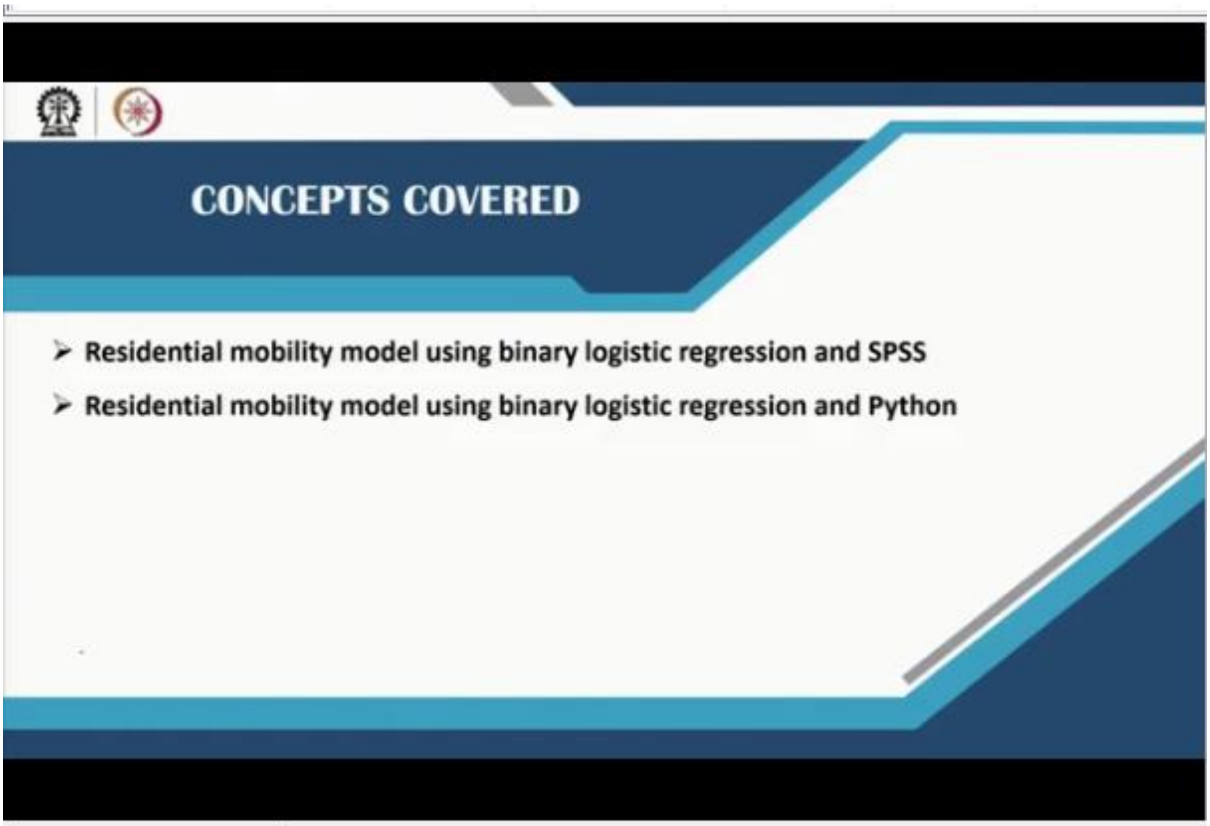
Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology-Kharagpur

Lecture-28
Residential Mobility Model Using Binary Logistic Regression

(Refer Slide Time: 00:33)

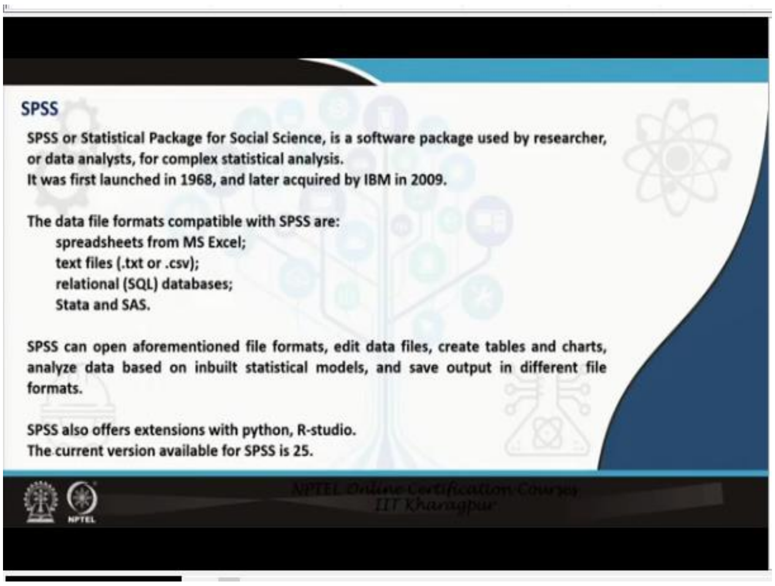


CONCEPTS COVERED

- Residential mobility model using binary logistic regression and SPSS
- Residential mobility model using binary logistic regression and Python

This lecture explains the residential mobility model using binary logistic regression in SPSS software and through Python programming.

(Refer Slide Time: 00:47)



SPSS

SPSS or Statistical Package for Social Science, is a software package used by researcher, or data analysts, for complex statistical analysis.
It was first launched in 1968, and later acquired by IBM in 2009.

The data file formats compatible with SPSS are:

- spreadsheets from MS Excel;
- text files (.txt or .csv);
- relational (SQL) databases;
- Stata and SAS.

SPSS can open aforementioned file formats, edit data files, create tables and charts, analyze data based on inbuilt statistical models, and save output in different file formats.

SPSS also offers extensions with python, R-studio.
The current version available for SPSS is 25.

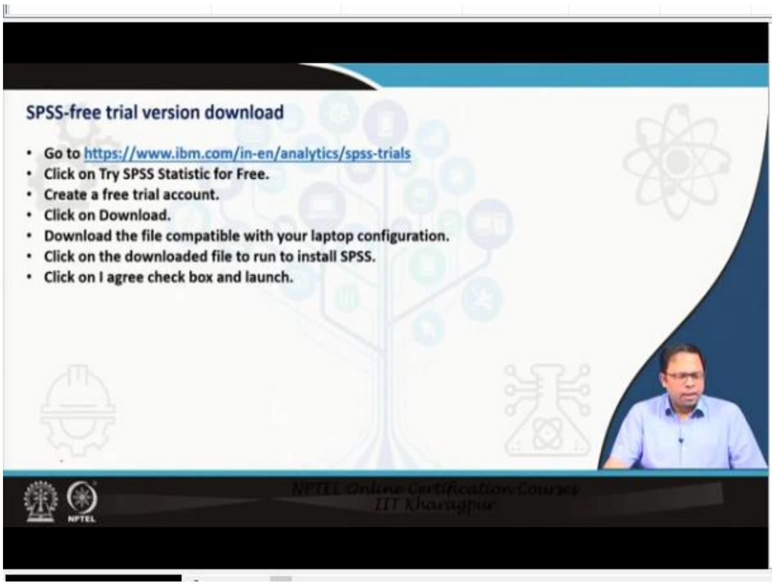
SPSS

SPSS is a statistical package for social science. It was developed in 1968 and then later acquired by IBM in 2009. It is a very popular software that is used in many institutions, and by data analysts for complex statistical analysis. The data file formats compatible with SPSS are spreadsheets, text files, relational databases, or STATA and SAS formats. Stata and SAA are other software for statistical analysis.

In SPSS, one can edit the data files, create tables, charts, and analyze the data based on different statistical models that are already inbuilt into that. The data and SPSS output can be saved in different file formats. SPSS also offers extensions with Python and R-studio, which are open source programming languages.

SPSS is a proprietary statistical analysis software. So, it is not free of cost. A free trial version of SPSS can also be accessed and downloaded using the web link given in the subsequent section. For this course, steps for downloading the free trial version is given to help students who have no access to the SPSS software package.

(Refer Slide Time: 03:13)



SPSS-free trial version download

- Go to https://www.ibm.com/in-en/analytics/spss-trials
- Click on Try SPSS Statistic for Free.
- Create a free trial account.
- Click on Download.
- Download the file compatible with your laptop configuration.
- Click on the downloaded file to run to install SPSS.
- Click on I agree check box and launch.

NPTEL Online Certification Courses
IIT Kharagpur

The free trial version can be accessed using the web link **https://www.ibm.com/in-en/analytics/spss-trials**. The steps to download the free trial version are as follows:

Step 1: Click on Try SPSS Statistics for free.

Step 2: Create a free trial account.

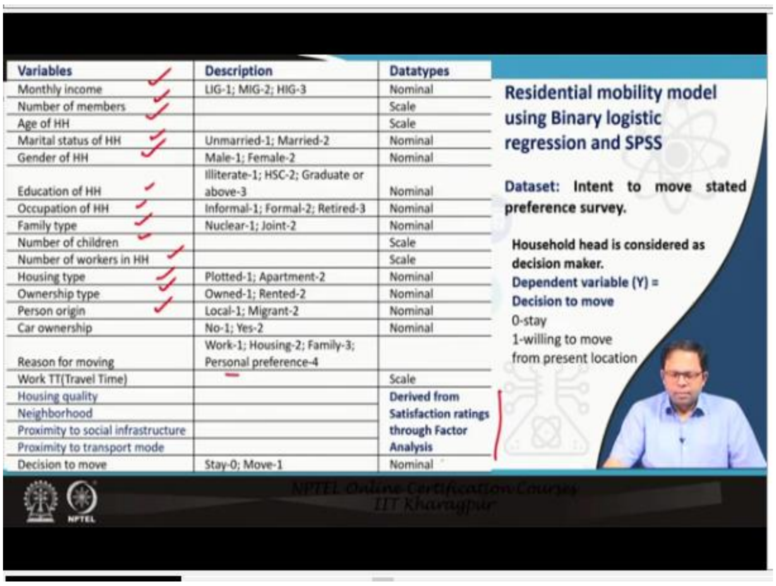
Step 3: Click on download.

Step 4: Download the file compatible with your laptop configuration.

Step 5: Click on the downloaded file to run to install SPSS

Step6: Click on the I agree check box and launch.

(Refer Slide Time: 03:44)



Variables	Description	Datatypes
Monthly income ✓	LIG-1; MIG-2; HIG-3	Nominal
Number of members ✓		Scale
Age of HH ✓		Scale
Marital status of HH ✓	Unmarried-1; Married-2	Nominal
Gender of HH ✓	Male-1; Female-2	Nominal
Education of HH ✓	Illiterate-1; HSC-2; Graduate or above-3	Nominal
Occupation of HH ✓	Informal-1; Formal-2; Retired-3	Nominal
Family type ✓	Nuclear-1; Joint-2	Nominal
Number of children ✓		Scale
Number of workers in HH ✓		Scale
Housing type ✓	Plotted-1; Apartment-2	Nominal
Ownership type ✓	Owned-1; Rented-2	Nominal
Person origin ✓	Local-1; Migrant-2	Nominal
Car ownership	No-1; Yes-2	Nominal
Reason for moving	Work-1; Housing-2; Family-3; Personal preference-4	
Work TT(Travel Time)		Scale
Housing quality		Derived from Satisfaction ratings through Factor Analysis
Neighborhood		
Proximity to social infrastructure		
Proximity to transport mode		
Decision to move	Stay-0; Move-1	Nominal

Residential mobility model using Binary logistic regression and SPSS

Dataset: Intent to move stated preference survey.

Household head is considered as decision maker.
Dependent variable (Y) = Decision to move
0-stay
1-willing to move from present location

NPTEL Online Certification Courses
IIT Kharagpur

Residential mobility model using binary logistic regression and SPSS

To explain the residential mobility model using SPSS, an example of the intention to move model is considered. The dataset used to develop this model was collected through a stated preference survey conducted in Bidhannagar Municipal Corporation (W.B.) in 2019. For the study, the household head is considered as the decision maker. So, the feedback was taken from the household head. Generally, it is assumed that the feedback given by the household head represents the feedback of the whole family.

In regression modelling like linear regression, or logistic regression, there are two types of variables. The first variable is the dependent variable which is being tested or measured. The second variable is independent variables which have a direct impact on the dependent variable. There is a cause and effect relationship between the dependent variable and independent variables.

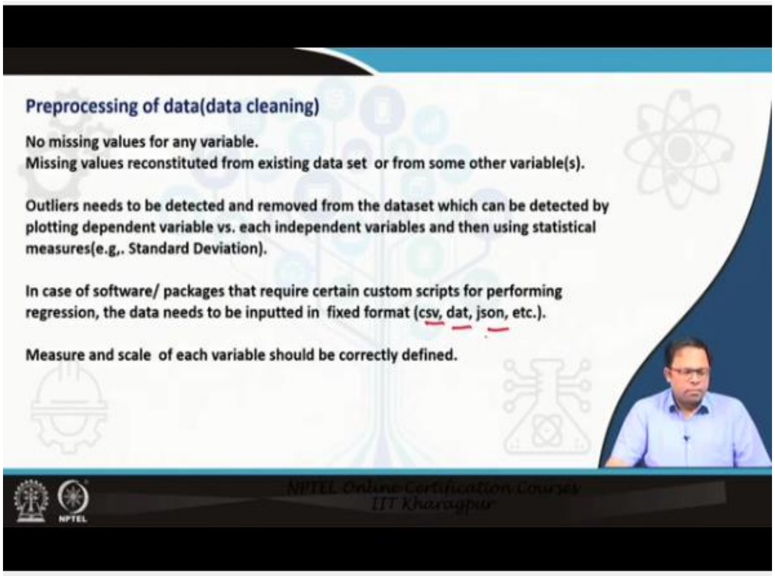
The alternatives/choices in the present model are the decision to move or the decision to stay. So, this will be the dependent variable (Y). The value of the dependent variable is 0 if the household will stay in that particular area, and 1 if a household is willing to move from the present location. The independent variables are household characteristics, dwelling characteristics, and location characteristics. The change in any of these independent variables may affect the decision to move or stay.

Household characteristics included in the model are monthly income, number of members, age of the household head, marital status of the household, education of the household head, the occupation of the household head, family type, number of children in the family, number of workers in the household, and person origin (migrant or local). Dwelling characteristics are housing type (plotted or apartment), and ownership type. In addition to the above characteristics, variables like car ownership and reason for moving are also included in the model.

Locational characteristics include satisfaction with housing quality, neighbourhood quality, proximity to social infrastructure, and proximity to transport mode. The locational characteristics are basically factors derived from satisfaction variables using factor analysis. Many household decisions are based on perceptions, like satisfaction with housing price, quality of structure. These perceptions/satisfaction values are generally measured on a Likert scale. Also, there may exist some correlation between these variables. If these variables are used in the model then due to correlation the model may generate wrong results. To avoid this, factor analysis is carried out to narrow down the perception variables into uncorrelated factors.

All these variables are either categorical (or nominal) or continuous (or scale). For example, gender is a categorical variable with two categories 'male and 'female'. Travel time to work is a continuous variable with values ranging from 5 min to 60 min.

(Refer Slide Time: 08:15)



Preprocessing of data(data cleaning)

No missing values for any variable.
Missing values reconstituted from existing data set or from some other variable(s).

Outliers needs to be detected and removed from the dataset which can be detected by plotting dependent variable vs. each independent variables and then using statistical measures(e.g,. Standard Deviation).

In case of software/ packages that require certain custom scripts for performing regression, the data needs to be inputted in fixed format (csv, dat, json, etc.).

Measure and scale of each variable should be correctly defined.

NPTEL Online Certification Courses
IIT Kharagpur

Pre-processing of data

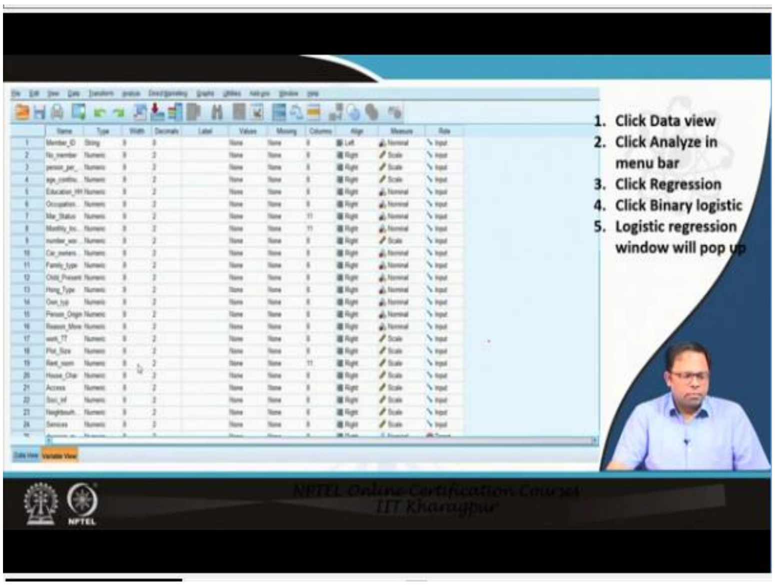
Pre-processing of data means preparing the raw data into a useful and efficient format. It is also known as data cleaning. So, before starting the development of a model, the raw data needs to be cleaned.

In this stage, a check of missing values and outliers is done. The data should not have any missing values for any variable. If missing values are there, then missing values are reconstituted from the existing data set or some other variables. For example, replacing a missing value of a particular variable with the average value of that variable.

Outliers are unusual values in the dataset. These values generally skew the model results. So, it should be detected and removed from the data set. Outlier values can be detected by plotting dependent variables versus independent variables and then using statistical measures like standard deviation.

In the case of software packages that require certain custom scripts for performing regression, the data needs to be imputed in a fixed format. It means that the format of data files needs to be created based on the input requirement of a particular software. For example, csv, dat, json. Also, the measure and scale of each variable should be correctly defined.

(Refer Slide Time: 10:02)



File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Member_ID	String	8	0		None	None	8	Left	Nominal	Input
2	No_member	Numeric	8	2		None	None	8	Right	Scale	Input
3	person_per_...	Numeric	8	2		None	None	8	Right	Scale	Input
4	age_contribu...	Numeric	8	2		None	None	8	Right	Scale	Input
5	Education_HH	Numeric	8	2		None	None	8	Right	Nominal	Input
6	Occupation...	Numeric	8	2		None	None	8	Right	Nominal	Input
7	Mar_Status	Numeric	8	2		None	None	11	Right	Nominal	Input
8	Monthly_Inc...	Numeric	8	2		None	None	11	Right	Nominal	Input
9	number_wor...	Numeric	8	2		None	None	8	Right	Scale	Input
10	Car_owners...	Numeric	8	2		None	None	8	Right	Nominal	Input
11	Family_type	Numeric	8	2		None	None	8	Right	Nominal	Input
12	Child_Present	Numeric	8	2		None	None	8	Right	Nominal	Input
13	Hsng_Type	Numeric	8	2		None	None	8	Right	Nominal	Input
14	Own_typ	Numeric	8	2		None	None	8	Right	Nominal	Input
15	Person_Origin	Numeric	8	2		None	None	8	Right	Nominal	Input
16	Reason_Move	Numeric	8	2		None	None	8	Right	Nominal	Input
17	work_TT	Numeric	8	2		None	None	8	Right	Scale	Input
18	Plot_Size	Numeric	8	2		None	None	8	Right	Scale	Input
19	Rent_room	Numeric	8	2		None	None	11	Right	Scale	Input
20	House_Char	Numeric	8	2		None	None	8	Right	Scale	Input
21	Access	Numeric	8	2		None	None	8	Right	Scale	Input
22	Soci_inf	Numeric	8	2		None	None	8	Right	Scale	Input
23	Neighbourh...	Numeric	8	2		None	None	8	Right	Scale	Input
24	Services	Numeric	8	2		None	None	8	Right	Scale	Input

1. Click Data view
2. Click Analyze in menu bar
3. Click Regression
4. Click Binary logistic
5. Logistic regression window will pop up

Model development

This section describes the model development process in SPSS. In the present study, the dependent variable is dichotomous or binary. When the dependent variable has two levels/categories, the binary logit model is used. Therefore, the current dataset will be analysed using a binary logit model for the development of the intention to move model.

Data Input

There are many ways to input data in SPSS. If the dataset is in SPSS file format (.sav), one can directly start working on SPSS. However, if the dataset is in a different file format (.txt, .xls), then the file needs to be imported. In this demonstration, the dataset is in excel spreadsheet format, which is imported in the SPSS. To import the file, first make sure the file is formatted as follows:

a) The variable name should be given on the top row of the spreadsheet. The data should start from the first column and second row of the spreadsheet.

b) The variable name should only contain alphabets, underscore, and numerals. For example, *Age or Marital_status or Person_origin_1*. It should not contain any special character like space. For example, *Marital Status.* Also, it should not start with a number. For example, *1_person_origin.*

To import the dataset file, click on *File>>Open>>Data*, and then select the dataset file. Once the file gets imported, a window is displayed with two view tabs i.e. *Data view*, and *Variable view*. The *Data view* window displays the current dataset whereas the *Variable view* window displays the information about the attributes of each variable.

In the *Variable view*, the different variables used in the present study are listed. The attributes of the variable are variable name (No_member, Car_ownership), 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 also be modified.

Model development and specifications

To develop the binary logit model, the following step should be followed.

Step 1:

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

Step 2:

Drag and drop the dependent variable (decision to move) to the *Dependent* box, and all the independent variables (person_per_room, monthly_income, and other) to the *Covariate* box.

Step 3:

Click on the *Categorical* tab to define the categorical variables, since SPSS does not define categorical variables in logistic regression itself. A *Logistic Regression: Define Categorical Variables* window gets open as shown in the video.

Step 4:

Drag and drop the categorical variables (Person_origin, Marital_status, and Monthly_income) in the *Categorical covariates* box, and then define the reference category (first or last) for each categorical variable.

Step 5:

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

Step 6:

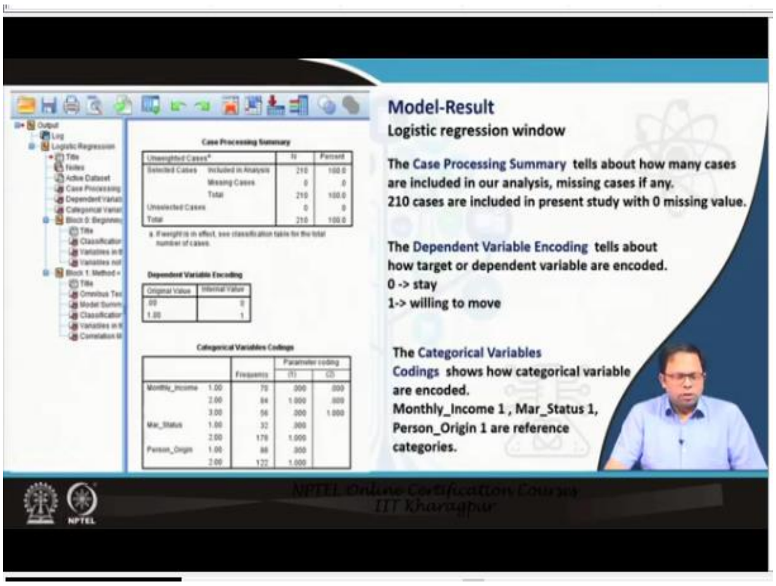
Click on the *options* tab to define model specifications. A *Logistic Regression: Options* window gets open. Here, check the appropriate statistics and plots.

Step 7: Click on the *continue* tab. You will be returned to the *Logistic regression* window.

Step 8: Click the *Ok* tab to run the model.

After following the above mentioned steps, the output for the binary logit model will be generated.

(Refer Slide Time: 12:46)



Case Processing Summary

Unweighted Cases^a		N	Percent
Selected Cases	Included in Analysis	210	100.0
	Missing Cases	0	.0
	Total	210	100.0
Unselected Cases		0	.0
Total		210	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
.00	0
1.00	1

Categorical Variables Codings

		Frequency	Parameter coding	
			(1)	(2)
Monthly_Income	1.00	70	.000	.000
	2.00	84	1.000	.000
	3.00	56	.000	1.000
Mar_Status	1.00	32	.000	
	2.00	178	1.000	
Person_Origin	1.00	88	.000	
	2.00	122	1.000	

Model-Result

Logistic regression window

The Case Processing Summary tells about how many cases are included in our analysis, missing cases if any.
210 cases are included in present study with 0 missing value.

The Dependent Variable Encoding tells about how target or dependent variable are encoded.
0 -> stay
1-> willing to move

The Categorical Variables Codings shows how categorical variable are encoded.
Monthly_Income 1 , Mar_Status 1, Person_Origin 1 are reference categories.

NPTEL Online Certification Courses
IIT Kharagpur

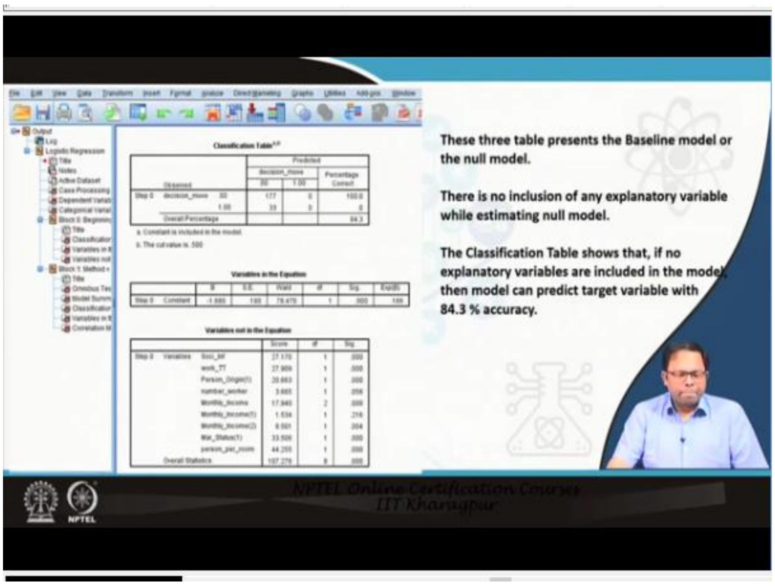
Model results

SPSS produces many tables of output when carrying out binomial logit regression. These tables are important to understand and interpret the results of the regression. The first table is the *case processing summary* which tells about how many cases are included in the analysis, how many are the missing cases, and how many cases are not included in the analysis. In the present model, 210 cases are included with 0 missing cases.

The second table is *dependent variable encoding*, which tells about how the target or dependent variable is encoded. So, original values 0.0 and 1.0 are coded as 0 and 1 in the model. 0 represent ‘stay’ and 1 represents ‘willing to move’.

The third table is *categorical variables coding*. It shows how independent categorical variables are coded, and which category of the variable is considered as a reference category. In the present model, marital status, person origin, and monthly income are the categorical variable and the respective coding is given in the table. For example, person origin has two categories i.e. 1 (local) and 2 (migrant), so person origin 1 is considered as reference and the other is coded as 1 in the model.

(Refer Slide Time: 14:02)



Classification Table[a,b]

Observed			Predicted decision_move .00	Predicted decision_move 1.00	Percentage Correct
Step 0	decision_move	.00	177	0	100.0
		1.00	33	0	.0
	Overall Percentage				84.3

a. Constant is included in the model.
b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.680	.190	78.470	1	.000	.186

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Soci_Inf	27.170	1	.000
		work_TT	27.909	1	.000
		Person_Origin(1)	20.663	1	.000
		number_worker	3.665	1	.056
		Monthly_Income	17.940	2	.000
		Monthly_Income(1)	1.534	1	.216
		Monthly_Income(2)	8.501	1	.004
		Mar_Status(1)	33.506	1	.000
		person_per_room	44.255	1	.000
	Overall Statistics		107.276	8	.000

These three table presents the Baseline model or the null model.

There is no inclusion of any explanatory variable while estimating null model.

The Classification Table shows that, if no explanatory variables are included in the model then model can predict target variable with 84.3 % accuracy.

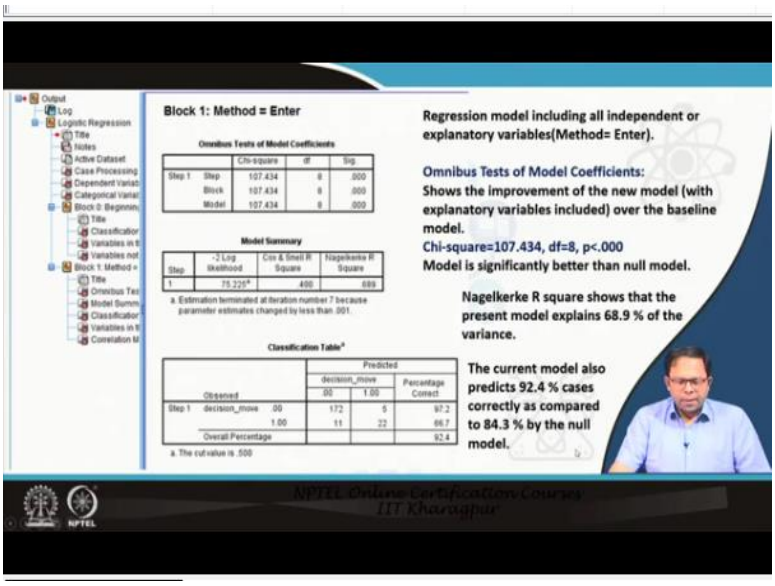
The three tables illustrated in the slide represent the baseline model or null model. In the null model, there is no inclusion of explanatory or independent variables for estimation. In the present case, if no explanatory variables are chosen, then 84.3 % of the cases are predicted correctly by the model. In other words, the target variable is predicted with 84.3 % accuracy.

A general regression equation can be written as follows:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i * x_i \qquad i = 1\ to\ n$$

Where Y is the dependent variable, xi is the independent or explanatory variable, βi are the coefficients of variables, and β0 is the coefficient for the constant. The constant includes the effect of all the variables which are not there in the model but influences the final decision. So, the *variable in the equation* table represents the value of the coefficient of this constant only.

(Refer Slide Time: 15:22)



Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	107.434	8	.000
	Block	107.434	8	.000
	Model	107.434	8	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	75.225ᵃ	.400	.689

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

Classification Tableᵃ

Observed			Predicted		
			decision_move		Percentage Correct
			.00	1.00	
Step 1	decision_move	.00	172	5	97.2
		1.00	11	22	66.7
	Overall Percentage				92.4

a. The cut value is .500

Regression model including all independent or explanatory variables(Method= Enter).

Omnibus Tests of Model Coefficients:
Shows the improvement of the new model (with explanatory variables included) over the baseline model.
Chi-square=107.434, df=8, p<.000
Model is significantly better than null model.

Nagelkerke R square shows that the present model explains 68.9 % of the variance.

The current model also predicts 92.4 % cases correctly as compared to 84.3 % by the null model.

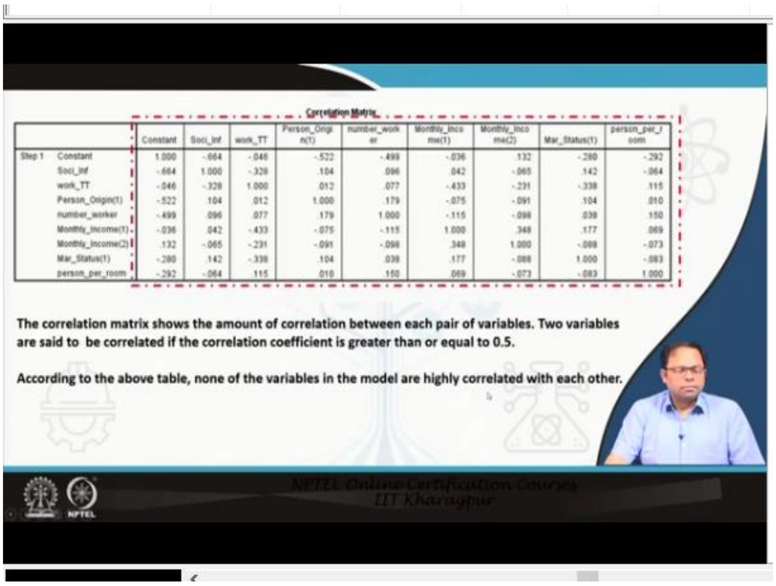
The next set of tables are the outcome of the regression model when all the independent variables are included in the model.

The first table is the outcome of omnibus tests of model coefficients. It is used to check if the new model, where all the explanatory variables are included, is an improvement over the baseline model or not. This test used the chi-square test to check if there is a difference between the log-likelihood of both the model. In the present case, the significance value is less than 0.05. It means that the new model is significantly better than the baseline model.

The model summary gives a value of -2log likelihood and pseudo-R-square for the new model. Here, nagelkerke R square is important, which suggests the percentage of variance explained by the new model. So, the present model explained around 68.9% of the variance in the outcome.

The *classification table* describes how many cases are correctly predicted by the model for each category of the dependent variable. The current model predicts around 92.4% of the cases correctly as compared to 84.3% of the previous model.

(Refer Slide Time: 17:53)



Correlation Matrix

		Constant	Soci_Inf	work_TT	Person_Origin(1)	number_worker	Monthly_Income(1)	Monthly_Income(2)	Mar_Status(1)	person_per_room
Step 1	Constant	1.000	-.664	-.046	-.522	-.499	-.036	.132	-.280	-.292
	Soci_Inf	-.664	1.000	-.328	.104	.096	.042	-.065	.142	-.064
	work_TT	-.046	-.328	1.000	.012	.077	-.433	-.231	-.338	.115
	Person_Origin(1)	-.522	.104	.012	1.000	.179	-.075	-.091	.104	.010
	number_worker	-.499	.096	.077	.179	1.000	-.115	-.098	.038	.150
	Monthly_Income(1)	-.036	.042	-.433	-.075	-.115	1.000	.348	.177	.069
	Monthly_Income(2)	.132	-.065	-.231	-.091	-.098	.348	1.000	-.088	-.073
	Mar_Status(1)	-.280	.142	-.338	.104	.038	.177	-.088	1.000	-.083
	person_per_room	-.292	-.064	.115	.010	.150	.069	-.073	-.083	1.000

The correlation matrix shows the amount of correlation between each pair of variables. Two variables are said to be correlated if the correlation coefficient is greater than or equal to 0.5.

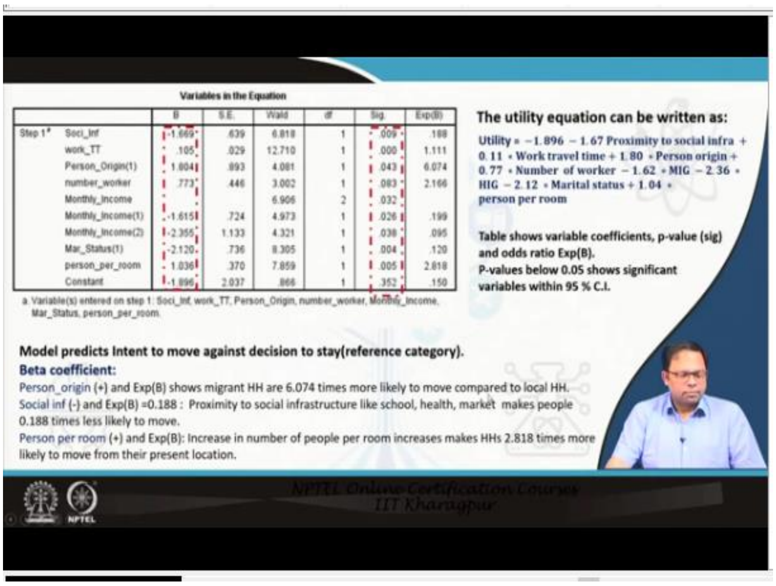
According to the above table, none of the variables in the model are highly correlated with each other.

The given table in the slide is a correlation matrix. It shows the amount of correlation between each pair of variables. In the present table, each cell value represents the correlation coefficient of corresponding variables in row and column. For example, the correlation coefficient of the number of workers and work travel time is 0.77.

If the value of the correlation coefficient is high, then the two variables are said to be highly correlated. If it low, then the two variables are uncorrelated. Generally, the value of a correlation coefficient greater than 0.5 is considered as high. If highly correlated variables are included in the model, then it violates the assumption of the MNL model which states that the independent variables should be uncorrelated. So, one variable is removed from the model.

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

(Refer Slide Time: 18:44)



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Soci_Inf	-1.669	.639	6.818	1	.009	.188
	work_TT	.105	.029	12.710	1	.000	1.111
	Person_Origin(1)	1.804	.893	4.081	1	.043	6.074
	number_worker	.773	.446	3.002	1	.083	2.166
	Monthly_Income			6.906	2	.032	
	Monthly_Income(1)	-1.615	.724	4.973	1	.026	.199
	Monthly_Income(2)	-2.355	1.133	4.321	1	.038	.095
	Mar_Status(1)	-2.120	.736	8.305	1	.004	.120
	person_per_room	1.036	.370	7.859	1	.005	2.818
	Constant	-1.896	2.037	.866	1	.352	.150

a. Variable(s) entered on step 1: Soci_Inf, work_TT, Person_Origin, number_worker, Monthly_Income, Mar_Status, person_per_room.

The utility equation can be written as:

$$\text{Utility} = -1.896 - 1.67\ \text{Proximity to social infra} + 0.11 * \text{Work travel time} + 1.80 * \text{Person origin} + 0.77 * \text{Number of worker} - 1.62 * \text{MIG} - 2.36 * \text{HIG} - 2.12 * \text{Marital status} + 1.04 * \text{person per room}$$

Table shows variable coefficients, p-value (sig) and odds ratio Exp(B).
P-values below 0.05 shows significant variables within 95 % C.I.

Model predicts Intent to move against decision to stay(reference category).

Beta coefficient:

Person_origin (+) and Exp(B) shows migrant HH are 6.074 times more likely to move compared to local HH.

Social inf (-) and Exp(B) =0.188 : Proximity to social infrastructure like school, health, market makes people 0.188 times less likely to move.

Person per room (+) and Exp(B): Increase in number of people per room increases makes HHs 2.818 times more likely to move from their present location.

The *variable in equation* table is the most important part of the model result. This table explains the contribution of each variable in the model. Initially, all the variables are included in the model, but many variable are found to be insignificant. After many iterations, the significant variables are identified. So, the present *variable in equation* table shows the output values for significant variables only.

The table provides regression coefficient estimates (B or β), significance value (sig.), and odds ratio (Exp(B)) for each category of variables. 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. It also signifies whether the dependent and independent variables are positively related or negatively related. The odds ratio signifies the influence of the independent variable on the likelihood of occurrence of one outcome. The significant variables are satisfaction with social infrastructure, work travel time, person origin, number of workers in the family, monthly income category, marital status of the household head, and persons per room. These variables influence the household's willingness to move.

Based on the results, a utility equation can be written as:

$$\text{Utility} = -1.896 - 1.67\ \text{Proximity to social infra} + 0.11 * \text{Work travel time} + 1.80 * \text{Person origin} + 0.77 * \text{Number of worker} - 1.62 * \text{MIG} - 2.36 * \text{HIG} - 2.12 * \text{Marital status} + 1.04 * \text{person per room}$$

The probability of a particular family to move in the future time period can be calculated using this utility equation.

Interpretation of the results

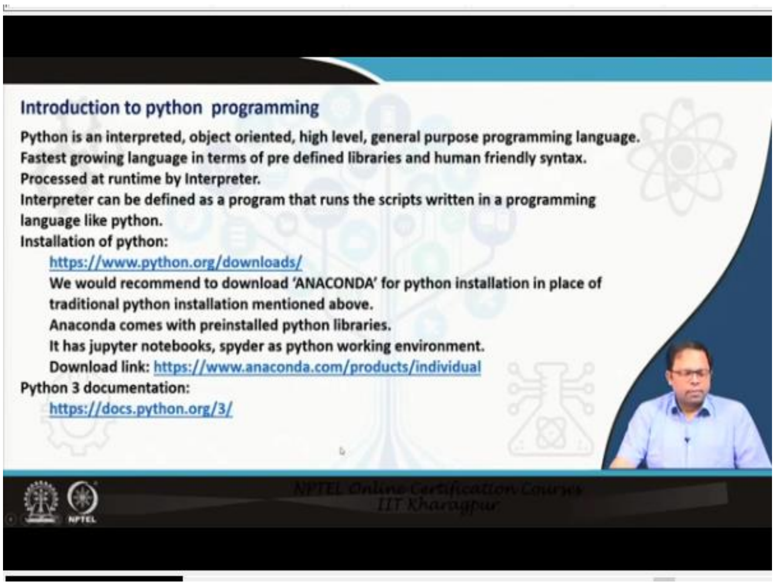
The present model predicts the probability of intention to move against the decision to stay. The decision to stay is a reference category, so all the results will be interpreted with reference to the decision to stay.

The regression coefficient and odds ratio for person origin 1 (migrant) are 1.804 and 6.074 respectively. This variable is positively related to the decision to move. It means that migrant households are more willing to move as compared to local households. Also, the odds ratio signifies that the migrant household is 6.074 more likely to move from the current location as compared to the local household.

The other variable is satisfaction with the proximity to social infrastructure (school, health, and markets). The regression coefficient and odds ratio for satisfaction with the proximity to social infrastructure are -1.669 and 0.188 respectively. So, this variable is negatively related to the decision to move, which means that increase in the satisfaction with the proximity to social infrastructure will make a household to stay in its current location. The value of the odds ratio signifies that the household is 0.188 times less likely to move. Logically, the proximity to health institutions, markets, and schools makes households less likely to move from a location.

Similarly, the other variables like person per room, monthly income, marital status, and work travel time can be interpreted.

(Refer Slide Time: 22:24)



Introduction to python programming

Python is an interpreted, object oriented, high level, general purpose programming language.
Fastest growing language in terms of pre defined libraries and human friendly syntax.
Processed at runtime by Interpreter.
Interpreter can be defined as a program that runs the scripts written in a programming language like python.
Installation of python:

- https://www.python.org/downloads/
- We would recommend to download 'ANACONDA' for python installation in place of traditional python installation mentioned above.
- Anaconda comes with preinstalled python libraries.
- It has jupyter notebooks, spyder as python working environment.
- Download link: https://www.anaconda.com/products/individual

Python 3 documentation:

- https://docs.python.org/3/

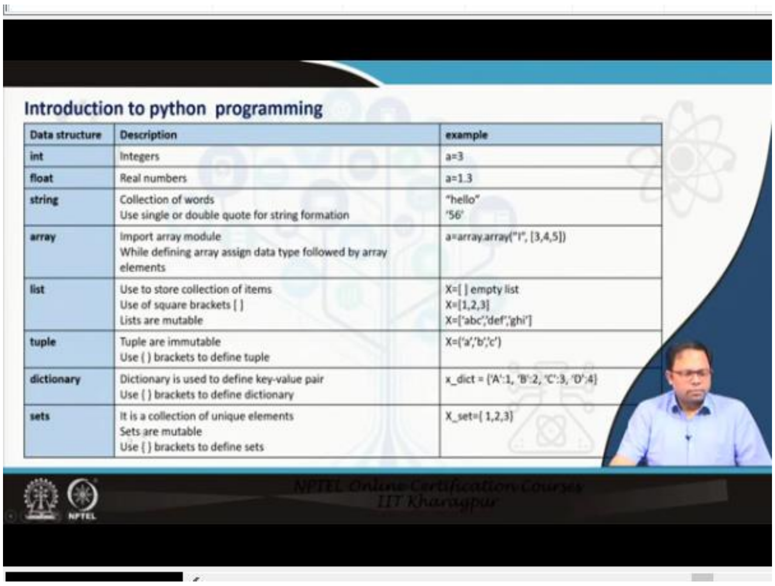
Introduction to python programming

The model developed using SPSS, can also be developed using programming languages like python, R-programming, C, or C++. In this lecture, model development using python has been discussed.

Python is a programming language. It is an interpreted, object-oriented, high level, and general purpose programming language. It is the fastest growing language in terms of pre-defined libraries and human friendly syntax. It is very easy to develop code on this programming language. The coding is almost like writing normal English language. The code is processed at runtime by the interpreter. Interpreter is a program that runs the scripts written in a programming language like Python. So, the interpreter executes the actual code.

Python can be installed from **https://www.python.org/downloads/**, or ANACONDA **https://www.anaconda.com/products/individual.** ANACONDA is an open source distribution platform. It comes with a lot of pre installed Python libraries. Also, it has a python working environment like Jupyter notebook and spyder. The Python language has an extensive body of documentation. Python 3 documentation can be accessed from **https://docs.python.org/3/.**

(Refer Slide Time: 24:05)



Introduction to python programming

Data structure	Description	example
int	Integers	a=3
float	Real numbers	a=1.3
string	Collection of words Use single or double quote for string formation	"hello" '56'
array	Import array module While defining array assign data type followed by array elements	a=array.array("I", [3,4,5])

list	Use to store collection of items
	Use of square brackets [ ]
	Lists are mutable
		X=[ ] empty list
		X=[1,2,3]
		X=['abc','def','ghi']

tuple	Tuple are immutable Use ( ) brackets to define tuple	X=('a','b','c')
dictionary	Dictionary is used to define key-value pair Use { } brackets to define dictionary	x_dict = {'A':1, 'B':2, 'C':3, 'D':4}
sets	It is a collection of unique elements Sets are mutable Use { } brackets to define sets	X_set={ 1,2,3}

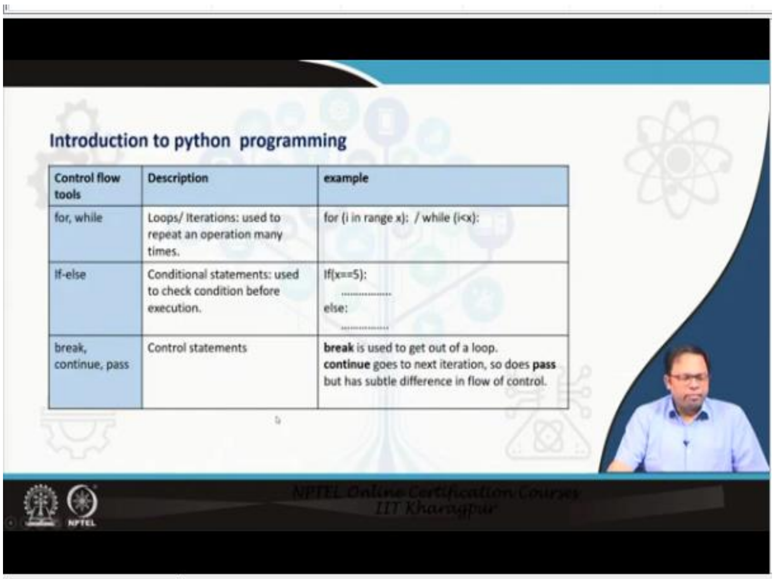
Data types

In computer programming, a physical problem requires certain variables to translate it into an analytical problem. These variables can be of integer type, character type, a group of characters (string), a set of numbers (arrays), etc. A compiler or interpreter has predefined datatypes to identify these variables. In other words, datatypes are predefined keywords used to identify variables of different types and allocate physical memory for the same. The various data structures/data types with examples are given in the table.

The first data type is integer, it represents numeric data (whole numbers) like 3, 4, or 18. The second data type is a float which represents a real number (specified with a decimal) like 1.3. Integer and float represent numeric data. The collection of words or characters are string type data. String is defined using a single quote or double quote like "hello", '56'. These datatypes can be created in python by assigning them to a variable. For example, a=3, a=1.3, and a="hello".

Array, list, tuple, dictionary, and sets are built-in data structure in python. These are used to store data. In python, the array can be created by importing the array module. The syntax for creating an array is array (data_type, value_list). For example, a=array.array("I",[3,4,5]). List is defined by square brackets, for example [1, 2, 3]. Tuple is similar to the list but it is immutable. It can be created by placing values within brackets ( ), like (1,2,3). Dictionary is a sequence of items, where each item has a key and a value. It can be created using curly brackets { }. Another data type is set, which is a collection of unique elements. It is created by placing the items within curly brackets. The example for each datatype and data structure is given in the table.

(Refer Slide Time: 25:34)



Introduction to python programming

Control flow tools	Description	example
for, while	Loops/ Iterations: used to repeat an operation many times.	for (i in range x): / while (i<x):

If-else	Conditional statements: used to check condition before execution.	If(x==5):
		……………
		else:
		……………

break, continue, pass	Control statements	**break** is used to get out of a loop. **continue** goes to next iteration, so does **pass** but has subtle difference in flow of control.

Control structures

Control structures are the method to control the flow of the programs. The three basic control flow in any programming language are sequential, selection, and repetition. Sequential is the default control structure, selection are conditional statements like if-else, and repetition are the loops like for loop, and while loop.

The 'for' loop and 'while' loop is used to repeat an operation many times. The syntax for 'for' loop and 'while' loop are as follows:

```
for i in x:
        expression
while i<x:
        expression
```

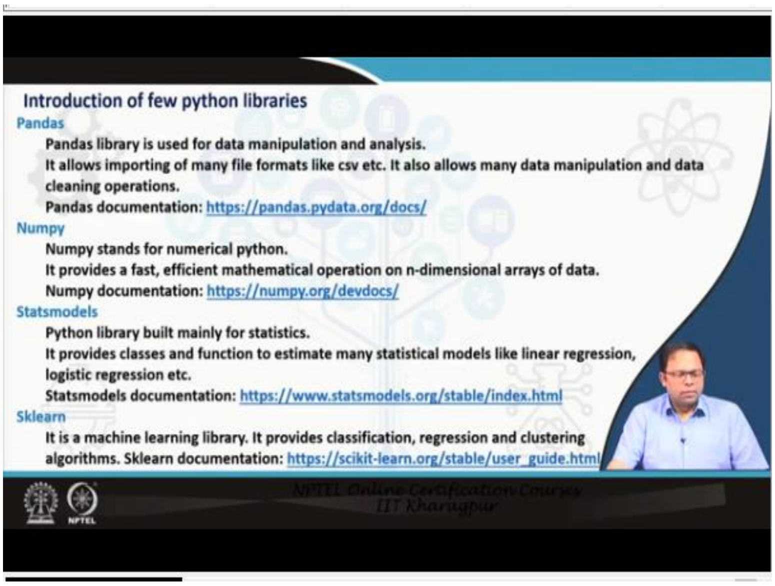
The 'if' and 'else' are used to check the condition before executing the code. The 'if' statement checks the condition and runs the expression if the condition is true. Whereas 'else' executes the expression when the condition of 'if' is false. The syntax for 'for' and 'while' loop are as follows:

```
if i==x:
        expression
else:
```

expression

In addition to the above, there are a few loop control statements such as break, continue, and pass. The break is used to terminate the loop or get out of the loop, whereas continue forces to execute the next iteration. The pass statement is used when the programmer does not want to execute a certain function within the code.

(Refer Slide Time: 26:26)



Introduction of few python libraries

Pandas

Pandas library is used for data manipulation and analysis.

It allows importing of many file formats like csv etc. It also allows many data manipulation and data cleaning operations.

Pandas documentation: https://pandas.pydata.org/docs/

Numpy

Numpy stands for numerical python.

It provides a fast, efficient mathematical operation on n-dimensional arrays of data.

Numpy documentation: https://numpy.org/devdocs/

Statsmodels

Python library built mainly for statistics.

It provides classes and function to estimate many statistical models like linear regression, logistic regression etc.

Statsmodels documentation: https://www.statsmodels.org/stable/index.html

Sklearn

It is a machine learning library. It provides classification, regression and clustering algorithms. Sklearn documentation: https://scikit-learn.org/stable/user_guide.html

Python libraries

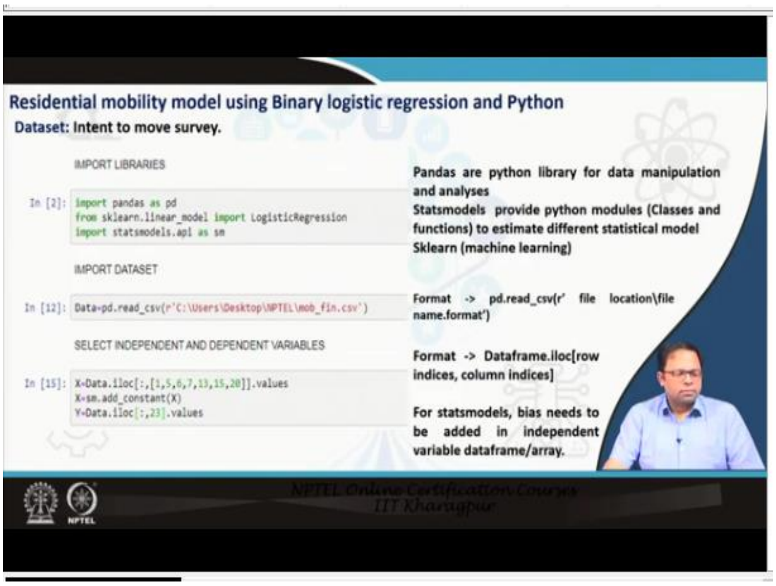
Python uses several libraries for statistical modelling, data manipulation, etc. For example, pandas, numpy, sklearn, statsmodels, and others. These libraries will be used to develop a residential mobility model and the residential location choice model.

Pandas library is used for data manipulation and analysis. It allows the importing of many file formats like csv. It also allows many data manipulation and data cleaning operations. Numpy stands for numerical Python, it allows many mathematical operations on n-dimensional arrays of data.

Statsmodels is a python library built mainly for statistics. It provides classes and functions to estimate many statistical models like linear regression, logistic regression, and others. Similar to statsmodel, sklearn is also used for statistical models. It is a machine learning library that provided classification, clustering, and regression algorithms.

The documentation for each library can be accessed from the links mentioned in the above slide. These libraries are basically groups of functions which are already written, one needs to understand which function does what, and the data entry format for that particular function.

(Refer Slide Time: 28:11)



Residential mobility model using Binary logistic regression and Python

Dataset: Intent to move survey.

IMPORT LIBRARIES

```
In [2]: import pandas as pd
        from sklearn.linear_model import LogisticRegression
        import statsmodels.api as sm
```

IMPORT DATASET

```
In [12]: Data=pd.read_csv(r'C:\Users\Desktop\NPTEL\mob_fin.csv')
```

SELECT INDEPENDENT AND DEPENDENT VARIABLES

```
In [15]: X=Data.iloc[:,[1,5,6,7,13,15,20]].values
         X=sm.add_constant(X)
         Y=Data.iloc[:,23].values
```

Pandas are python library for data manipulation and analyses

Statsmodels provide python modules (Classes and functions) to estimate different statistical model

Sklearn (machine learning)

Format -> pd.read_csv(r' file location\file name.format')

Format -> Dataframe.iloc[row indices, column indices]

For statsmodels, bias needs to be added in independent variable dataframe/array.

Residential mobility model using Binary logit regression and Python

The concept of the binary logit model can be implemented in python programming as well. In the present example, the intention to move model developed in SPSS is again developed using python. The format of the dataset is also the same, which means that each column represents a single variable and each row represents a case. The general steps to develop a binary logit model in python are:

Step 1: Import libraries such as pandas, statsmodels.

Step 2: Import dataset, and define independent and dependent variables.

Step 3: Create a model.

Data input

At the first step, libraries or packages like *pandas, statsmodel* are imported. The libraries are loaded with alias *pd* and *sm* respectively. The pandas library is for reading the dataset, and statsmodels library is for calling logistic regression function. The code for importing libraries is:

import pandas as pd

import statsmodels.api as sm

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. So, the file extension for the present dataset is .csv. A CSV file is a text file with values separated by commas. 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\mob_fin.csv')

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

Dataframe.iloc[row indices, column indices]

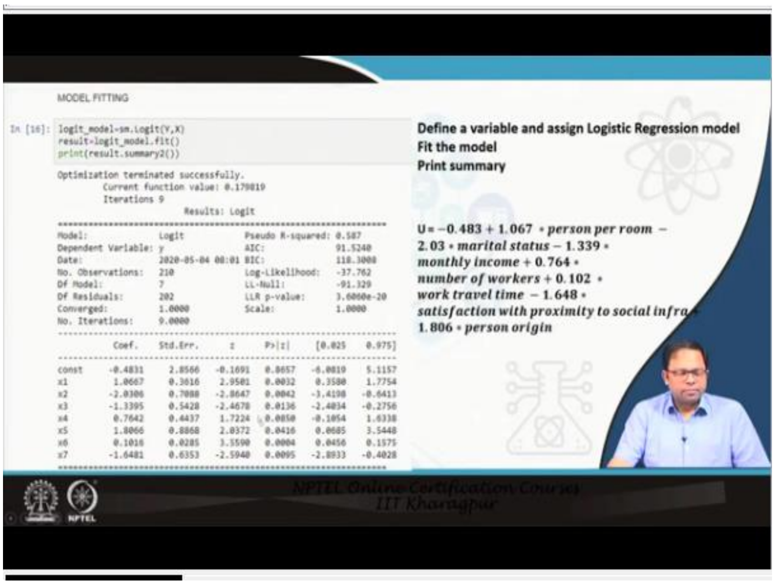
So, the independent variables (person origin, monthly income, marital status, and others) are assigned to variable 'X' and the dependent variable (decision to move or stay) is assigned to variable 'Y'. Also, a bias needs to be added in the independent variable data frame for statsmodels.

X=Dtata.iloc[:,[1,5,6,7,13,15,20]].values

X=sm.add_constant(X)

Y=Data.iloc[:,23].values

(Refer Slide Time: 30:00)



MODEL FITTING

In [16]:
```
logit_model=sm.Logit(Y,X)
result=logit_model.fit()
print(result.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.179819
         Iterations 9
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.587
Dependent Variable: y                AIC:              91.5240
Date:               2020-05-04 08:01 BIC:              118.3008
No. Observations:   210              Log-Likelihood:   -37.762
Df Model:           7                LL-Null:          -91.329
Df Residuals:       202              LLR p-value:      3.6060e-20
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
------------------------------------------------------------------
          Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
------------------------------------------------------------------
const    -0.4831   2.8566  -0.1691  0.8657  -6.0819   5.1157
x1        1.0667   0.3616   2.9501  0.0032   0.3580   1.7754
x2       -2.0306   0.7088  -2.8647  0.0042  -3.4198  -0.6413
x3       -1.3395   0.5428  -2.4678  0.0136  -2.4034  -0.2756
x4        0.7642   0.4437   1.7224  0.0850  -0.1054   1.6338
x5        1.8066   0.8868   2.0372  0.0416   0.0685   3.5448
x6        0.1016   0.0285   3.5590  0.0004   0.0456   0.1575
x7       -1.6481   0.6353  -2.5940  0.0095  -2.8933  -0.4028
=================================================================
```

Define a variable and assign Logistic Regression model
Fit the model
Print summary

$$U = -0.483 + 1.067 * \textit{person per room} - 2.03 * \textit{marital status} - 1.339 * \textit{monthly income} + 0.764 * \textit{number of workers} + 0.102 * \textit{work travel time} - 1.648 * \textit{satisfaction with proximity to social infra} + 1.806 * \textit{person origin}$$

Model fitting

The final step is to create a model. In order to perform logistic regression, Statsmodels provides a Logit ( ) function. This function takes dependent (Y) and independent (X) variables to produce output. So, first, a variable *logit_model* is defined and a logistic regression model is assigned to it. Then, the model is fitted to the data. Finally, the output is printed. The code for calling the Logit function, fitting the model, and printing the model summary is given below.

```
logit_model=sm.Logit(Y,X)
result=logit_model.fit( )
print(result.summary2( ))
```

Model result and interpretation

It is important to mention that the reference category for each categorical variable is the first category like in SPSS. So, the reference category for person origin is local, monthly income is lower income group, and marital status is unmarried. The model predicts the decision to move in reference to the decision to stay.

The model summary includes the estimates of regression coefficients, model specifications such as log-likelihood for null and full model, and basic summary like model type, number of iterations, degree of freedom etc.

In the present model, the model type is Logit, the dependent variable is ‘y’ (decision to move or stay), the number of observation is 210, the number of iteration when the model converged or log likelihood maximized is 9, the maximized value of log likelihood for full model is -37.762, the maximized value of log likelihood for the null model is -91.329, and the pseudo R-square value is 0.587.

The coefficient values for each independent variable given in the model summary is shown below.

Variable	Variable name	Coefficients (β values)	Sig. value (p-value)
Constant		-0.4831	0.8657
X1	Person per room	1.0667	0.0032
X2	Marital status	-2.0306	0.0042

X3	Monthly income	-1.3395	0.0136
X4	Number of workers	0.7642	0.0850
X5	Work travel time	1.8066	0.0416
X6	Satisfaction with proximity to social infrastructure	0.1016	0.0004
X7	Person origin	-1.6481	0.0095

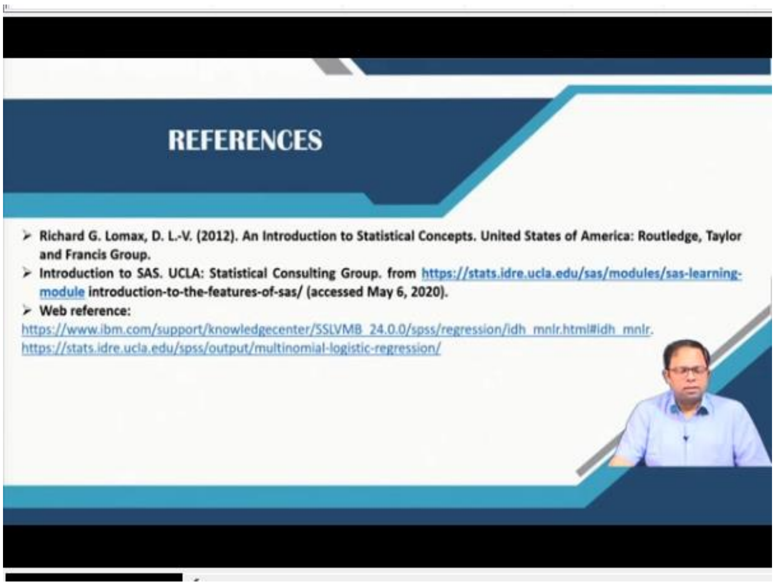
All the independent variables are found to be significant i.e. p-value <0.05. The variables such as marital status, monthly income, and person origin are negatively related to the decision to move. For example, an unmarried household head is more willing to move as compared to a married household head. The variables such as satisfaction with proximity to social infrastructure, person per room, work travel time, and number of workers are positively related to the decision to move. For example, an increase in work travel time will increase the willingness to move from the current location. Similarly, other variables can be interpreted.

Also, based on the results, the utility equation can be written as:

$$\text{Utililty} = -0.483 + 1.067 * \text{person per room} - 2.03 * \text{marital status} - 1.339 * \text{monthly income} + 0.764 * \text{number of workers} + 0.102 * \text{work travel time} - 1.648 * \text{satisfaction with proximity to social infra} + 1.806 * \text{person origin}$$

The probability of a particular family to move in the future time period can be calculated using this utility equation.

(Refer Slide Time: 31:51)



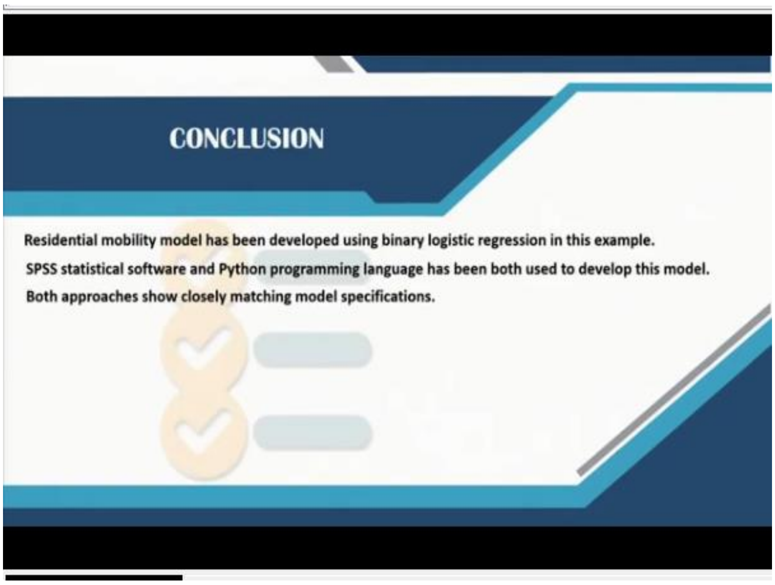
REFERENCES

- Richard G. Lomax, D. L.-V. (2012). An Introduction to Statistical Concepts. United States of America: Routledge, Taylor and Francis Group.
- Introduction to SAS. UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/sas/modules/sas-learning-module introduction-to-the-features-of-sas/ (accessed May 6, 2020).
- Web reference:

https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/regression/idh_mnlr.html#idh_mnlr.
https://stats.idre.ucla.edu/spss/output/multinomial-logistic-regression/

Reference are listed in the above slide.

(Refer Slide Time: 31:54)



CONCLUSION

Residential mobility model has been developed using binary logistic regression in this example.
SPSS statistical software and Python programming language has been both used to develop this model.
Both approaches show closely matching model specifications.

Conclusion

The residential mobility model has been developed using binary logistic regression in this example. SPSS statistical software and Python programming language has been both used to develop this model. Both approaches show closely matching models specification.