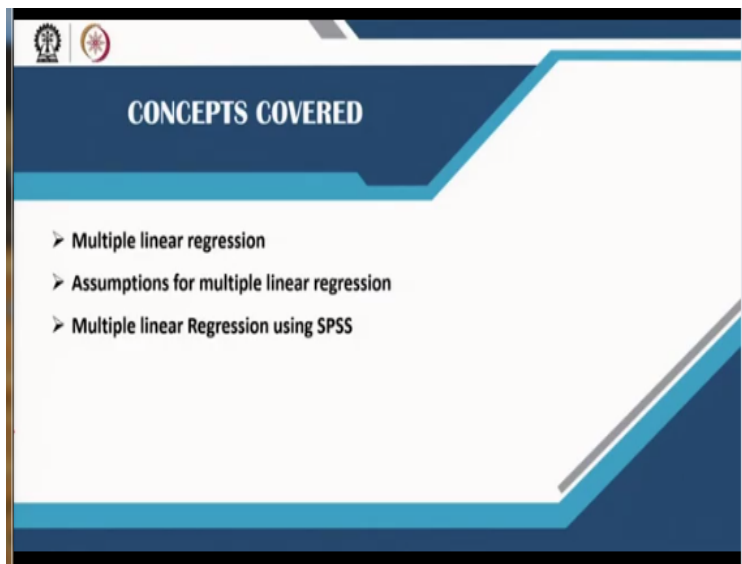


Urban Land use and Transportation Planning
Prof. Debapratim Pandit
Agricultural and Regional Planning Department
Indian Institute of Technology-Kharagpur

Lecture-32
Multiple Linear Regression

In lecture 32 we will cover multiple linear regression. The different concepts that we will cover are: Multiple linear regression, Assumptions for multiple linear regression, and Multiple linear regression using SPSS.

(Refer Slide Time: 00:32)



(Refer Slide Time: 00:40)

Multiple linear regression

The statistical methodology for prediction of a (continuous) dependent variable from multiple predictor variables which may be categorical or continuous in nature.

For,
 y = response/dependent variable and,
 X_1, X_2, \dots, X_k as a set of explanatory variables.

Simple linear regression equation:
 $E(y) = a + b_1x_1$

Out of infinite number of straight lines that can be drawn through a scatter plot between y and x_1 the goal is to find the values of a (intercept) and b (slope) which best fits the data.

Ordinary Least square (OLS) method:
 Minimize the sum of the square of residuals/ error/deviation.

$$S = a_0 \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

NPTEL Online Certification Course
 IIT Kharagpur

Multiple Linear Regression (MLR):

Multiple linear regression (MLR) is a statistical methodology for prediction of a continuous dependent variable from multiple predictor variables which may be categorical or continuous in nature. A simple linear regression equation has only one predictor variable and would look like the equation below.

$$Y = a + b_1 \cdot X_1$$

In case of simple linear regression, for a given set of values of dependent variable **Y**, **a** is the intercept and **b₁** is the slope of the straight line (linear graph), which is an attempt to predict or fit the values of **Y** using a straight line. Fitting a line essentially means finding the best line to explain the scatter plot (data) of **Y**. For example, in the given figure there are two variables, age and size. So an attempt has been made to predict size using age as predictor variable. The line represented by **Size= 0.83 + 0.96 Age** predicts size using age. Here, 0.83 is the intercept and 0.96 is the slope of the line ($\frac{\text{Rise}}{\text{Run}}$).

As it can be seen in the figure, the scatter plot of the dependent variable i.e. size follow a direction but is scattered along it. If the data is to be represented by a straight line, it will touch some of the points of the scatter plot, but will miss many of the points. The difference between each of the observations of **Y** and the corresponding projection on the straight line is called error or residual, and is represented by the following equation.

$$\epsilon_i = Y_i - \hat{Y}_i$$

Where ϵ_i is the residual or error for i^{th} observation, Y_i is the observed value of the dependent variable for i^{th} value of \mathbf{X} and \hat{Y}_i represents the value on the fitted line for the same value of \mathbf{X} . In order to determine the best line that fits the data, the sum of the squared errors (SSE) over each and every observation is minimized. In many texts, SSE is also called sum of squared residuals (SSR).

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \beta_i X_i)^2$$

In other words, out of infinite number of straight lines that could be drawn through a scatter plot between \mathbf{Y} and \mathbf{X}_1 . The goal is to find the values of intercept ' \mathbf{a} ' and slope ' \mathbf{b} ' which results in the least sum of squared errors, alternatively, best fits the data. The process of finding the values for ' \mathbf{a} ' and ' \mathbf{b} ' to get the minimum SSE is called ordinary least square (OLS) estimation.

(Refer Slide Time: 04:19)

Multiple regression equation is an extension of the OLS method:
 $E(y) = a + b_1x_1 + b_2x_2 + \dots + b_px_p$
 $a = E(y)$ when $x_1 = x_2 = \dots = x_p = 0$.
 b_1, b_2, \dots, b_p are called partial regression coefficients.
 There is a linear relationship between $E(y)$ and x_i (while other explanatory variables are controlled).
Assumptions:
 Linear relationship ✓
 Multivariate normality ✓
 No or little multicollinearity ✓
 No auto-correlation ✓
 Homoscedasticity ✓
 Sample size: at least 20 cases per independent variable

Contrary to simple linear regression, MLR has multiple variables and if \mathbf{Y} is the response of the dependent variable and $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$ are a set of explanatory variables, and each of these has n observations, the equation would look like the following equation, where the value of all the partial regression coefficients $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_p$, and the intercept \mathbf{a} , are estimated by OLS estimation using the n observations.

	Dependent	Independent					
	Y	X ₁	X ₂	X _j	X _p
Observations	y ₁	x ₁₁	x ₁₂	x _{1j}	x _{1p}
	y ₂	x ₂₁	x ₂₂	x _{2j}	x _{2p}
	⋮
	y _i	x _{i1}	x _{i2}	x _{ij}	x _{ip}
	⋮
	y _n	x _{n1}	x _{n2}	x _{nj}	x _{np}

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 \dots \dots \dots + b_k \cdot X_p$$

Intercept is the value of Y when X₁, X₂ X_p (or X, in case of simple linear regression) is equal to 0. Each of the variables X₁, X₂ X_p are linearly related to Y while all the other variables are controlled or kept fixed.

There are some assumptions for multiple linear regression:

- Linear relationship
- Multivariate normality
- No multicollinearity
- No auto-correlation
- Homoscedasticity

Adequate sample size is very crucial for a reliable estimate of MLR. It is recommended to have at least 20 observations per independent variable. For example, if a model has 4 variables, so the minimum number of samples for a reliable estimate of partial regression coefficients and the constant is 80.

(Refer Slide Time: 06:47)

Linear Relationship(relationship between the independent and dependent variables has to be linear)

- Scatter plots with individual variables.
- Outliers are important.

Multivariate normality(variables should be multivariate normal)

- Tested with a Histogram, Q-Q-Plot (points lie on a diagonal line when data is normal) or a P-P plot.
- Goodness of fit test to test normality.
- Non-linear transformation (e.g., log-transformation) when not normal.

Multicollinearity(When the independent variables are too highly correlated with each other)

- Correlation matrix (Pearson's Bivariate Correlation among all independent variables).
Absolute values of correlation coefficients need to be smaller than 0.3 to be called weakly correlated. (Source: <https://ibguides.library.kent.edu/SPSS/PearsonCorr>)
- Variance Inflation Factor (VIF)
 - $VIF > 10$; there is multicollinearity
 - Remove independent variables with high VIF values

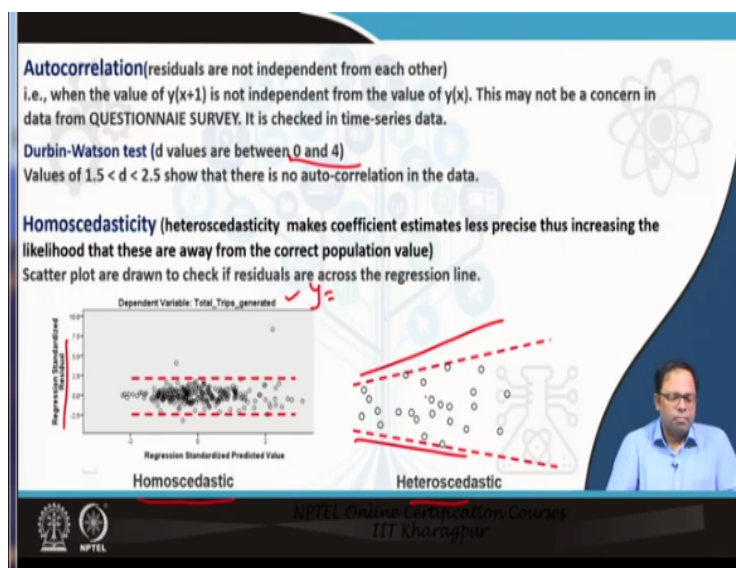
Following is the detailed explanation of each of the assumptions:

Linear relationship: Relationship between the dependent variable and the independent variables is assumed to be linear. This can be checked by a scatter plot of individual independent variables vs. the dependent variable. In the scatter plots, some observations might be located very far from the cluster where majority of the observations or data points lie. Such observations are called outliers. Presence of outliers in the data may influence the model in an undesirable way by skewing the fitted regression line. That means because these points are far away from rest of the points of dataset, these will actually change the orientation of the regression line. So it is better to get rid of the outliers based on certain conditions like, outliers lying beyond the standard deviation of 3, or lying in certain quartile if we divide the data into different quartiles.

Multivariate normality: All the independent variables should be multivariate normally distributed. This could be tested by drawing a histogram of the dataset. Other methods like Q-Q (quantile- quantile) plot and P-P (Probability plot) can also be employed to test is a variable is normally distributed. In both of Q-Q plot and P-P plot, if the points or observations of a variable lie on the diagonal line, the variable is normally distributed. Goodness of fit test is also there to test the normality of a given data. Sometimes a given data may not seem to be normally distributed. In such cases, a non-linear transformation (E.g.: log transformation) can make the data normally distributed and the transformed data can be used in the regression.

No Multicollinearity: Multicollinearity means the high correlation between independent variables. In such cases, a change in one of the variables, results in a change in the correlated variable(s). A regression model should not have multicollinearity and in case of such an instance, only one of the correlated variables should be included in the model. In order to test the correlation between each pair of variables, a matrix is made using Pearson's bivariate correlation. Ideally the correlation coefficients for every pair of variables should be less than 0.3 to be called 'weakly correlated'. If the coefficients are between 0.3 to 0.5, they are moderately correlated, and coefficients higher than that represent high correlation. Another method to determine the presence of multicollinearity is using variance inflation factor (VIF). For any of the independent variables, If the VIF is between 5 to 10, multicollinearity exists. Independent variables with VIF towards 10, should be removed as they are highly correlated.

(Refer Slide Time: 11:06)



No Autocorrelation: Often MLR employs data, which might have dependent residuals. For example, the value of dependent variable at a point $(x+1)$ is also dependent of the value of the same dependent variable at a previous point (x) . This phenomenon is called autocorrelation. Although this is not an issue with survey data, but it might be there in time-series data. Durbin-Watson test is used to check for auto-correlation. Out of the range of 0 to 4 for the value of ' d ', a value between 1.5 and 2.5 suggests no auto-correlation in the data.

Homoscedasticity: A data is said to be homoscedastic if the error or residual is same across all values of the independent variables. MLR assumes the data to be homoscedastic and not heteroscedastic. Heteroscedasticity results in the regression coefficients to be less precise and hence away from the correct population values. To check if a given data is homoscedastic or not, a scatter plot of predicted values and residuals is done. If the points in the scatter plot can be seen to be lying between two parallel lines, then the data is homoscedastic. If the scatter plot lies between two non-parallel lines, heteroscedasticity is said to be there in the data.

(Refer Slide Time: 13:17)

Multiple linear regression with SPSS

Dataset of housing price in California.
Each observation is on a particular block group.
Each block group (600-3000 people)

Data Source: <https://github.com/ageron/handson-ml/tree/master/datasets/housing>

Data cleaning (undertaken before analysis)

Outliers.
Missing data must be dealt with appropriately.
e.g., if the total no. of missing data is less than 10% of the total data, the entries can be deleted.
If it is more, then we might need to use specialized techniques to predict the missing values based on the available complete observations.

The variable names must not have SPACE, SPECIAL CHARACTERS (other than '-') and FIRST LETTER as a NUMERAL.

Total observations: 20435	
Attributes	Datatype
Housing median age	Continuous
Total rooms	Continuous
Total bedrooms	Continuous
Population	Continuous
Median income	Continuous
Median house value	Continuous
Ocean proximity	Categorical

Multiple Linear Regression using IBM SPSS:

Out of the many statistical packages available in the market, IBM SPSS is one of them. In this demonstration, SPSS has been used to carry out MLR using the dataset of 'housing price in California'. The data is freely available on GitHub (<http://github.com/ageron/handson-ml/tree/master/datasets/housing>) and has 20435 observations on seven variables, namely; Housing median age, Total rooms, Total bedrooms, Population, Median income, Median house value, Ocean proximity, in which except for Ocean proximity, all are continuous data and the mentioned variable is categorical data. Each of the 20435 observations is data of a zone on the given variables. Each zone consists of a group of 600 to 3000 households. So each of the 20435 observations of the mentioned variables is the zonal average of the respective variables.

The first thing that is required to be done is data cleaning. In data cleaning, outliers need to be checked for and removed if found. Apart from that, there might also be some missing data due to human error by surveyor or due to the inability of the respondents to provide response for a few variables. Ideally if there are missing values, the whole observation needs to be deleted, but as we go on deleting observations, it might so happen the observations to be deleted is more than 10% of the data collected. In such a case, some techniques need to be employed to predict the values for the missing entries based on the data at hand. For example, for a particular group of values we can take average of similar kind of population group or we can take the average of the entire population, while some other variables might require specialized techniques. If the observations having missing values is less than 10%, the incomplete observations can be deleted.

Nomenclature of variables follows certain rules in SPSS, which otherwise might cause the program to interrupt the analysis. For example, space between the variable names, special characters, numbers at the start of variable name are not allowed. Underscore maybe used to separate words in the name of a variable.

(Refer Slide Time: 16:09)

Dummy variables:
Categorical variables cannot be used in regression directly. They need to be converted to dummy variables.

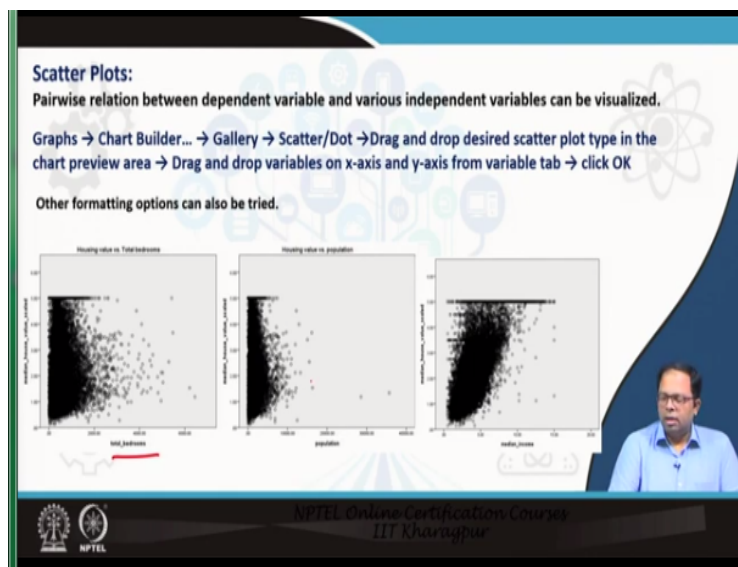
Proximity to Ocean (KEY)		S.No	Proximity to Ocean	Dummy coded data:				
Category	Key			0	1	2	3	4
Island	0	1	Island(0)	1	0	0	0	0
Bay area	1	2	Island(0)	0	1	0	0	0
Near Ocean	2	3	Near Ocean(2)	0	0	1	0	0
1 Hour from ocean	3	4	Bay area (1)	0	0	0	1	0
Inland	4	5	1 Hr. from ocean (3)	0	0	0	0	1

Dummy Variables:

Categorical variables cannot be directly used in regression. They are required to be coded into dummy variables. For example, in the California housing dataset, the data on proximity to ocean has 5 categories; Island, Bay area; Near ocean; 1 hour from ocean; and Inland. They are represented

by the numbers 0,1,2,3, and 4 respectively. For this example, 5 dummy variables are required to be made, each for a category of proximity. So if a house is situated in an island, the column 0 will have the value 1 and rest of the cells in the row will have 0. Similarly, if a house is situated in Bay area, column 2 will have the value 1, and rest of the cells in the row will have 0. If there are n categories, only **(n-1)** dummy variables will be used in the model because if none of the n-1 are having '1' in its value, it is understood that the observation belongs to the omitted class.

(Refer Slide Time: 17:50)



Scatter plots:

Scatter plots are done to visualize the pair wise relation between dependent variable and the various independent variables and scatter plots help us to visualize. The figures shown are for the variables total number of bedrooms, population, and median income, versus the dependent variable i.e. median house value. In order to make scatter plots in SPSS, following sequence of steps needs to be followed:

Graphs → Chart Builder → Gallery → Scatter or Dot →

(Drag and drop desired scatter plot type in the chart preview area) →

(Drag and drop variables on the x – axis and y – axis from the variable tab →

*Click **OK***

(Refer Slide Time: 18:54)

Correlation matrix
(Matrix of pair-wise correlation of variables)

	Area_under_beds	Area_under_beds_sq	Population	Households	Total_rooms	Total_rooms_sq	Area_under_beds_sq	Area_under_beds	Population	Households	Total_rooms	Total_rooms_sq	Area_under_beds_sq	Area_under_beds	Population	Households	Total_rooms	Total_rooms_sq	
Area_under_beds	1.00																		
Area_under_beds_sq	0.89	1.00																	
Population	0.78	0.61	1.00																
Households	0.98	0.97	0.98	1.00															
Total_rooms	0.93	0.87	0.87	0.93	1.00														
Total_rooms_sq	0.86	0.80	0.80	0.86	0.93	1.00													

It is a measure of the strength of relationship between variables. Two highly correlated variables should not be included in a regression model. This results in inconsistency in model interpretation.

Total_bedrooms is highly correlated to population (.878), households(.98) and total_rooms (.93) and the correlation coefficients are statistically significant as well.

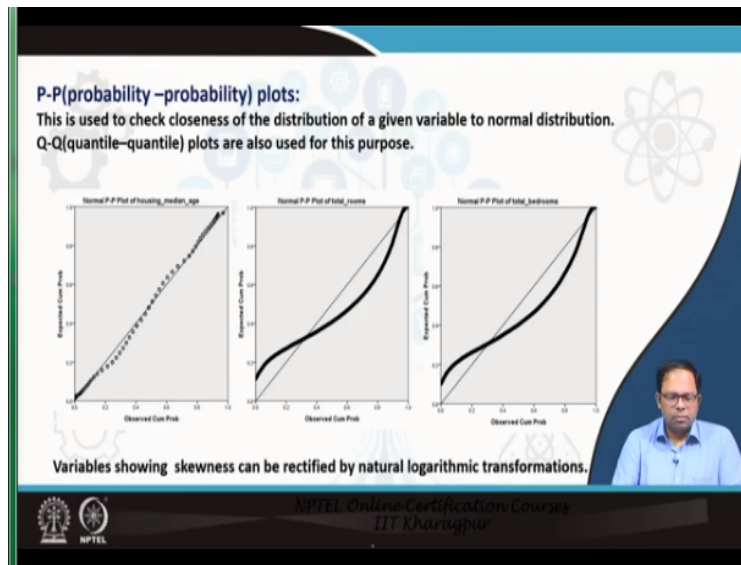
Total_bedroom represents the 3 correlated variables.

NPTEL Online Certification Course
IIT Kharagpur

Correlation matrix:

Correlation matrix is generated for determining pair wise correlation of variables. The matrix is a square matrix of the size equal to the number of independent variables. Each cell in the matrix has the correlation coefficient for the variables in corresponding row and column along with the significance level of the coefficients. Only one of the variables with high and statistically significant correlation should be included in the regression model. For example, total number of bedrooms is correlated with population (0.878), household (0.98) and total rooms (0.93). So, instead of taking all 4 variables in the equation, we can take only one of these in the equation and that would be good enough. So, this matrix helps us to decide which variables to include in the model and a more efficient way to build a model.

(Refer Slide Time: 20:29)



P-P plots:

As already discussed earlier in the assumption for normality, P-P plot is used to check the closeness of the distribution of a given variable to the normal distribution. Q-Q plots can also be used for the same purpose. For a particular variable, the more the data points are near to the diagonal line, closer is the variable to the normal distribution. For variables that show a lot of deviation from the diagonal line, natural logarithmic transformation of those variables may be considered for checking if it follows normal distribution, and the transformed values of the variables can be used in the regression model.

(Refer Slide Time: 21:13)

Steps to perform Linear Regression in SPSS:

Multiple linear regression is found in SPSS in Analyze/Regression/Linear...

Select Dependent variable → Select Independent variables
 Statistics : Make sure Model Fit and Estimates is .

Method:

- ✓ 'Enter' allows all the variables in the model at once.
- ✓ 'Stepwise' starts with one variable and keeps adding one variable after every iteration.
- ✓ 'Forward' includes variables (one at a time) based on their explanatory power.
- ✓ 'Backward' starts with all the variables and removes one at a time based on the poorest explanatory power.

Click OK

Other options available in the dialogue box for Linear Regression can also be tried.

NPTEL Online Certification Course
 IIT Kharagpur

Steps to perform Linear regression in SPSS:

Following are the steps to be followed to perform a linear regression in SPSS:

- From the menu, Click *Analyze* → *Regression* → *Linear* ...
- Choose the dependent variable and independent variables and add them to respective section in the linear regression window.
- **Method** needs to be selected which directs the software on the way variables are included in the model. '**Enter**' method allows all the selected independent variables in the model at a time. After each round of estimation the less significant variables can be removed and the model can be tested further. '**Stepwise**' method starts with one variable and keeps on adding one variable at a time if the addition results in improvement of the results. '**Forward**' method adds one variable at a time based on the greatest explanatory power. '**Backward**' method starts with all the variables and keeps on removing one variable at a time based on the lowest explanatory power.
- After method is selected, OK needs to be clicked for the software to carry out the estimation.

(Refer Slide Time: 22:38)

Model 1 Interpretation:

Model	Unstandardized Coefficients			Standardized Coefficients		Sig.
	B	Std. Error	Beta	1	Sig.	
1	(Constant)	-124.7	.019		-5.409	.000
	housing_median_age ^a	.017	.000	.191	38.613	.000
	median_income	.432	.003	.711	144.082	.000

a. Dependent Variable: median_house_value_scaled

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.714 ^a	.510	.510	80848

a. Predictors: (Constant), median_income, housing_median_age
b. Dependent Variable: median_house_value_scaled

- The model predicts Median House value using independent variables (We start with two).
- Constant (intercept) is the default value of house price when all other variables are 0.
- The Model summary shows the model fit (% of variance explained).
- Adjusted R-Square is penalized for complexity of the model (more variables) and more appropriate.
- The coefficients for each variable is their regression weight and their sign show the direction.
- If F-test is significant, then the model explains a significant amount of the variance.

Each coefficient has a significance (last column) called p-value. Significance of 0.05 means 1-0.05=0.95 or 95% confidence interval. As a rule of thumb, we consider a variable to be significant if its coefficient is significant at or less than 0.05.

NPTEL Online Certification Course
IIT Kharagpur

Model 1 Interpretation:

In model 1, **median house value** is predicted using **median housing age**, and **median income**. The estimation result shows the partial regression coefficients of the two variables and a constant

also. Constant is the intercept that has been explained in the introduction. It represents the median housing value when value of both the variables is zero.

The model summary shows R-square (0.51) which essentially represents the percentage of variance in the dependent variable explained by the model. Due to the way R-square is estimated, it is sensitive to the number of variables in the model. In other words, with the increase in the number of variables in the model, R-square increases. To avoid this ambiguity, adjusted R-square (0.51) is used, which is a penalized version of R-square. In other words, as the number of variables in a model increases or the model complexity increases, the R-square value is penalized to make sure sensitivity to number of variables is taken care of. In the given model, R-square is same as adjusted R square because the included variables are highly significant. F value is another indicator that helps to determine the overall integrity of a particular model. A statistically significant ($p\text{-value} \leq 0.05$) F-test signifies that the model explains a significant amount of variance.

The coefficients for each variable is their regression weight and the sign i.e. positive or negative, corresponding to each of these coefficients indicates the relationship of that variable with the dependent variable. This means if the coefficient of any independent variable is negative, with the increase in the value of the variable, the value of dependent variable decreases and vice versa. In the given model, both housing median age, and median income are positively related to house value.

The last column of the coefficient table shows the p-value or level of significance of each coefficient. Level of significance is used to infer the confidence interval at which the coefficients have been found to be significant. As a rule of thumb, $p\text{-value} \leq 0.05$ or $C.I = (1-p) \times 100\%$ i.e. 95% is taken as significant. There can be instances where variables having theoretical evidence of being important may not be significant at 95% C.I, but maybe significant at 90% C.I. In such cases, we might relax this rule to include such variables in the model, with proper justification.

(Refer Slide Time: 25:14)

Model 2 Interpretation:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta	Std.		
1	(Constant)	.514	.021			24.955	.000
	housing_median_age	.009	.000	.102		21.309	.000
	median_income	.382	.003	.628		134.863	.000
	island	1.847	.328	.025		5.634	.000
	bay	.132	.018	.038		7.525	.000
	near_ocean	.173	.016	.050		10.699	.000
	inland	-.717	.013	-.289		-57.114	.000

a. Dependent Variable: median_house_value_scaled

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.773 ^a	.597	.597	73284

a. Predictors: (Constant), inland, island, housing_median_age, near_ocean, median_income, bay
b. Dependent Variable: median_house_value_scaled

R-Square increases from .510 to .597 as the number of variables are increased in the model.

Proximity to Ocean (KEY)

Island	0'
Bay area	1
Near Ocean	2
1 Hour from ocean	3
Inland	4

The model can be interpreted as :
Cost of house in a block increases with age of houses and income of people in the block.
Additionally, houses on 'island' are most costly followed by houses in 'bay area' and 'near ocean'.
If the house is located 'inland' the cost of house decreases.

NPTEL Online Certification Course
IIT Kharagpur

Model 2 interpretation:

In Model 2, the variables included are the dummy variables representing proximity to ocean, apart from the variables used in model 1. In addition to median age and median income, categorical variable island, bay, near ocean, and inland are included. Adding these variables result in the increase in the percentage of total variance explained i.e. R-square from 0.51 to 0.597. Adjusted R-square also increases and is same as R-square again as the categorical variables are highly significant. The model can be interpreted as; Value or cost of house increases as the median age of house in the zone and the median income of residents in the zone increases. Houses on island are most costly, followed by houses near ocean and in bay area. Cost decreases if a house is in inland areas.

(Refer Slide Time: 26:29)

Model 3 Interpretation:

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients		Sig.	
	B	Std. Error	Beta	t		
1	(Constant)	.316	.024		21.550	.000
	housing_median_age	-.009	.000	-.101	-20.234	.000
	median_income	.302	.003	.628	134.507	.000
	inland	1.647	.320	.025	5.153	.000
	bay	.132	.010	.206	7.523	.000
	near_ocean	.173	.016	.000	10.630	.000
	inland	-.717	.013	-.260	-58.727	.000
	population	-.015927	.000	-.001	-.116	.908

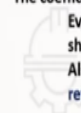
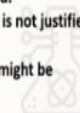






a. Dependent Variable: median_house_value_scaled

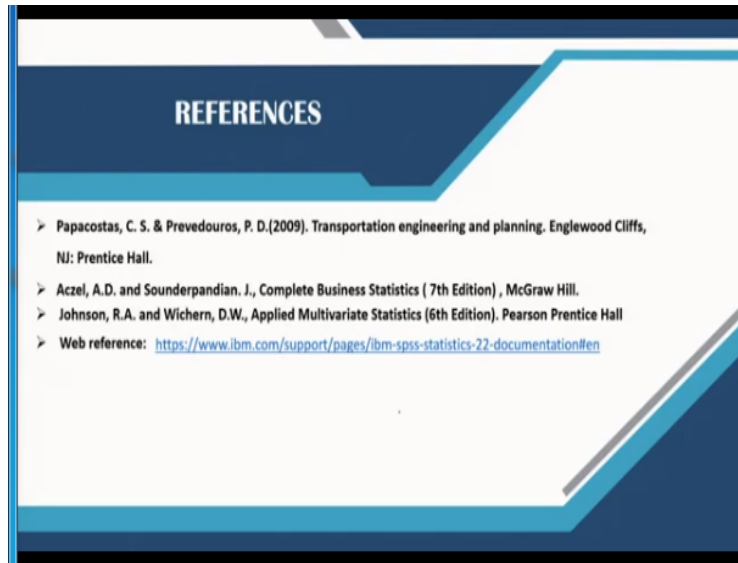
Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.773 ^a	.597	.597	73204

a. Predictors: (Constant), population, median_income, inland, near_ocean, bay, housing_median_age, inland
b. Dependent Variable: median_house_value_scaled

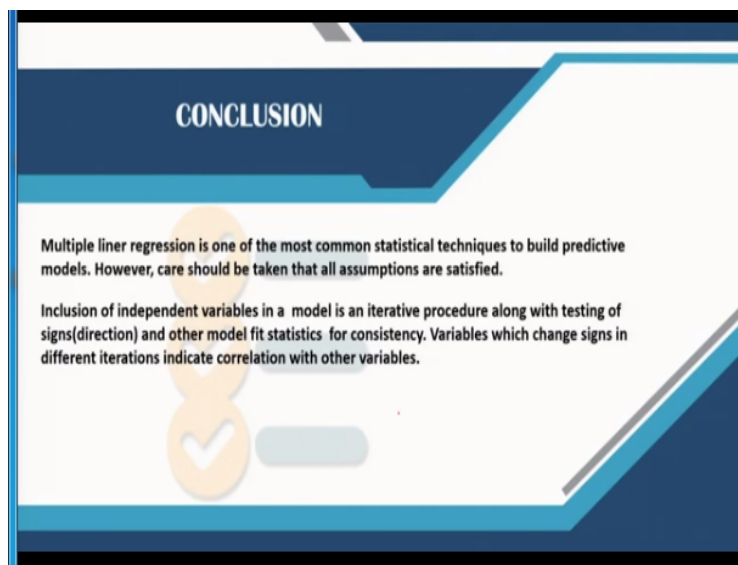
On inclusion of another variable 'population' the overall R-square does not change from the previous model.
 The coefficient is also not significant at 0.05 and thus may be dropped.
 Even if any variable has very less p-value but theoretically it is not justified, it should not be included in the model.
 Also, if any variable is important from policy perspective, it might be retained in spite of having comparatively higher p-value.



So, these are some of the references you can use.

(Refer Slide Time: 27:42)



In conclusion, it can be said that multiple linear regression is one of the most common statistical techniques to build predictive models. However, care should be taken that all assumptions are satisfied. Inclusion of independent variables in a model is an iterative procedure along with testing of signs that is the positive and negative relationship with the dependent variable, and the model fit statistics for consistency. Variables which change signs in different iterations indicate correlation with other variables and they may be required to be omitted from the model.