Zero		-7309.601		-7309.601
Log-likelihood at Constant				4132.916
Log-likelihood at Convergence	-7309.601	-4132.916		-3626.186
Rho-Squared w.r.t. Zero	NA	0.4346		0.5039
Rho-Squared w.r.t. Constants	NA	NA		0.1226

Software used: ALOGIT, LIMDEP and ELM




Many versions of the model were made using many combinations of the variables. The results shown have been estimated using LIMDEP, in the ALOGIT module, and ELM was also used. LIMDEP is statistical software that needs to be used. As an alternative, python can be used to estimate the model which is freely available. In multinomial logistic regression using python, there is scope of defining utility equations for different available alternatives, and availability of alternatives can also be mentioned for each observation.

In the results shown, the first model is a zero coefficient model, which has all the variables and the alternative specific constants as 0. The log likelihood is -7309.601 for this model. The second model is a constant only model, which has the alternative specific constants only. The log likelihood is -4132.916 and R-square is 0.4346 for this model. This significant reduction in log-likelihood value implies that the inclusion of the constants enhanced the prediction power of the model. The third model is the base model, which has all the generic variables, the alternative specific income coefficients, and the alternative specific constants. The log likelihood is -3626.186 and the R-square is 0.5039 for this model. The further decrease in the log-likelihood for the base model signifies even greater prediction power than both the previous models.

(Refer Slide Time: 33:53)

Variables	Model 7W	Model 11W	Constants	
Travel Cost by Income (1990 cents)	-0.004 (-17.2)	-0.004 (-17.4)	0	0
Travel Time (minutes)			Drive Alone (Base)	
Motorized Modes Only	-0.042 (-11.8)	-0.038 (-10.7)	Shared Ride 2	-2.188 (-22.3) -1.594 (-12.1)
Non-Motorized Modes Only	-0.048 (-8.6)	-0.047 (-8.4)	Shared Ride 3+	-3.518 (-28.6) -3.14 (-17.0)
OVT by Distance (mi) Motorized Modes	-0.181 (-10.1)	-0.181 (-9.8)	Transit	-0.042 (-0.3) 0.963 (4.8)
			Bike	-2.687 (-8.1) -1.831 (-4.5)
Income (1,000's of 1990 dollars)			Walk	-1.023 (-3.5) -0.238 (-0.7)
Drive Alone (Base)	0	0	Log-likelihood at Zero	-7309.601 -7309.601
Shared Ride 2	-0.001 (-1.0)	-0.002 (-1.2)	Log-likelihood at Constant	-4132.916 -4132.916
Shared Ride 3+ = Shared Ride 2	-0.001 (-1.0)	-0.002 (-1.2)	Log-likelihood at Convergence	-3547.34 -3489.23
Transit	-0.007 (-3.8)	-0.006 (-3.0)	Rho-Squared w.r.t. Zero	0.5147 0.522
Bike	-0.012 (-2.3)	-0.012 (-2.2)	Rho-Squared w.r.t. Constants	0.1417 0.125
Walk	-0.008 (-2.6)	-0.008 (-2.5)	Adjusted Rho-Squared w.r.t. Zero	0.5122 0.502
			Adjusted Rho-Squared w.r.t. Constants	0.1411 0.154
Autos Ownership		(Autos per worker)	Likelihood ratio Test vs. Model 7W	NA 216.25<0.001
Drive Alone (Base)		0	Adj. LRT vs. Model 10W	NA <0.001
Shared Ride 2		-0.433 (-5.6)		
Shared Ride 3+		-0.267 (-2.4)		
Transit		-0.990 (-8.6)		
Bike		-0.673 (-2.7)		
Walk		-0.628 (-3.9)		

(Source: Koppelman and Bhat, 2006)

Non-motorized mode(physically stressful) users are more sensitive to travel time than motorized modes users.
People are more sensitive to out-of-vehicle travel time (OVT) than to in-vehicle travel time (IVT).

The tables show another two variations of the work mode choice model for the same area. These new models are termed as 7W and 11W. As compared to the base model, the difference in both 7W and 11W is that, instead of using a generic travel time, travel time for motorized modes, travel time for non-motorized modes, and out-of-vehicle travel time for motorized modes has been

considered. Out-of-vehicle travel time does not exist for non-motorized modes as they offer last mile connectivity. The motive behind segregating the travel times is that people travelling by motorized modes feel the passage of time in a very different way than it is felt by people travelling by non-motorized modes; be it comfort, exposure to external environment, etc. Due to these segregations in travel time, the database also needs to be modified. Travel time for motorized modes and for non-motorized modes needs to be made into separate lists. Further, motorized mode travel time needs to be split into in-vehicle travel time and out-of-vehicle travel time.

Travel cost remains as a generic variable in both the 7W and 11W models. Income is an alternative specific variable in 7W and 11W, same as in the base model. A new alternative specific variable, 'auto ownership' has been introduced, with DA as the reference alternative, in the model 11W. This results in the modification of the utility equation of the alternatives.

Comparing the model 7W and 11W, the log likelihood value reduces from -3547.34 in 7W to -3489.236 in 11W, which is not a very significant decrease. The R-square value increases from -0.5147 in 7W to 0.5227 in 11W, which is a slight increase. Travel time in general is a disutility, meaning more is the travel time, less in the probability to choose the mode. Non-motorized mode users are more sensitive to travel time as compared to motorized modes which can be seen from the travel time coefficients in both 7W and 11W. The absolute value of coefficient of travel time for non-motorized mode is greater than that of motorized mode in the both models. This is justified because travelling in non-motorized mode is physically more stressful as compared to travelling in a motorized mode.

People are found to be more sensitive to out of vehicle travel time than to in vehicle travel time. This is also understandable as during the in-vehicle travel time duration, people are comfortably seated, whereas during the out-of-vehicle travel, they are subjected to physical labour and harsh environmental conditions.

(Refer Slide Time: 36:41)

Multinomial Logit and Multinomial Choice using LIMDEP and NLOGIT

$$Prob(Y_{i,mode} = 1) = \frac{\exp(\alpha_{mode} + \beta_{time} \cdot TIME_{i,mode} + \beta_{cost} \cdot COST_{i,mode} + \gamma_{mode} \cdot INCOME_i)}{\sum_{modes} \exp(\alpha_{mode} + \beta_{time} \cdot TIME_{i,mode} + \beta_{cost} \cdot COST_{i,mode} + \gamma_{mode} \cdot INCOME_i)}$$

	MODE	TIME	INVC	INVT	GC	HINC	PSIZE
1	0	69	59	100	70	35	1
2	0	34	31	372	71	35	1
3	0	35	25	417	70	35	1
4	1	0	10	180	30	35	1
5	0	64	58	68	68	30	2
6	0	44	31	354	84	30	2
7	0	53	25	399	85	30	2
8	1	0	11	255	50	30	2
9	0	69	115	125	129	40	1
10	0	34	98	892	195	40	1
11	0	35	53	882	149	40	1
12	1	0	23	720	101	40	1
13	0	64	49	68	59	70	3
14	0	44	26	354	79	70	3
15	0	53	21	399	81	70	3
16	1	0	5	180	32	70	3

CLOGIT command in LIMDEP

CLOGIT ; Choices = list of names for the choices
Lhs = the choice variable
Rhs = attributes that vary across the choices
Rh2 = characteristics that do not vary across choices \$

SAMPLE	1-840 \$
Choices	= air, train, bus, car
Lhs	= mode
Rhs	= invt
Rh2	= hinc \$

(Source: A Quick Start Introduction to NLOGIT 5 and LIMDEP 10)

Multinomial Logit and Multinomial Choice using LIMDEP and NLOGIT:

In LIMDEP the command used to invoke multinomial logit model is **CLOGIT**. The format of the equation used to calculate the choice probability of a particular alternative is as follows:

$$Prob(Y_{i,mode} = 1) = \frac{e^{(\alpha_{mode} + \beta_{time} \cdot TIME_{i,mode} + \beta_{cost} \cdot COST_{i,mode} + \gamma_{mode} \cdot INCOME_i)}}{\sum_{all\ modes} e^{(\alpha_{mode} + \beta_{time} \cdot TIME_{i,mode} + \beta_{cost} \cdot COST_{i,mode} + \gamma_{mode} \cdot INCOME_i)}}$$

There is a user manual for LIMDEP which provides the theory behind the models and the way to use the software. As shown in the table, the data entry has been done in the format already discussed in the beginning of this section. For a given set of alternatives, utility equations are estimated for each of them using the variables in the database. In the given data, the probability of an alternative being chosen can also be referred as the probability of the ‘**mode**’ column value to be 1, for that particular alternative. The mode column is the same as ‘alternative chosen’ column in the ‘case-alternative’ table discussed before. In the mode column, the actual choice is 1, and 0 is for all the other alternatives are not chosen.

In LIMDEP, code needs to be written to instruct the software about the number of samples, as shown in the table. The different choices available needs to be specified; the ‘mode’ variable or the variable that mentions the choices made, needs to be specified corresponding to ‘**Lhs**’; all the independent variables that vary across all the modes (or choices), needs to be specified corresponding to ‘**Rhs**’ like travel time, travel cost, etc. ; and all the independent variables that

remain same for all the modes (or choices), needs to be specified corresponding to ‘Rhs2’ like household income, age, etc.

(Refer Slide Time: 37:58)

Variables that vary across the choices
 TTME = terminal time (waiting time to begin the journey)
 INVC = in-vehicle cost
 INVT = in vehicle time
 GC = a generalized cost measure

Variables that do not vary across choices
 HINC = household income and
 PSIZE = party size
 These are characteristics of the person (traveler).

Discrete choice (multinomial logit) model
 function -249.25650 Estimation based on N = 210, K = 8 Inf. Cr. AIC = 514.5 AIC/N = 2.450 R2 = 1-Log L/ Log L * Log-L fncn R- sqrd R2Adj Constants only -283.7588. 1216.
 1103 Chi-squared [5] = 69.00454 Prob [chi squared > value] = .00000 Response data are given as ind. choices Number of obs. =210. skipped 0 obs

Mode	Coefficient	Standard Error	z	Prob z >*	95% Confidence Interval	
INVT	-.00350***	.00075	-4.69	.0000	-.00496	-.00204
INVC	-.00858	.00626	-1.37	.1707	-.02084	.00369
A AIR	-1.15318	.70809	-1.63	.1034	-2.54101	.23465
AIR_HIN1	.00243	.01045	.23	.8162	-.01806	.02292
A TRAIN	2.07165***	.43004	4.82	.0000	1.22879	2.91451
TRA_HIN2	-.05090***	.01207	-4.22	.0000	-.07456	-.02723
A_BUS	.81928	.50127	1.63	.1022	-.16319	1.80176
BUS_HIN3	-.03268**	.01297	-2.52	.0117	-.05810	-.00727

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

(Source: A Quick Start Introduction to NLOGIT 5 and LIMDEP 10)

The table shown is the output generated by LIMDEP for a multinomial choice model which has the variable names, the coefficients, the z-statistic, and the p-value or the level of significance. Some of the widely used variables that vary across modes are: **terminal time**, which refers to the waiting time; **in-vehicle cost**, which means the fare (in case of transit, shared modes) or fuel cost (in case of personal vehicle); **in-vehicle time**, which means the travel time during which a person is in the vehicle; **generalized cost**, which is a combined cost of other monetary or non-monetary expenses. Among the many variables that remain constant across modes, some are: household income; party size i.e., the size of the group of a cohesive people a person travels with regularly; etc.

The output also shows other model fit statistics like; Akaike Information Criterion (AIC), log-likelihood, R-squared value, and the model chi-square. These measures can be used to know how well the model fits the data and how powerful it is in terms of prediction.

(Refer Slide Time: 39:30)

REFERENCES

- Koppelman, F.S., & Bhat, C.R. (2006). A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models.
- Johnson, Richard A, and Dean W. Wichern. Applied Multivariate Statistical Analysis. Englewood Cliffs, N.J.: Prentice Hall, 1992. <https://www.statisticssolutions.com/the-logistic-regression-analysis-in-spss/>
- Bhui, S & Pandit, D. (2019). Investigating factors responsible for reduction in bus ridership due to fixed route paratransit modes. Conference proceedings of International Conference of Future Cities- 2019.
- A Quick Start Introduction to NLOGIT 5 and LIMDEP 10. Retrieved May 16, 2020, from <http://people.stern.nyu.edu/wgreene/Lugano2013/A%20Quickstart%20Introduction%20to%20NLOGIT%20and%20LIMDEP.pdf>

These are some of the references that can be referred for further reading.

(Refer Slide Time: 39:34)

CONCLUSION

Mode choice model specifications depend on policy under investigation as well as model fit statistics.

A complicated model does not always give better results.

Utility equations should be framed carefully considering both generic and alternative specific variables.

RP data does not offer much variation whereas SP data offers higher variability but may not reflect actual choices.

Several software including Python libraries are available to undertake mode choice modeling.

From the lecture it can be concluded that, mode choice model specifications depend on policies under investigation as well as model fit statistics. A complicated model does not always give better results. Utility equations should be framed carefully considering both generic and alternative specific variables. RP data reflects the real choice making behaviour but does not offer much variation, whereas SP data offers higher variability but may not reflect actual choices. Several software including python libraries are available to undertake mode choice modeling.