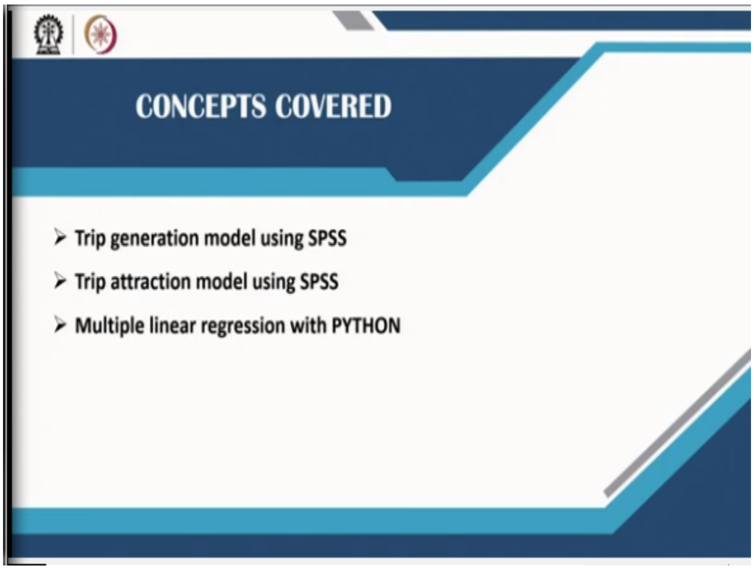
Urban Land use and Transportation Planning
Prof. Debapratim Pandit
Agricultural and Regional Planning Department
Indian Institute of Technology-Kharagpur

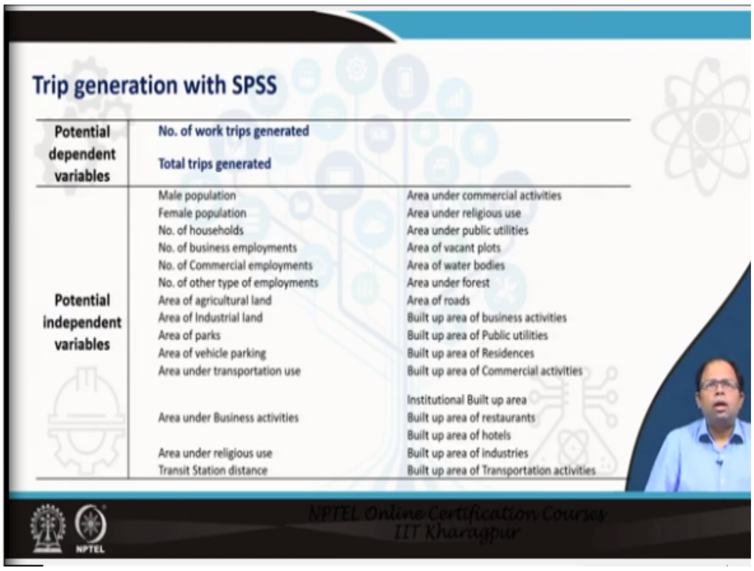
Lecture-34
Trip Production and Attraction 2

This lecture is a continuation of the trip production and attraction and covers development of trip generation model using SPSS; development of trip attraction using SPSS; and performing MLR in python.

(Refer Slide Time: 00:29)



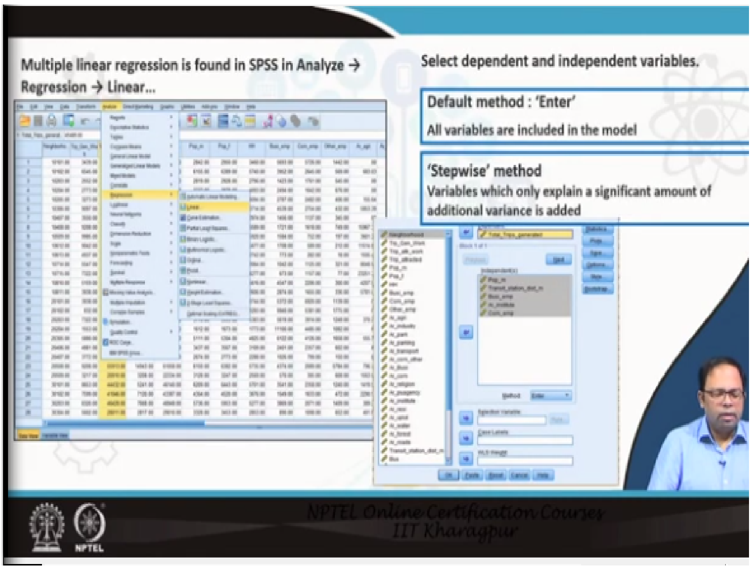
(Refer Slide Time: 00:44)



Trip Generation with SPSS:

In trip generation model, the intention is to predict is the total number of work trips generated or total trips generated from a particular zone. So, these are the potential dependent variables that are considered and the potential independent variables, as shown in the table are male population, female population, number of households, number of business employments, commercial employments, and other types of employments. Apart from these, areas under different land uses; built-up area of business activities, religious use, commercial activities; transit station distance are also considered. These are similar to the variables discussed during the use of regression model in trip generation.

(Refer Slide Time: 01:50)



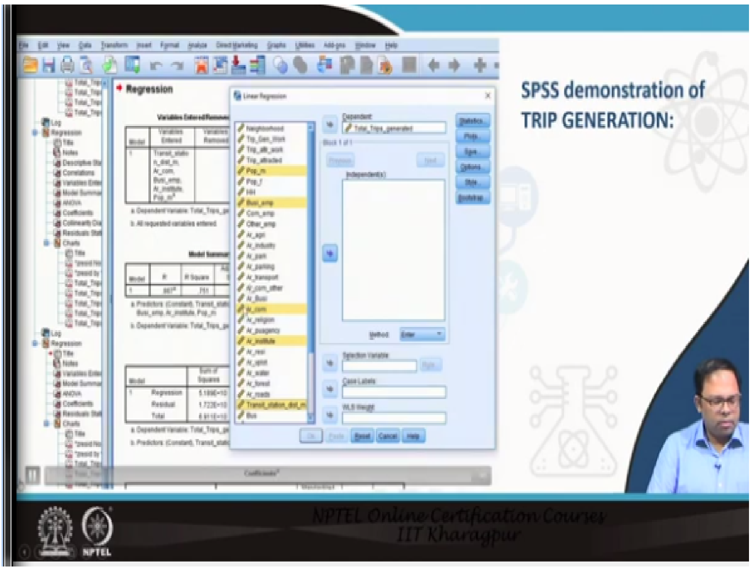
Trip generation demonstration in SPSS:

After doing the data entry in SPSS, the following steps are to be performed in order to develop a regression model of trip generation:

$Menu\ bar \rightarrow Analyze \rightarrow Regression \rightarrow Linear\ \dots \rightarrow$
$Select\ dependent\ and\ independent\ variables \rightarrow Specify\ method \rightarrow$
$Select\ statistics\ to\ be\ estimated \rightarrow (Optional)\ specify\ elements\ for\ X-axis\ and\ Y-$
$axis\ for\ plot(s) \rightarrow (Optional)\ specify\ elements\ to\ be\ saved\ separately \rightarrow$
$(Optional) select\ options\ for\ stepwise\ removal \rightarrow Click\ '\boldsymbol{OK}'$

From the menu bar, **analyze** option is to be selected. From the dropdown menu, **regression** needs to be selected, and then **linear** needs to be selected. The invoking of the linear regression results in a dialogue box popping up, which has three sections; a section with all the variables created in the database, a section of **dependent variable**, and another for **independent variables**. From the section with all the variable names mentioned, appropriate variable(s) are to be dragged and dropped in either of the other two sections i.e. dependent and independent sections. After that the method to be followed for including variables in the model needs to be specified from the dropdown menu adjacent to **method**. The default method is '**Enter'**, which allows all the selected independent variables in the model together. Another method is **'Step-wise'**, which includes the independent variables that explain a significant amount of variance in dependent variable. Other available methods are; step-wise, forward, backward, etc.

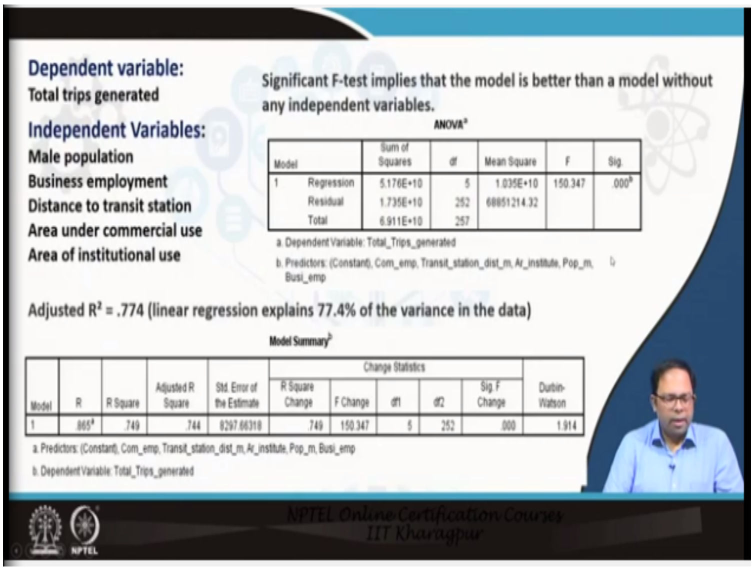
(Refer Slide Time: 02:46)



(Video Starts: 02:51)

After selecting the method of introducing variables, various statistics and options need to be specified from the different tabs at the top right corner of the regression dialogue box. In the '**Statistics**' tab, various statistics can be chosen to be estimated and displayed like, model fit, R-square changes, descriptives, Durbin-Watson test, etc. In the '**Plots**' tab, x-axis and y-axis can be specified for various plots like, predicted values, dependent variable, residuals, etc. This tab also has the option to see the PP-plots (normal probability plots) and histogram for each variable. In the '**Save**' tab, the values of various forms of distances, predicted values, residuals, etc. can be selected to be stored in separate columns for each observation in the dataset. In the '**Options**' tab, exclusion criteria can be specified. Other tabs like, style, bootstrap, selection variable, etc. can be explored and used as and when required. Finally, on selecting '**OK**', the package estimates the results based on the selections in the regression dialogue box and is displayed in an '**Output window**'. In the demonstration, descriptive statistics, correlations, predicted value vs. residual plot, table of coefficients, model summary, ANOVA test can be seen in the output window. From the predicted value vs. residual plot, heteroscedasticity in the data can be checked. **(Video Ends: 04:08)**

(Refer Slide Time: 04:09)

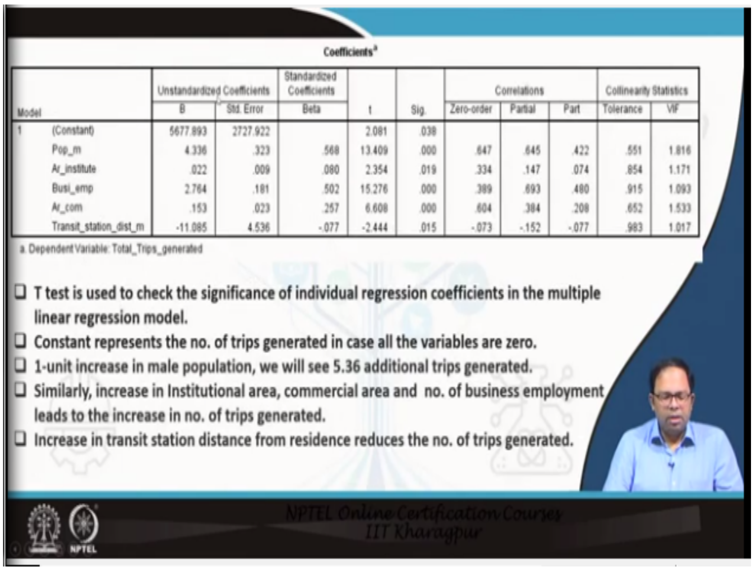


Model interpretation of trip generation:

In this demonstration, total trips generated was taken as the dependent variable. The independent variables chosen were; male population, business employment, distance to transit station, area under commercial use, and area of institutional use. Male population, rather the working age groups of male population have impact on trip generation. But as per the data in the demonstration, male population was found to be more appropriate. Total number of business employment actually determines the number of work trips that would be generated. As many people move from office to retail area for shopping purposes, area under commercial use and area under institutional use also plays a role in trip generation. The distance to transit station also plays a role in trip generation.

In the ANOVA table, the statistically significant value of F-statistic shows that the model is and it significantly better than the model without any independent variables. The model summary table shows that the adjusted R square value is 0.744, which implies that this regression model explains around 77.4% of the variance in the data. The Durbin Watson statistic which is slightly less than 2 indicates the presence of a slightly positive auto-correlation, but it is negligible.

(Refer Slide Time: 06:18)



The output also has the table of coefficients that has the coefficients (beta values), the t-test statistic, and the p-value of each independent variable and the constant as well. All the variables were found to be statistically significant and positively related to trips generated, except for 'distance from transit' which has a negative beta value. This implies that an increase in the distance from transit results in the decrease in the number of trips. The constant represents the number of trips generated in case all the variables is 0. Unlike the F-test testing the overall significance of the model, t-test tests the significance of individual coefficients. If the absolute value of t-statistic is 1.97≤, the p-value is ≥0.05.

A one unit change in the male population results in the increase in 4.33 increase in the number of trips. Similarly, one-unit increase in the area under institutional use, business employment, and area under commercial use results in the increase in 0.022, 2.76, 0.153 trips respectively. One-unit increase in the distance from transit station results in the reduction of around 11 trips. All of these results match with the theoretical underpinnings of the model and implies that the signs are what they should be. Hence it can be said that the model is valid.

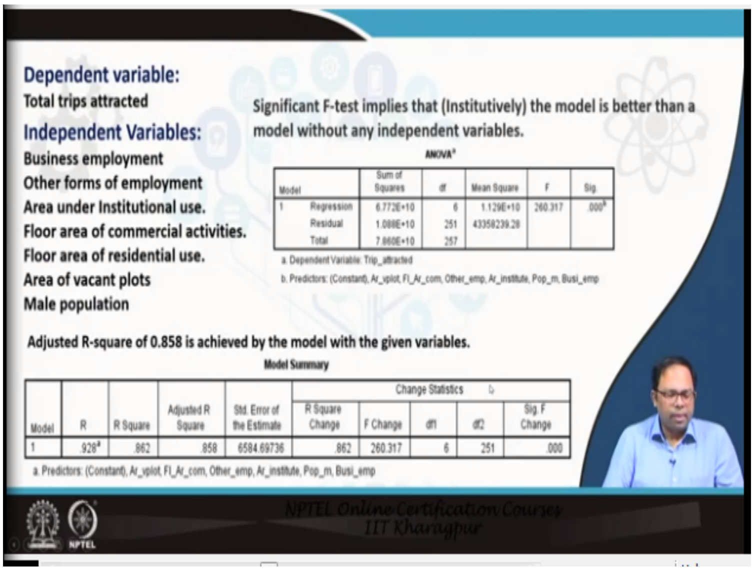
(Refer Slide Time: 08:03)

Trip attraction with SPSS:

In the trip attraction model, the potential dependent variables considered are total work trips that are attracted and total trips that are attracted. The potential independent variables are same as discussed in trip generation, which are; number of households; male population; female population; number of households; number of business employments; commercial employments; other types of employments; areas under different land uses; built-up area of business activities, religious use, commercial activities, roads; transit station distance; etc.

For some aspects like business or commercial, number of employment might be a better indicator as compared to built-up area or area under landuse. The subtleties of the appropriateness of the different kinds of the same variable needs to be understood or a combination of these can also be taken.

(Refer Slide Time: 09:29)

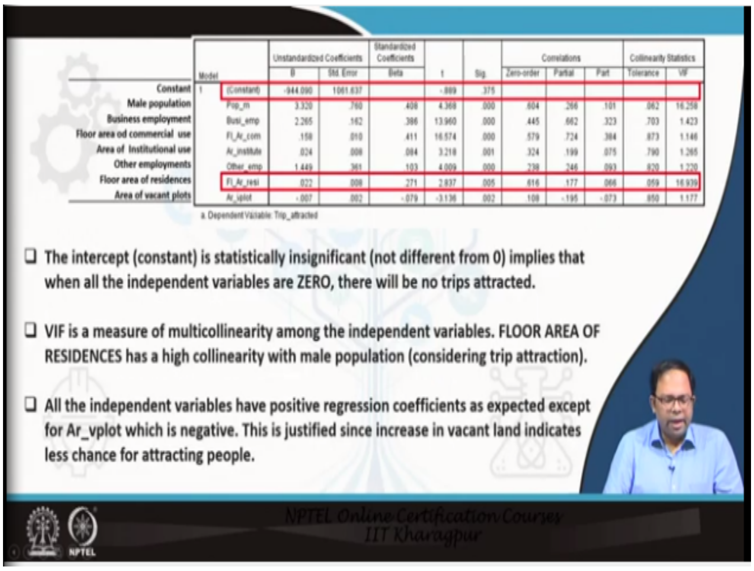


Model interpretation of trip attraction:

In the demonstration model for trip attraction, total trips attracted has been taken as the dependent variable. The independent variables used were business employment, other forms of employment, area under institutional use, floor area of commercial activities, floor area of residential use, area of vacant plots, and male population. It should be noted that business employment has been considered against business land use. Area of vacant plots represent the empty space in the respective zone.

In the ANOVA table, the statistically significant value of F-statistic shows that the model is and it significantly better than the model without any independent variables. The model summary table shows that the adjusted R square value is 0.858, which implies that this regression model explains around 85.8% of the variance in the data. Although many people have different opinions, but most people consider any value of adjusted R-square above 0.75 as a good fit for a dataset generated from natural environment.

(Refer Slide Time: 10:56)

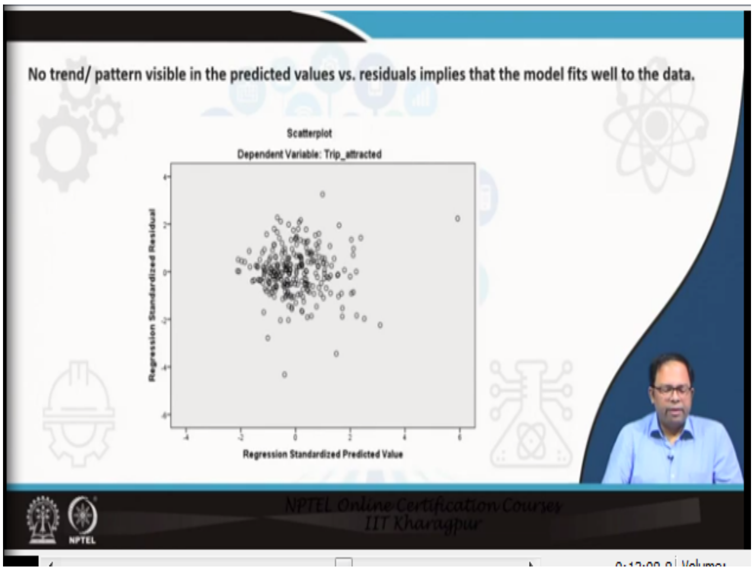


	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
			B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
Constant	1	(Constant)	-944.090	1061.637		-.889	.375					
Male population		Pop_m	3.320	.760	.408	4.368	.000	.604	.266	.101	.062	16.258
Business employment		Busi_emp	2.265	.162	.386	13.960	.000	.445	.662	.323	.703	1.423
Floor area od commercial use		Fl_Ar_com	.158	.010	.411	16.574	.000	.579	.724	.384	.873	1.146
Area of Institutional use		Ar_institute	.024	.008	.084	3.218	.001	.324	.199	.075	.790	1.265
Other employments		Other_emp	1.449	.361	.103	4.009	.000	.238	.246	.093	.820	1.220
Floor area of residences		Fl_Ar_resi	.022	.008	.271	2.837	.005	.616	.177	.066	.059	16.939
Area of vacant plots		Ar_vplot	-.007	.002	-.079	-3.136	.002	.108	-.195	-.073	.850	1.177

a. Dependent Variable: Trip_attracted

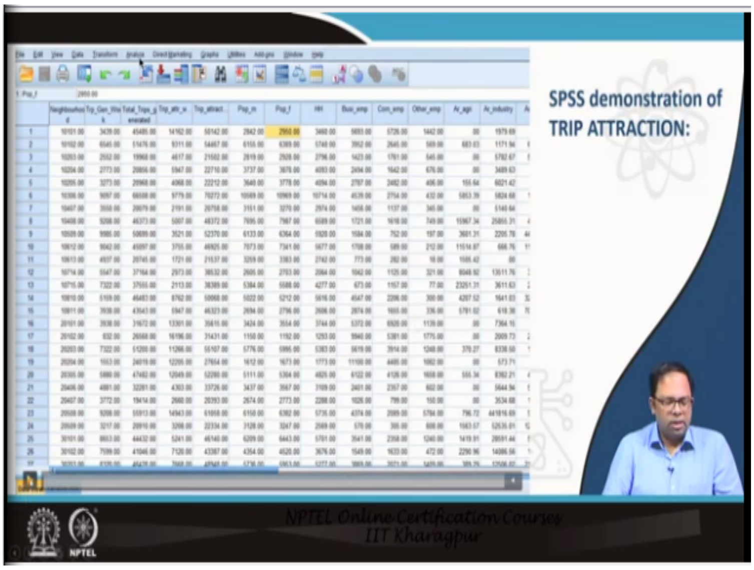
The estimated coefficient of the constant is not statistically significant. The estimated value of the constant implies that when all the other variables are 0, no trips will be attracted. It is interpreted so as the estimated constant is less than zero but in reality number of trips cannot be a negative entity. VIF is a measure of collinearity among independent variables. The variable 'floor area of residences' is found to have high collinearity with male population in this trip attraction model. So, even though the significance value for this variable is 0.005, since it is highly multicollinear with male population, it should be ignored and removed from the model. All independent variables have positive regression coefficients except for 'Ar_vplot' which is negative, all of which do have theoretical underpinnings. Increase in vacant plot implies no landuse and hence no trips attracted, ultimately reducing the total trips attracted to the respective zone.

(Refer Slide Time: 13:08)



The plot of residuals against the dependent variable shows no tend as such. Such a plot having no trend or pattern implies that the model fits properly with the given data.

(Refer Slide Time: 13:29)



Trip attraction demonstration in SPSS:

The course of actions to develop a trip attraction model in SPSS is similar to that for trip generation, apart from the choice of dependent and independent variables. In this case, the dependent variable is '**trip attracted**'. The different independent variables selected are same as discussed in the above section. One thing to note is that it is very unlikely that we will arrive at the best model in the first attempt. One might have to try numerous combinations of independent variables before a satisfactory model or the best model (given the dataset) is found.

Various statistics like correlation coefficients, R-square, etc. are used to judge a model and provide a basis for including or excluding variables from a model. For example, in the case of the model in demonstration, ‘floor area of residences’ cannot be kept in the model because of high correlation with male population. This is how a trip attraction model can be developed model using SPSS. **(Video Ends: 14:43)**

(Refer Slide Time: 14:44)

<u>Multiple linear regression with Python:</u>

Instead of using statistical packages like SPSS, python, R, etc. can be used to estimate multiple linear regression (MLR). The use of python for MLR is covered in this section. Python is a high-level, interpreted, general-purpose programming language which is very popular due to its readability and packages available. It can be freely downloaded from www.python.org or it can be installed through anaconda, which is a free and open source distribution of python for scientific computing. The introduction to python and a guide to developing a logistic regression model using it has been covered already in a different module.

In python, functions or a group of functions are bundled in entities called ‘**packages**’. Packages are imported to before writing a program and then functions are invoked from the code as and when required. One such package is ‘**Pandas**’, which has functions for data analysis and processing. Another package is ‘**scikit-learn**’ (Sklearn**)** that has popular machine learning models.

Linear regression is also a model in this package. Following is the code for linear regression using python:

Python code for trip generation and trip attraction model using SKLEARN.

```
import pandas as pd
from sklearn.linear_model import LinearRegression
```

- *Importing pandas and scikit-learn*

```
df=pd.read_csv(r'file.csv')
df.fillna(0,inplace=True)
X_gen=df.iloc[:,[5,8,17,21,39]]
Y_gen=df.iloc[:,2]
X_attr=df.iloc[:,[5,8,10,21,23,30]]
Y_attr=df.iloc[:,4]
```

- *Reading file into dataframe through **pandas.read***
- *Replacing blank spaces with '0'.*
- *Segregating dependent variables and independent variables for generation and attraction models by COLUMN IDs from the whole dataset.*

```
gen=LinearRegression().fit(X_gen,Y_gen)
b_gen=gen.coef_
b0_gen=gen.intercept_

attr=LinearRegression().fit(X_attr,Y_attr)
b_attr=attr.coef_
b0_attr=attr.intercept_
```

- *Invoking Linear regression for generation and storing the coefficients and intercept in **b_gen** and **b0_gen**.*
- *Invoking Linear regression for attraction and storing the coefficients and intercept in **b_attr** and **b0_attr**.*

```
print("TRIP GENERATION:\n","Intercept:",b0_gen)
varg=X_gen.columns.tolist()
for i in range(len(varg)):
   print(varg[i], b_gen[i])
rg=gen.score(X_gen,Y_gen)
print("R-squared=",rg,"\n")

print("TRIP ATTRACTION:\n","Intercept:",b0_attr)
vart=X_attr.columns.tolist()
for j in range(len(vart)):
   print(vart[j], b_attr[j])
ra=attr.score(X_attr,Y_attr)
print("R-squared=",ra)
```

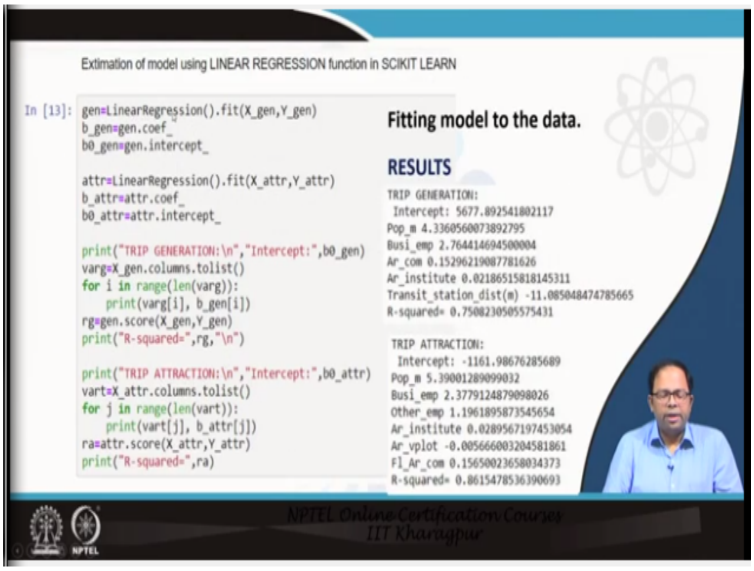
- *Printing the results….*

```
TRIP GENERATION:
 Intercept: 5677.892541802117
Pop_m 4.3360560073892795
Busi_emp 2.764414694500004
Ar_com 0.15296219087781626
Ar_institute 0.02186515818145311
Transit_station_dist(m) -11.085048474785665
R-squared= 0.7508230505575431

TRIP ATTRACTION:
 Intercept: -1161.98676285689
Pop_m 5.39001289099032
Busi_emp 2.3779124879098026
Other_emp 1.1961895873545654
Ar_institute 0.0289567197453054
Ar_vplot -0.005666003204581861
Fl_Ar_com 0.15650023658034373
R-squared= 0.8615478536390693
```

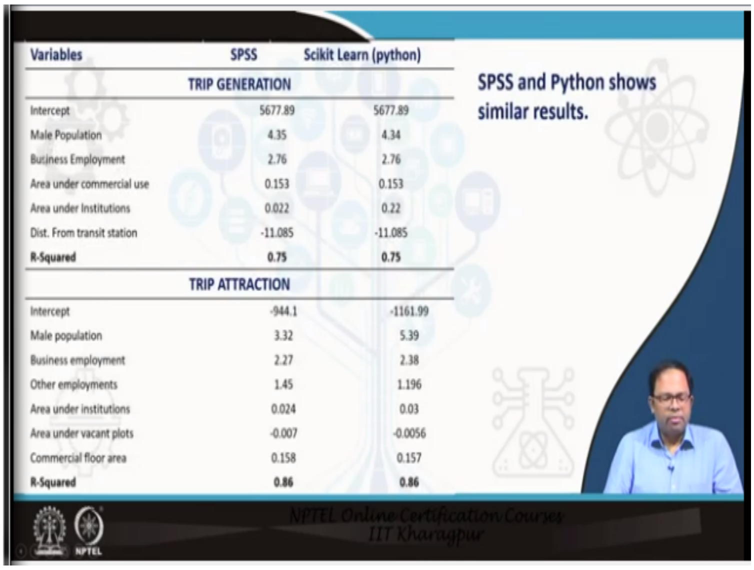
In the shown code, both trip generation and trip attraction is modelled. At first, the packages needs to be included in the code, the data file needs to be read in comma separated values (**csv**) format. It is read through '**pandas.read**' and stored in a dataframe. From this dataframe, appropriate independent variables and dependent variables for either of the models are stored in arrays by their column IDs in the original dataset. 'Y_gen' and 'Y_attr' are arrays that store the dependent variables for trip generation and trip distribution respectively. 'X_gen' and 'X_attr' are two-dimensional arrays that store the independent variables for either of trip generation and trip attraction respectively.

(Refer Slide Time: 16:52)



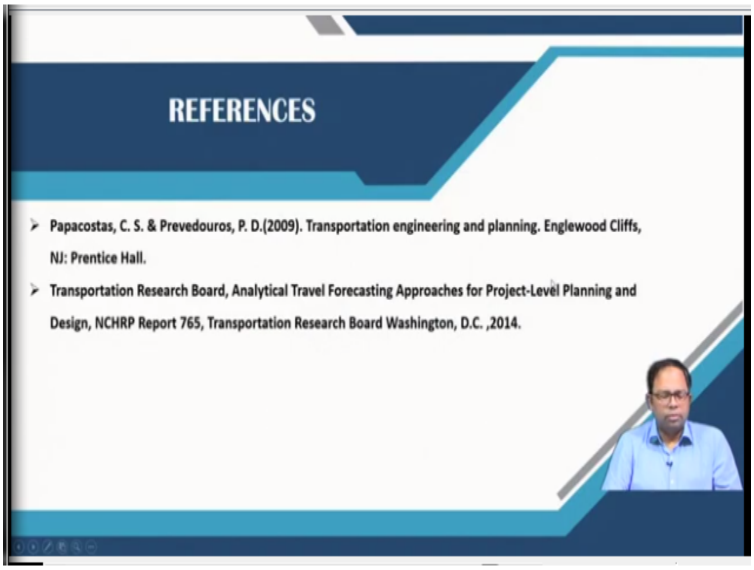
After these arrays are defined, the linear regression model is invoked and a MLR is fitted for trip generation using '**X_gen**' and '**Y_gen**', which are the independent and dependent variables respectively for trip generation. The estimated coefficients and the intercept is stored in '**b_gen**' and '**b0_gen**' respectively. R-squared value can be retrieved using the '**score**' function of the estimated MLR model represented by '**gen**'. Hence the trip generation model is estimated. A similar task is repeated for the trip attraction part and code is written to display the results at the end of operation.

(Refer Slide Time: 18:15)



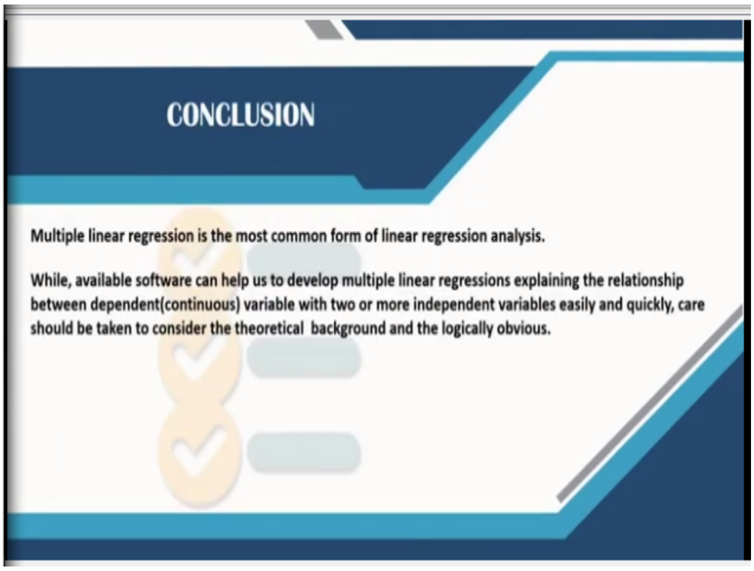
On comparing the results obtained from SPSS and the results obtained from python using scikit-learn, it is found that apart from a few floating point changes, both the results are same. The similarity was exhibited by the trip generation model. The trip attraction model shows some difference in the coefficients. The reason behind it maybe the existence of collinearity in the data. The R-square values of the models are same in both SPSS estimates and python generated estimates.

The model building process remains same in python as it is in SPSS. So it is very easy to write this kind of code in python. The only major difference is the interface of both the programs. Using python for MLR spares one from expending a lot of money on statistical software. **(Refer Slide Time: 20:09)**



And this is how we can do multiple linear regression. So, these are the references that can be looked up for further reading.

(Refer Slide Time: 20:17)



In conclusion, it can be said that, multiple linear regression is the most common form of linear regression analysis. While, available software can help us to develop multiple linear regression explaining the relationship between dependent continuous variables and two or more independent variables easily and quickly, care should be taken to consider the theoretical background and the logically obvious. To estimate the models very fast, we should be very careful about the theoretical background i.e., our a priori knowledge regarding this kind of model is not violated and at the same time the explanatory variables should be logically obvious.