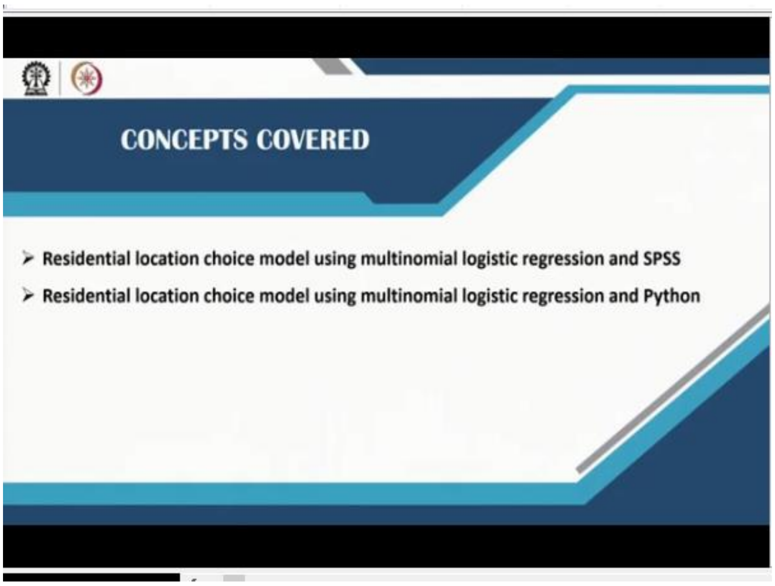
Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology-Kharagpur

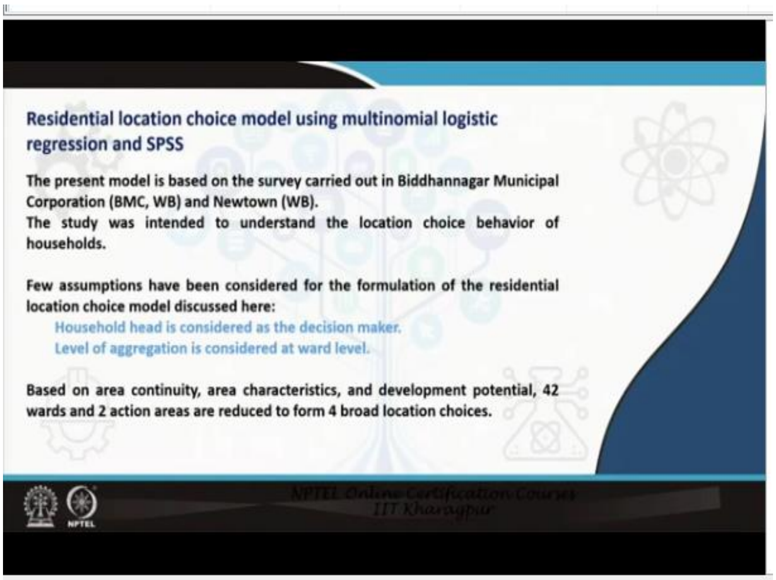
Lecture-30
Residential Location Choice Model Using Multinomial Logistic Regression

(Refer Slide Time: 00:30)



The different concepts covered in this lecture are the residential location choice model using multinomial logistic regression in SPSS, and Python.

(Refer Slide Time: 00:42)

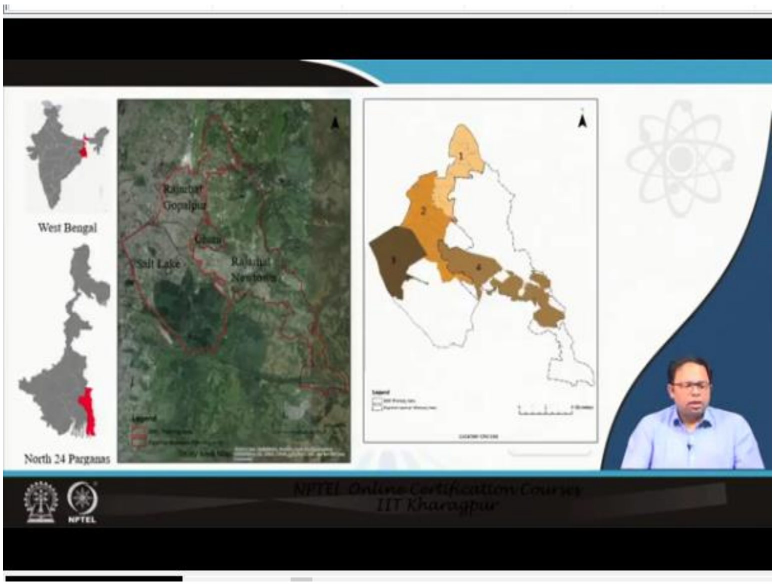


Residential location choice model using multinomial logistic regression and SPSS

The present model is based on a survey carried out in Biddhannagar Municipal Corporation (BMC) and Newtown in West Bengal in 2019. The study was intended to understand the location choice behaviour of households. Initially, few assumptions have been made for the formulation of this particular model. For example, the household head is considered to be the decision-maker and the level of aggregation is considered at the ward level. But the model results were found to be insignificant. Therefore, the location choices were narrowed down to four broad locations.

Earlier, the location choices were 42 wards within BMC and 2 action areas within Newtown. Later, these 44 location choices were narrowed down to 4 broad locations based on factors like area continuity, development potential and area characteristics. In the present example, the model is developed considering these 4 choices only. It is done for a better explanation of how to use SPSS and python for developing a multinomial logit model.

(Refer Slide Time: 02:18)

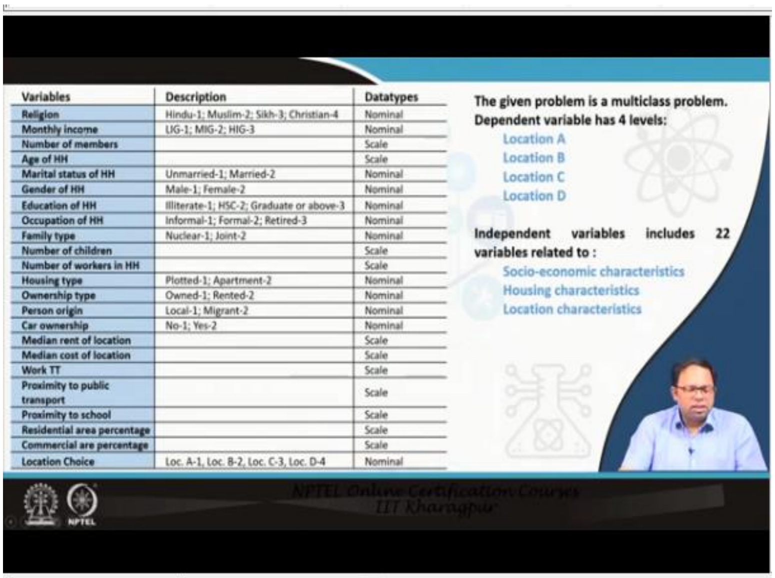


The study area map shows the location of Biddhanagar Municipal Corporation (BMC) and Newtown. Both the areas lie in North 24 Parganas district of West Bengal. BMC constitutes the area of Rajarahat Gopalpur and Salt Lake City whereas, Newtown includes three action areas. Salt Lake City and Newtown are developed as a satellite town for Kolkata city.

The location choice set map represents the 4 broad locations defined for the development of present location choice model. The development pattern in Rajarhat Gopalpur is mostly organic particularly in the wards constituting location 2, whereas wards in location 1 includes a mix of old and new developments. Location 3 covers Salt Lake City, which is a highly developed and planned area. Location 4 includes Newtown's action area 1 and 3. Action area 1 and 3 are newly developing areas but mostly HIG kind of development.

Since, Rajarhat Gopalpur, Salt Lake City, and Newtown shows a different urban character, therefore the wards within these areas are combined to form 4 location categories. Also, the present model considers the households who are located in these areas, and predict in which broad location category households are going to settle in.

(Refer Slide Time: 04:08)



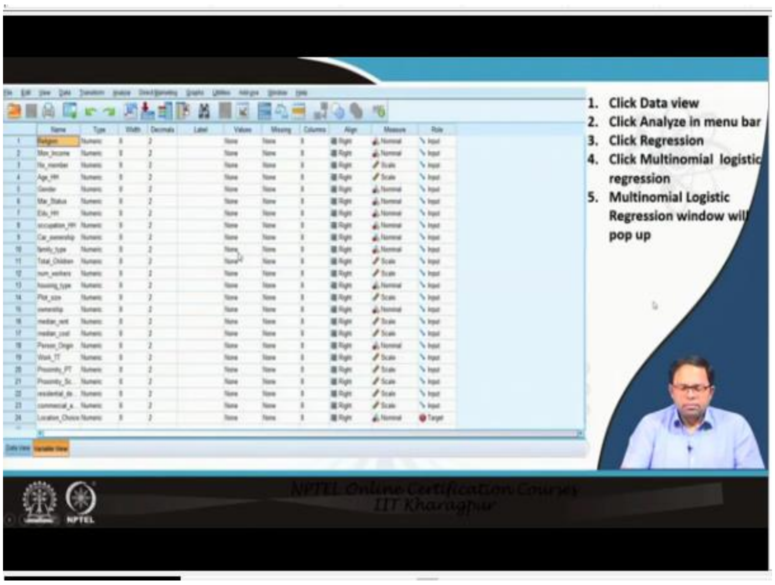
The regression models like logistic regression or linear regression have dependent and independent variables as discussed in the previous chapter. The present problem is a multiclass problem. It means that the dependent variable has more than two categories, which are four location (location 1, 2, 3, and 4 or Location 'A', 'B', 'C', and 'D'). The independent variables include household characteristics, location characteristics, and dwelling characteristics.

Household characteristics include religion, monthly income, age, education, marital status, gender, and occupation of household head, family type, number of children, car ownership, person origin, and number of workers. Dwelling characteristics include housing type, and ownership type. Location characteristics include median rent and cost of housing, work travel time, proximity to public transport, schools, percentage of residential area, and commercial area.

The independent variables are either categorical (or nominal), or continuous (scale). For example, monthly income is a categorical variable with three categories which are LIG, MIG, and HIG. The number of workers is a continuous variable, with values ranging from 0 to 4.

The present model predicts the probability of choosing location A, B, C, or D. This prediction is based on the independent or explanatory variables. For example, does the car ownership influence the choice of location A or B? So, there are 22 independent variables which are tested, and a dependent variable with 4 categories.

(Refer Slide Time: 05:54)



Model development

This section explains the model development process in SPSS software. The process of installation of SPSS, data pre-processing, and data formatting have been discussed in the previous lecture (refer to lecture 28).

In the present study, the dependent variable has four categories i.e. location ‘A’,’ B’,’ C’, and ’D’. When the dependent variable has more than two levels/categories, the multinomial logit model is used. Therefore, the current dataset will be analysed using a multinomial logit model for the development of the location choice model.

Data input

At first, the dataset file needs to be imported. To import the dataset file, consider the following path *File>>Open>>Data*, and then select the 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_HH), 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). These attributes can be also modified as per the requirement.

Model development and specifications

The different steps for the development of multinomial logistic regression are similar to binary logit model discussed in lecture 28. The steps are given below:

Step 1:

Click on *Analyse>>Regression>>Multinomial logistic*, on the menu ribbon. A *Multinomial Logistic regression* window will appear as shown in the video.

Step 2:

Drag and drop the dependent variable (location choice) to the *Dependent* box. Then, click on the *Reference Category* tab to define the reference category of the dependent variable. In SPSS, the default reference category is the last category of the variable. For the present model, location 4 is the reference category.

Step 3:

Drag and drop the categorical independent variables (Gender, Religion, Mon_Income) to the *Factor(s)* box, and the continuous variable (Age_HH, Work_TT) to the *Covariate(s)* box.

Step 4:

Click on the *Statistics* tab to define the model specifications. A *Multinomial Logistic Regression: Statistics* window gets open. Here, check the cell probabilities, classification table, goodness of fit etc.

Step 5:

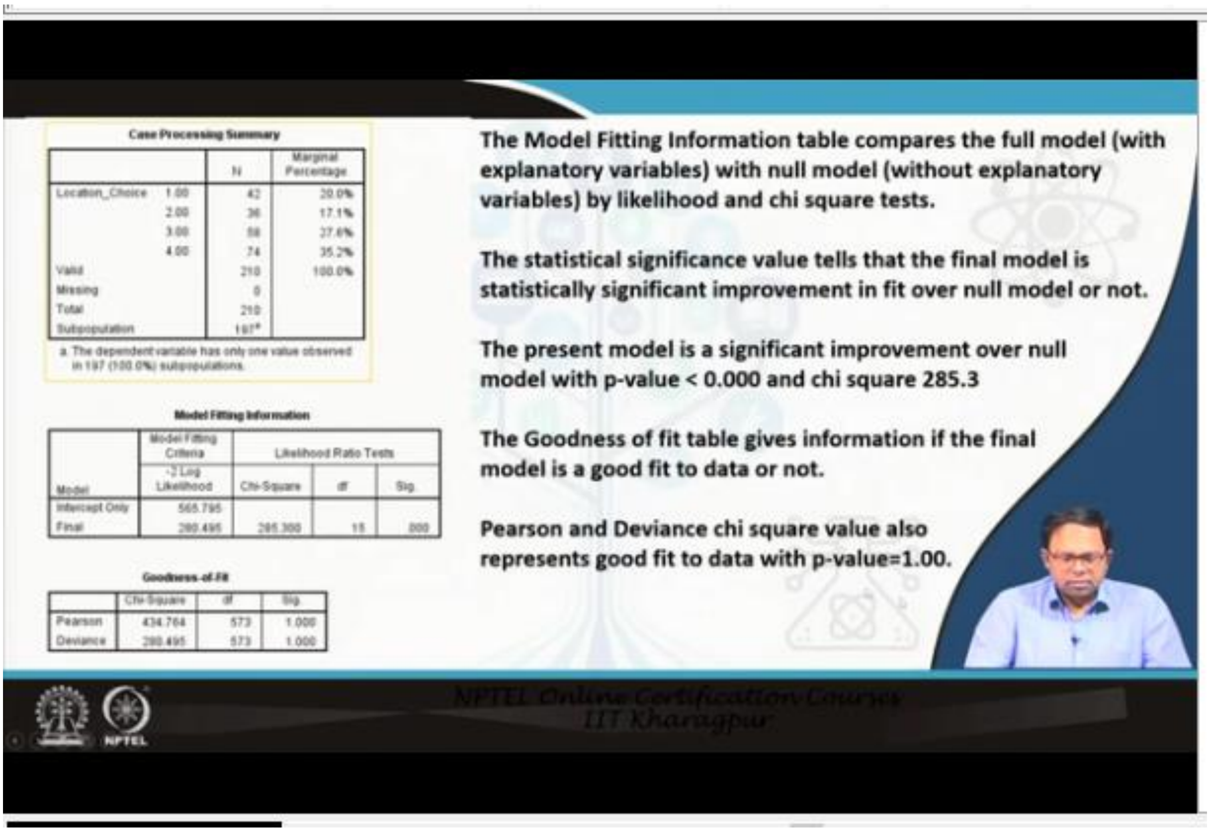
Click on the *continue* tab. You will be returned to the *Multinomial Logistic regression* window.

Step 6:

Click the *Ok* tab to run the model.

After following the aforementioned steps, the output for the multinomial logit model will be generated.

(Refer Slide Time: 08:16)



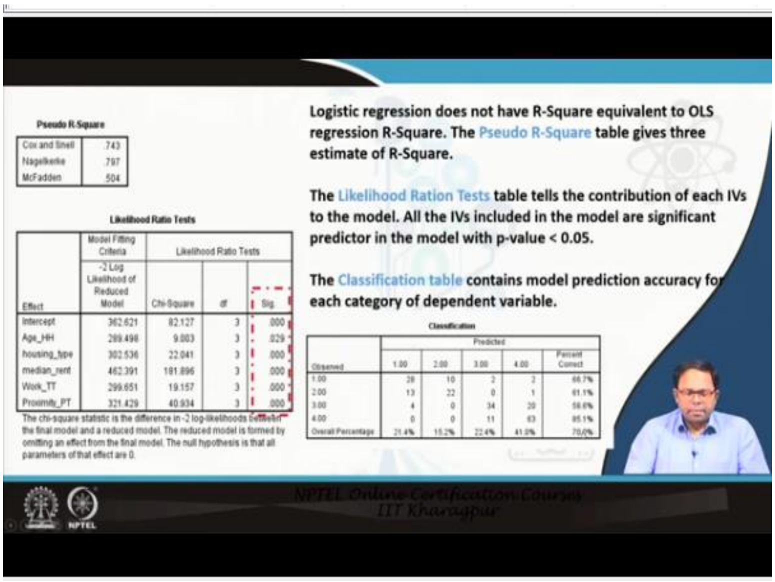
Model result

SPSS produces many tables of output for multinomial logistic regression. These tables are important to understand and interpret the results of the regression. The first table *Case Processing Summary* shows how many cases are included in the analysis for each category of the dependent variable, how many cases are missing etc.

The second table is *Model Fitting Information.* This table compares the full model (with explanatory variable) with the null model (without explanatory variable) using likelihood and chi-square test. The value of -2log likelihood of the full model should be lower than the value of intercept only (or null model). The present model is a significant improvement over the null model with p-value <0.000, chi-square 285.3 and lower log-likelihood value.

The third table is *Goodness of fit*, which gives information if the final model is a good fit to the data or not. The good fit of the model is represented by higher p-value i.e. >0.05 for both Pearson and Deviance statistics. The present model is a good fit to the data with p-value=1.00.

(Refer Slide Time: 09:31)

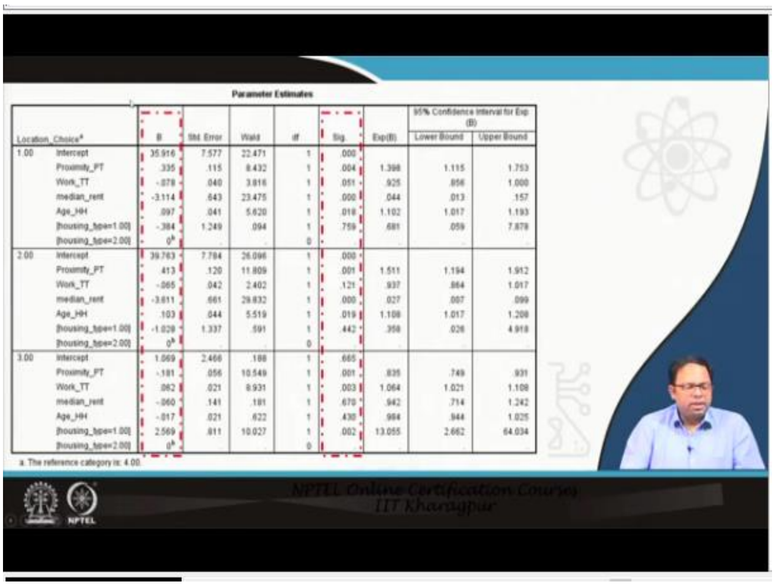


The next table is *Pseudo R-Square*, which gives three estimate of R-Square. The logistic regression does not have R-square equivalent to OLS regression R-Square. So, it only gives an idea of how good the model is.

The *Likelihood Ratio* Tests table tells the contribution of each independent variable to the model. All the independent variables included in the model are significant predictor in the full model with p-value<0.05.

The *Classification* table contains model prediction accuracy for each category of the dependent variable. In other words, it tells how well the model classifies the cases for each category of the dependent variable. The present model correctly classifies 66.7% cases for location 1, 61.1% cases for location 2, 58.6% cases for location 3, and 85.1% cases for location 4. The overall percentage is the most important part of this table, which is 70% for the present model.

(Refer Slide Time: 11:12)



Parameter Estimates

Location_Choice[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
1.00	Intercept	35.916	7.577	22.471	1	.000			
	Proximity_PT	.335	.115	8.432	1	.004	1.398	1.115	1.753
	Work_TT	-.078	.040	3.816	1	.051	.925	.856	1.000
	median_rent	-3.114	.643	23.475	1	.000	.044	.013	.157
	Age_HH	.097	.041	5.620	1	.018	1.102	1.017	1.193
	[housing_type=1.00]	-.384	1.249	.094	1	.759	.681	.059	7.878
	[housing_type=2.00]	0[b]	.	.	0	.	.	.	.
2.00	Intercept	39.763	7.784	26.096	1	.000			
	Proximity_PT	.413	.120	11.809	1	.001	1.511	1.194	1.912
	Work_TT	-.065	.042	2.402	1	.121	.937	.864	1.017
	median_rent	-3.611	.661	29.832	1	.000	.027	.007	.099
	Age_HH	.103	.044	5.519	1	.019	1.108	1.017	1.208
	[housing_type=1.00]	-1.028	1.337	.591	1	.442	.358	.026	4.918
	[housing_type=2.00]	0[b]	.	.	0	.	.	.	.
3.00	Intercept	1.069	2.466	.188	1	.665			
	Proximity_PT	-.181	.056	10.549	1	.001	.835	.749	.931
	Work_TT	.062	.021	8.931	1	.003	1.064	1.021	1.108
	median_rent	-.060	.141	.181	1	.670	.942	.714	1.242
	Age_HH	-.017	.021	.622	1	.430	.984	.944	1.025
	[housing_type=1.00]	2.569	.811	10.027	1	.002	13.055	2.662	64.034
	[housing_type=2.00]	0[b]	.	.	0	.	.	.	.

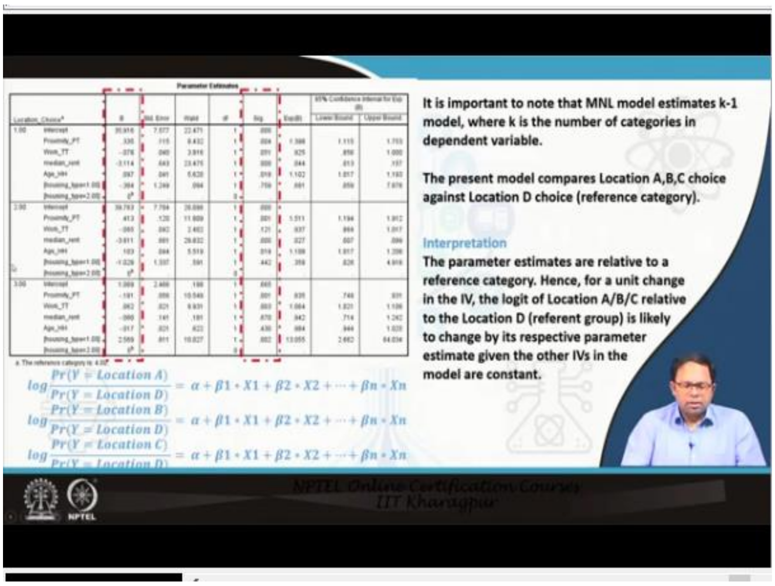
a. The reference category is: 4.00.

The most important part of the model result is *Parameter Estimates* table. This table explains the influence of each independent variable on the probability of choosing a location. Initially, all the independent variables are included in the model, but many variable were found to be insignificant. After many iterations, the significant variables are identified. So, the present *Parameter Estimates* table shows the parameter estimate for significant variables only. The significant variables are housing type, age of household head, median rent of the location, work travel time, and proximity to public transport.

The table gives regression coefficient estimates (B or β), significance value (sig.), and odds ratio (Exp(B)) for each category of variables. It also presents parameter estimates and other values in three sets. The first set of values are for location 1 with reference to location 4. Similarly, the second set and third set of values are for location 2, and 3 (with reference to location 4) respectively.

The regression coefficient (B or β) signifies the value by which the dependent variable will change if there is a unit change in the independent variable. Also, it signifies the effect (+ or -) of the independent variable on the dependent variable. The odds ratio signifies the influence of the independent variable on the likelihood of occurrence of the outcome.

(Refer Slide Time: 12:31)



Result interpretation

The MNL model estimates k - 1 model where k is the number of categories in the dependent variable. As discussed earlier, the present model compares location A, B, and C against location D which is the reference category. The parameter estimates are relative to a reference category. Hence, for a unit change in the independent variable, the logit of location A/B/C relative to the location of D is likely to change by its respective parameter estimate given other independent variables in the model are constant. Based on the values of parameter estimates, the logit of location can be mathematically expressed as:

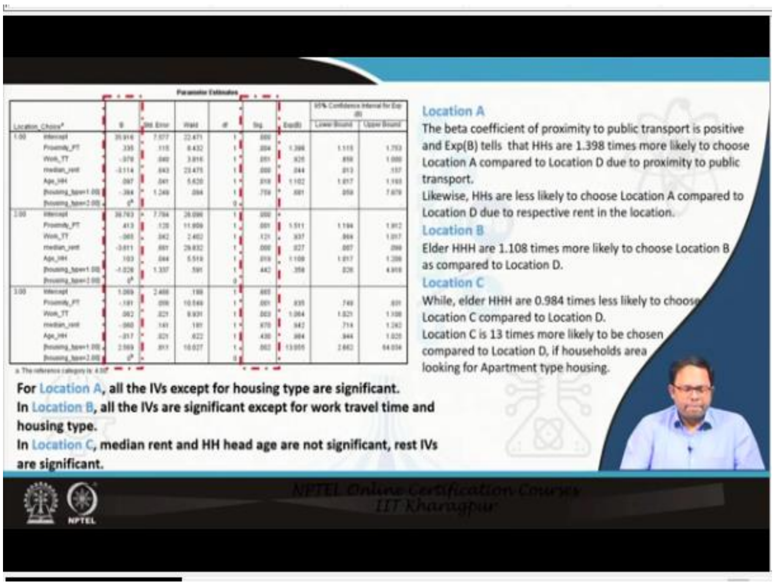
$$\log\frac{\Pr(Y=\text{Location A})}{\Pr(Y=\text{Location D})} = 35.916 + 0.335 * \text{proximity to public transport} - 0.078 * \text{work travel time} - 3.114 * \text{median rent} + 0.97 * \text{age of household head}$$

$$\log\frac{\Pr(Y=\text{Location B})}{\Pr(Y=\text{Location D})} = 39.76 + 0.413 * \text{proximity to public transport} - 3.611 * \text{median rent} + 0.103 * \text{age of household head}$$

$$\log\frac{\Pr(Y=\text{Location C})}{\Pr(Y=\text{Location D})} = 1.069 - 0.181 * \text{proximity to public transport} + 0.062 * \text{work travel time} + 2.569 * \text{housing type 1}$$

These equations includes only the significant variables for each locations.

(Refer Slide Time: 14:24)

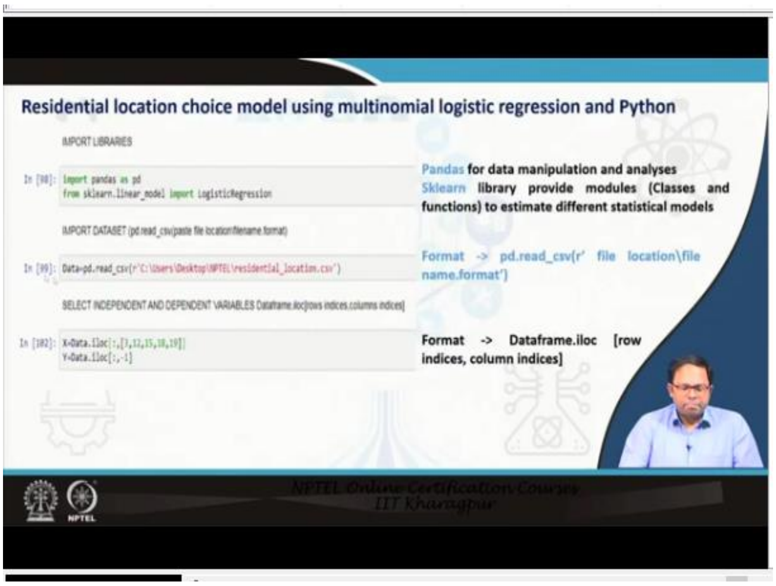


For location ‘A’, the significance value is less than 0.05 for proximity to public transport, work travel time, median rent, and age of household head. So, these variables are significant for location ‘A’. Whereas p-value is greater than 0.05 for housing type, therefore it is an insignificant variable. The beta coefficient and odds ratio for proximity to public transport are 0.335 and 1.398 respectively. It means that the households are 1.398 times more likely to choose location ‘A’ compared to location ‘D’ due to proximity to public transport. Also, the beta coefficient and odds ratio for median rent are -3.114 and 0.044 respectively. So, the households are 0.044 times less likely to choose location ‘A’ compared to location ‘D’ due to median rent in location ‘A’.

Similarly, for location ‘B’, independent variable such as public transport, median rent, and age of household head are significant. Housing type and work travel time are insignificant. The beta coefficient of household head is positive, and the odds ratio is 1.108. So, the elder household heads are 1.108 times more likely to choose location ‘B’ compared to location ‘D’. The beta coefficient and odds ratio for median rent are -3.611 and 0.027 respectively. It means that the households are 0.027 times less likely to choose location ‘B’ as compare to location ‘D’ due to median housing rent in location ‘B. Likewise, other significant variables can be interpreted for location ‘A’, ‘B’, and ‘C’.

This is how the MNL model is interpreted. It is important to mention that if the choices are more, for example, 1000 choices, then 999 different models are to be developed in SPSS. It generally gives insignificant results or the model becomes unstable. Also, the random selection of a subset is not possible in SPSS. So, in such cases, other analysis software's are used.

(Refer Slide Time: 14:27)



Residential location choice model using multinomial logistic regression and Python

The concept of multinomial logistic regression can be implemented in python programming as well. Python is an open source programming language and can be used for statistical analysis. It has inbuilt libraries for statistical models. These libraries are imported for analysis and modelling.

In the present example, residential location choice behaviour of household in BMC area is being modelled. The general steps to develop a multinomial logit model in python are similar to binary logit model discussed in intention to move model. These steps are as follows:

Step 1: Import libraries such as pandas, statsmodels or sklearn.
Step 2: Import dataset, and define independent and dependent variable.
Step 3: Create a model.

Data input

At the beginning, pandas library with alias 'pd' is imported. The pandas library is for reading the dataset. In addition to it, the logistic regression function is imported from sklearn library. Sklearn library is a machine learning library. The code for importing libraries is:

import pandas as pd

from sklearn.linear_model import LogisticRegression

The second step is to import the dataset. There are several ways to import the datasets. The present example shows how to read a csv file into a pandas data frame. The file extension for the present dataset is .csv. The format to read a csv file into pandas dataframe is as follows:

Data = pd.read_csv(r' file location\file name.format')

Data = pd.read_csv(r'C:\Users\Desktop\NPTEL\residential_location.csv')

The next step is to select the independent and dependent variables. In pandas data frame 'Data', the column represents a single variable/attributes (person origin, monthly income), and the row represents values of these variables/attributes. The format to select columns and rows in a pandas data frame is as follows:

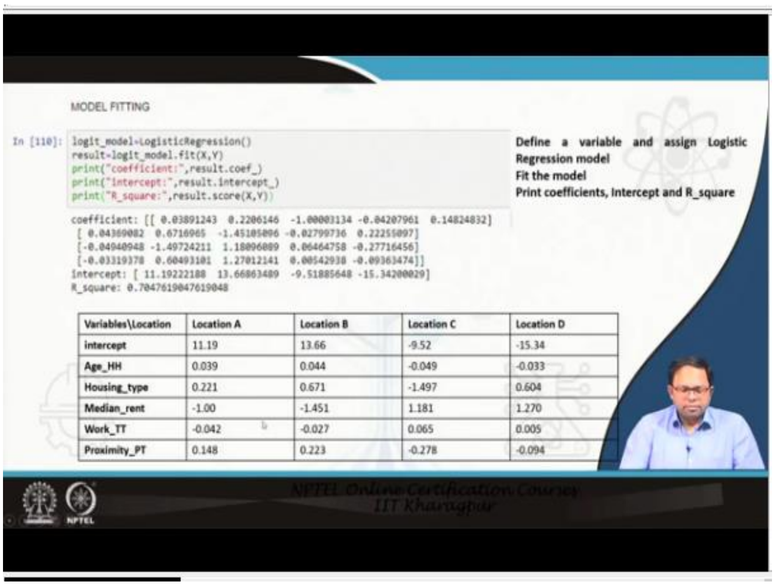
Dataframe.iloc[row indices, column indices]

So, the independent variables (person origin, monthly income, marital status, and others) are assigned to variable 'X' and the dependent variable (location choice) is assigned to variable 'Y'.

X=Dtata.iloc[:,[3,12,15,18,19]].values

Y=Data.iloc[:,-1].values

(Refer Slide Time: 15:25)



Model fitting

The final step is to create the model. The logistic regression function is already imported from sklearn library. This function takes the dependent variable (Y) and independent variable (X) while fitting the model. So, a variable *logit_model* is defined and *LogisticRegression* function is assigned to it. Then, the model is fitted to the data. Finally, coefficients, intercept, and R-square is printed. One can also print other model statistics. The code for calling the function, fitting the function, and printing the coefficients, intercept and R-square value is given below.

```
logit_model=LogisticRegression ( )
result=logit_model.fit(X,Y)
print (“coefficients:”,result.coef_)
print(“intercept:”,result.intercept_)
print(“R-square:”,result.score(X,Y))
```

Model result and interpretation

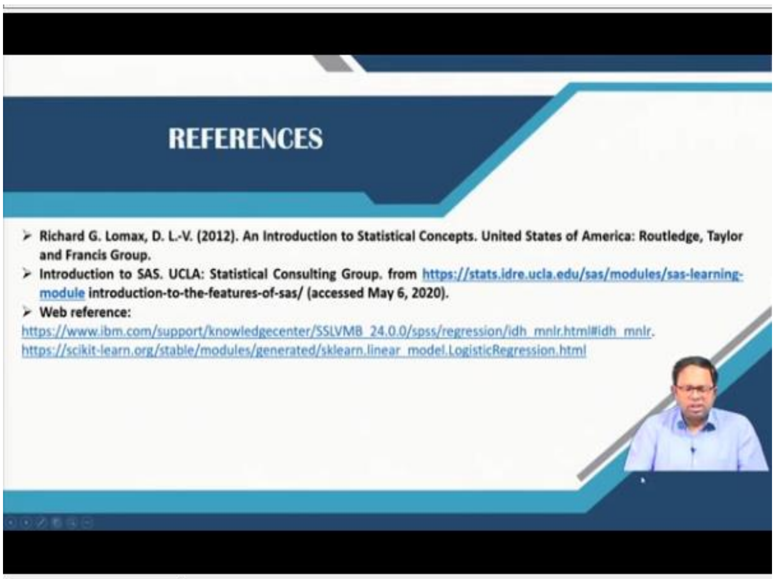
The output shows values for coefficients, intercept, and R-square. The coefficient includes 4 sets of values. Each set represents coefficient values of 5 independent variable for each location. The intercept and coefficients value with the variable name for each location is shown in the table below:

Variables\Locations	**Location A**	**Location B**	**Location C**	**Location D**
Intercept	11.19	13.66	-9.52	-15.34
Age of household head	0.039	0.044	-0.049	-0.033
Housing type	0.221	0.671	-1.497	0.604
Median rent	-1.00	-1.451	1.181	1.270
Work travel time	-0.042	-0.027	0.065	0.005
Proximity to public infrastructure	0.148	0.223	-0.278	-0.094

Based on the above values, the utility equation can be formulated for each location. Also, the probability of choice of a particular location can be calculated. The variable such as age of household head, housing type, and proximity to public transport are positively related to the choice of location ‘A’. Whereas, median rent and work travel time are negatively related. It means that household are less likely to move to location ‘A’ if the median rent and work travel time increases, and more likely to choose location ‘A’ if proximity to public transport and age of household hear increase, or the household type is plotted. Similarly, the variables can be interpreted for location 'B', 'C', and 'D'.

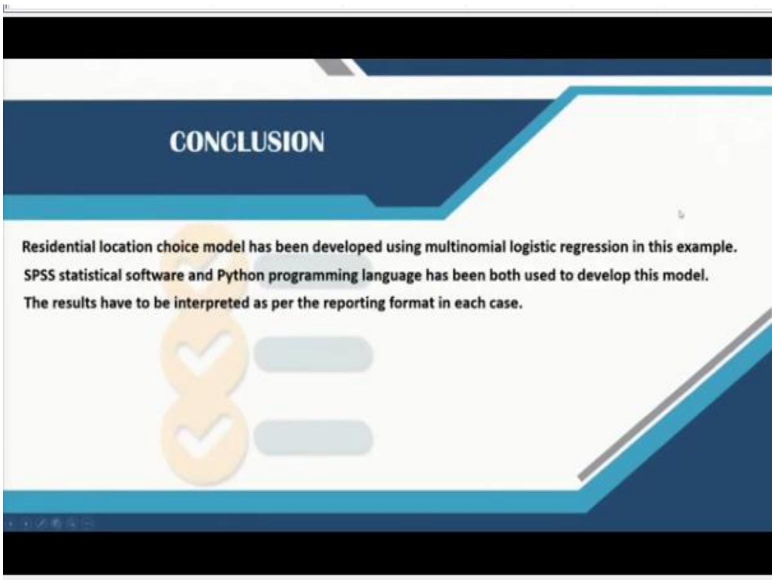
So, this is how the multinomial logistic regression can be implemented in Python, and the model results can be interpreted.

(Refer Slide Time: 17:23)



A list of references is given in the slide.

(Refer Slide Time: 17:27)



Conclusion

Residential location choice model has been developed using multinomial logistic regression in this example. SPSS statistical software and Python programming language has been used to develop this model. Also, the result is interpreted as per the reporting format in each case.