

Six Sigma
Prof. Jithesh J Thakkar
Department of Industrial and Systems Engineering
National Institute of Technology, Kharagpur

Lecture – 32
Correlation and Regression Analysis

Hello friends, hope you are doing well in your studies and in your profession. And I wish you all the best with the ongoing Six Sigma journey for your present and future endeavor. I hope the concepts you have studied so far must have broaden your horizon and must be in fact helping you in your day to day life as well as the profession. Because the concept of six sigma is quite universal and the principles and its application can be visualised can be seen in all the facets domain of personal and professional life.

So, let us advance in our six sigma journey and to remind you I would like to say that we have completed define measure improve. And now we are discussing various topics with hypothesis testing started in the phase of analyse. So, today we will see the concepts of Correlation and Regression Analysis as a part of analyzed phase and this is our lecture 32.

(Refer Slide Time: 01:33)

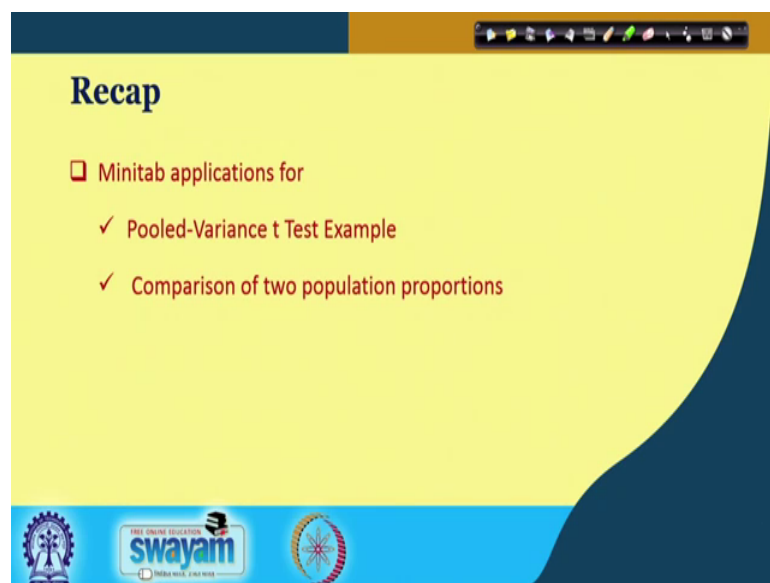


So, I would just like to start with a funny cartoon that many a times you may like to correlate say non related things. And this is where you really need to appreciate that correlation is not equal to causation. So, if you just see this cartoon it says that there is a

growth in sales and there is the growth in moustache as say length. So, moustache length and sales in a way you can correlate.

That yes if there is increase in moustache length of the manager then there is an increase in sales, but you cannot really make any meaningful say outcome or any meaningful insight through such kind of causation. So, correlation is something to be more meaningful about predicting the behaviour of two different variables. And there should be some sense when you are evaluating the two different phenomena.

(Refer Slide Time: 02:39)



So, we had just as a small recap we had the Minitab application as I said which is quite useful for both university student and the managers practicing manager. And we have demonstrated the application of Minitab for two different cases. One is pooled variance t test and another was comparison of two proportion two population proportions.

(Refer Slide Time: 03:07)

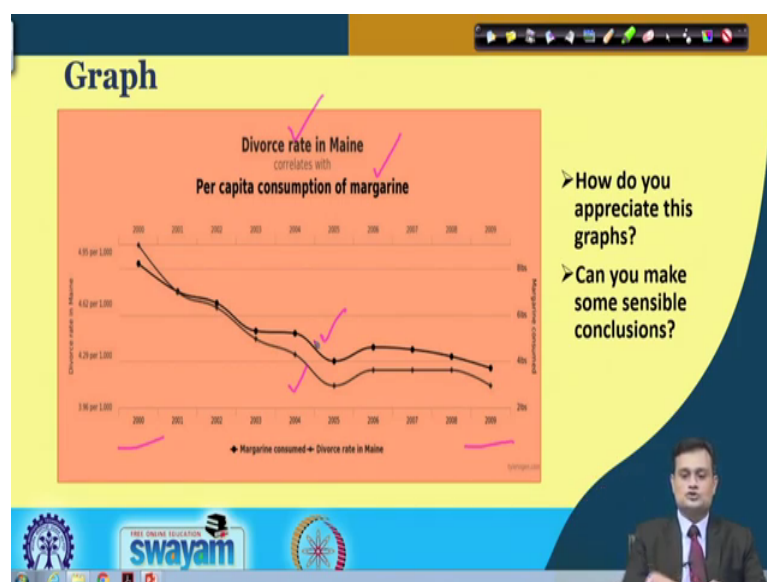
CONCEPTS COVERED

- Correlation analysis
- Regression Analysis: Basics (Predicting the value of a dependent variable based on an independent variables)

swayam

So, this particular lecture will basically focus on correlation analysis and regression analysis. Some basics about predicting the value of a dependent variable based on the independent variable.

(Refer Slide Time: 03:23)

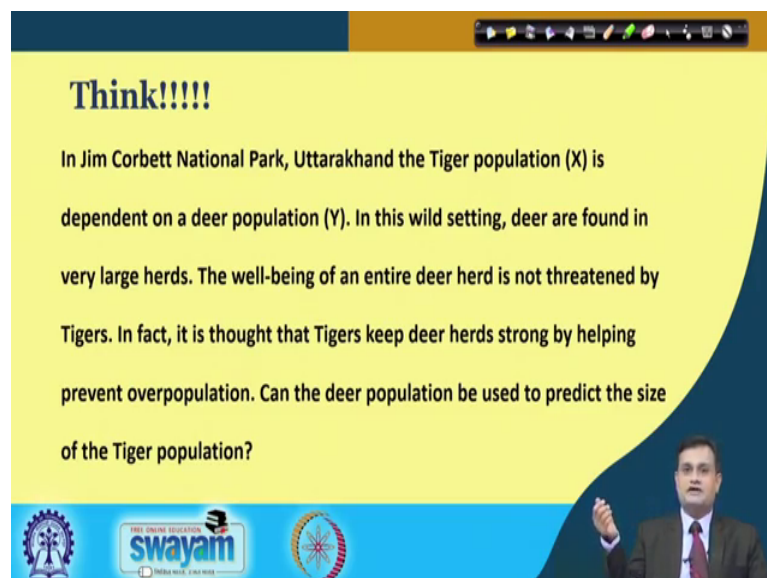


So, just see this and what do you appreciate. I have just borrowed this graph and I am just putting it here does it really make sense. The graph says something like this that there is say a phenomena like diverse rate in Maine. And this correlates with per capita consumption of margarine. So, margarine is a kind of butter that you use to flavour your

dish. And these two phenomena are captured over a period of time maybe from 2000 to 2009 and you have these two graphs resulting.

So, like our moustache example here you can easily say that yes if the consumption of margarine increases then divorce rate will increase. So, two different phenomena which has no relationship with each other if you put it on a plot you may find that there is a relationship. So, this is really not worthy of analysing because these two phenomena they cannot have any real relationship with each other. And we will just be manipulating with the say information in order to draw some wrong false conclusions.

(Refer Slide Time: 05:01)



Think!!!!

In Jim Corbett National Park, Uttarakhand the Tiger population (X) is dependent on a deer population (Y). In this wild setting, deer are found in very large herds. The well-being of an entire deer herd is not threatened by Tigers. In fact, it is thought that Tigers keep deer herds strong by helping prevent overpopulation. Can the deer population be used to predict the size of the Tiger population?

THE OPEN EDUCATION SWAYAM
eGangotri

So, here my point is to say that there should be some relevance. Now I want to put some think kind of situation and the situation is like this. Let us say in India in Jim Corbett National Park, Uttarakhand the Tiger Population X is dependent on a deer population Y. Now in this particular wild setting deer are found in a very large herds and the well being of an entire deer herd is not threatened by tiger's fine. So, that is something which is in between. In fact, it is thought that tigers keep deer herd strong.

So, this is something which is not really say common sense. And there is something interesting to investigate. So, it is believed that tiger keeps the deer heart strong by helping prevent overpopulation and can the deer population be used to predict the size of the tiger population. So, something is happening in the wild setting and this is related

with the population of the tiger population of the deer. And presence of one kind of say animal has an impact on the population of the other kind of animal.

(Refer Slide Time: 06:26)

Think!!!!

- Can you find an equation for the best-fitting line relating X and Y?
- Can you use this relationship to predict the size of the Tiger population when you know the size of the deer population?
- What fractional part of the variability in y can be associated with the variability in X?
- What fractional part of the variability in y is not associated with a corresponding variability in X?

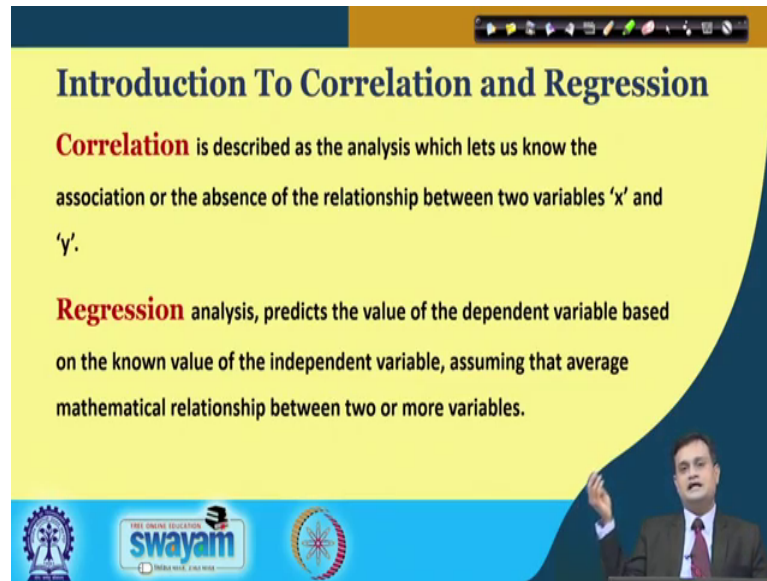
Logos: Swamyam, Free Online Education, and other educational institutions.

Video inset: A man in a suit speaking.

Now, here you may like to investigate can you find an equation for the best fitting line relating X and Y tiger and deer can you use this relationship to predict the size of the tiger population when you know the size of the deer population what fractional part of variability in Y can be associated with the variability in X and what fraction part of the variability in Y is not associated with the corresponding variability in x.

So, you would see that there is a great scope to investigate different kinds of quotients and you may be interested to answer couple of them as a part of this investigation. So, I am just giving this preliminary example and not going into the data analysis to motivate you that when we talk about some kind of prediction when we talk about some kind of relationship between two phenomena typically our correlation and regression analysis then that can really bring some interesting fruitful insights for the organization for the business for the individual.

(Refer Slide Time: 07:44)



Introduction To Correlation and Regression

Correlation is described as the analysis which lets us know the association or the absence of the relationship between two variables 'x' and 'y'.

Regression analysis, predicts the value of the dependent variable based on the known value of the independent variable, assuming that average mathematical relationship between two or more variables.

The slide features a yellow background with a blue wave-like shape on the right side. At the bottom, there are logos for "THE ONLINE EDUCATION swayam" and "INDIA WIDE, 24x7 WIDE", along with a small image of a presenter in a suit.

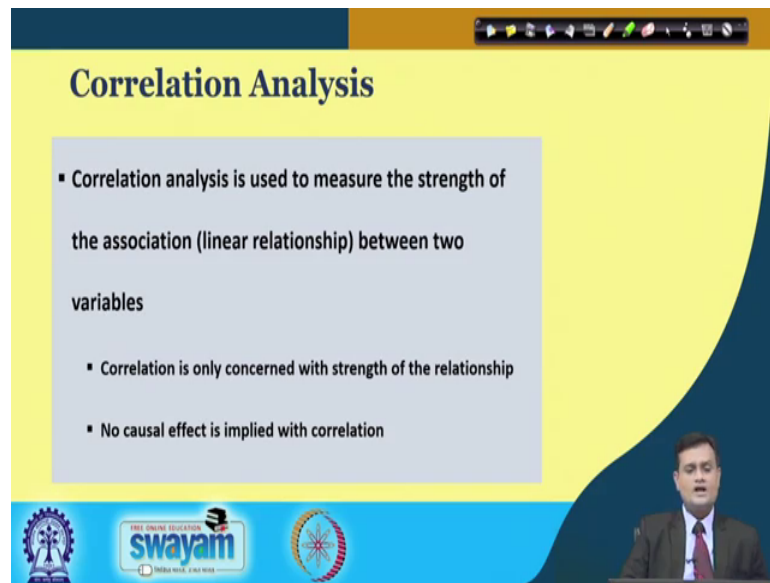
So, what is then correlation and what is then regression that the technical definition if you see then correlation is described as the analysis which let us know the association or the absence of the relationship there could be relationship there could not be. So, if there is relationship then what is that relationship between variable X and variable Y regression.

On the other side predicts the value of the dependent variable based on the known value of the independent variable assuming that the average mathematical relationship between two or more variables is prevailing. So, you have an independent variable for example, you want to check that the what are the you want to check that there are factors which has an impact on the quality defect you items produced in a production setup.

So, here your number of defective items produce is typically a dependent variable, that have the dependence on many other independent factors like training and skill of the worker, the efficiency of the maintenance system the production system capability. We have seen process capability analysis moral of the worker and many other points. Finally, which will result in the number of defectives.

So, you may like to investigate that what are the factors that are really contributing towards the number of defective items produced. And you may like to fit the relationship between the dependent variable that is number of defective. And independent variables like say process capability worker moral skill and other things.

(Refer Slide Time: 09:41)



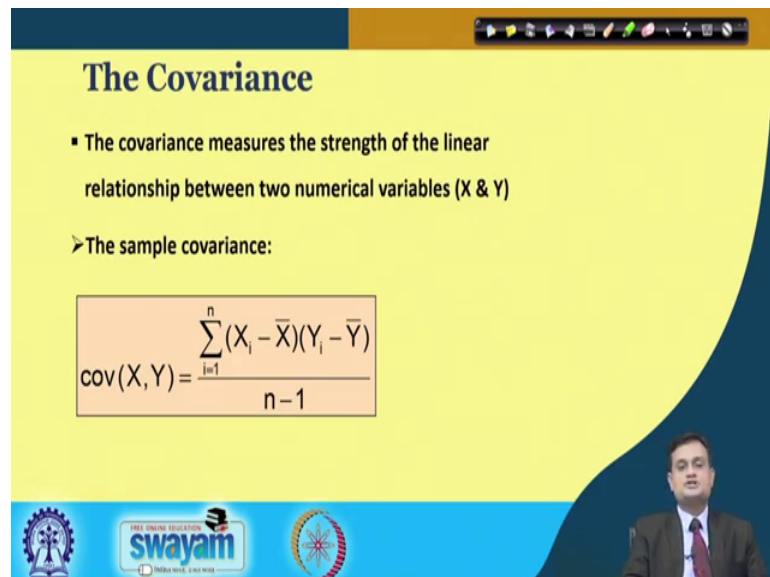
Correlation Analysis

- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables
- Correlation is only concerned with strength of the relationship
- No causal effect is implied with correlation

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there are logos for 'swayam' and other educational institutions, along with a small video inset of a presenter in a suit.

So, a correlation analysis basically say used to measure the strength of the association linear relationship between two variables. So, you have correlation analysis and it is only concerned with the strength of the relationship, no causal effect is implied with the correlation as we had seen in couple of funny examples initially discussed.

(Refer Slide Time: 10:09)



The Covariance

- The covariance measures the strength of the linear relationship between two numerical variables (X & Y)

➤ The sample covariance:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there are logos for 'swayam' and other educational institutions, along with a small video inset of a presenter in a suit.

So, just see that we had discussion on covariance when we talked about the measures of dispersion. And we had seen that there is a measure like covariance so, X and Y. So, how the variable X and Y they vary with respect to each other. So, this could be easily

determined by having the expression $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ in the numerator.

(Refer Slide Time: 10:45)

Interpreting Covariance

➤ **Covariance** between two variables:

- $\text{Cov}(X,Y) > 0 \Rightarrow$ X and Y tend to move in the same direction ✓
- $\text{Cov}(X,Y) < 0 \Rightarrow$ X and Y tend to move in opposite directions ✓
- $\text{Cov}(X,Y) = 0 \Rightarrow$ X and Y are independent ✓

➤ **The covariance has a major flaw:** ✓

- It is not possible to determine the relative strength of the relationship from the size of the covariance

So, we have the conclusions based on correlation sorry covariance analysis like this. That if my covariance is greater than 0 X Y, X and Y tend to move in the same direction. If I have less than 0 they will move in the negative directions it will tell me about the direction nature of relationship. And if it is 0 they are independent there is no point in investigating the relationship. So, this covariance has some major flow and it is not possible to determine the relative strength of the relationship from the size of the covariance.

It will only say that whether they are going in the positive direction or they are changing simultaneously in the negative direction or there is no relationship there are in they are independent. But it is not possible for me to say that variable X and variable Y they are strongly related with each other weakly related or say moderately related. So, I cannot comment on the strength of the relationship.

(Refer Slide Time: 12:04)

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two numerical variables

➤ **Sample coefficient of correlation:**

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y}$$

where

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$
$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

swayam

So, here now; obviously, I would like to investigate the strength of the relationship also and for that you have a measure which is called r coefficient of correlation. So, coefficient of correlation basically uses the value of covariance $X Y$ divided by $S X$ that is sample standard deviation of variable X and $S Y$ sample standard deviation of variable Y . So, you can compute the covariance using the formula we have seen $S X$ because I apply the central limit theorem I will divide it this particular by n minus 1 n minus 1. And you can just plug in the values in this equation and this will give you the value of r .

(Refer Slide Time: 12:49)

Features of the Coefficient of Correlation

- The population coefficient of correlation is referred as p .
- The sample coefficient of correlation is referred to as r .
- Both of them have following features:
 - Unit free
 - Ranges between -1 and 1
 - The closer to -1 , the stronger the negative linear relationship
 - The closer to 1 , the stronger the positive linear relationship
 - The closer to 0 , the weaker the linear relationship

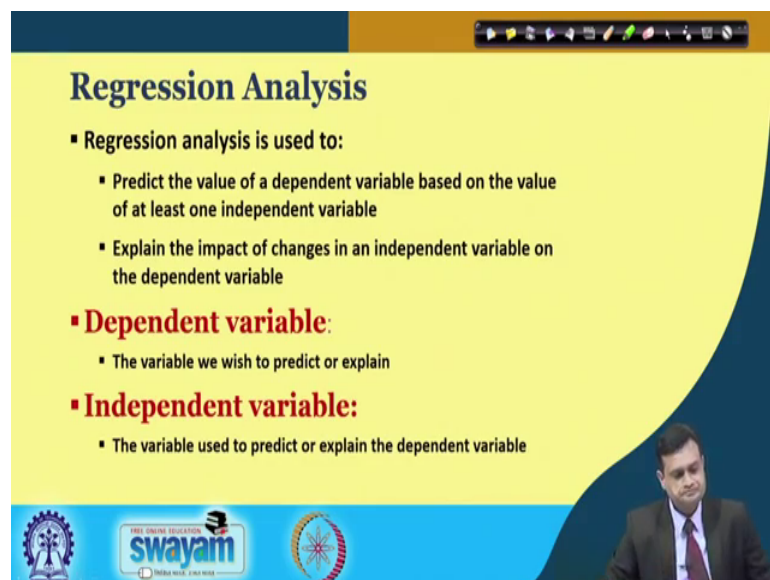
swayam

So, I have the relationship between two variables which is predicted using the r sometimes also it is called ρ . So, you can have for population coefficient of correlation as ρ . And maybe for sample you can use it as r these are the standard nomenclatures. So, you may have the features like this they are unit free and ranges from minus 1 to 1.

And if it is closer to minus one stronger negative relationship closer to 1 then it is stronger positive relationship. And if it is 0 then they work say weaker linear relationship or there is no relationship. So, many a times I would like to analyze such thing if you just recall our qfd discussion. We had created the roof matrix and roof matrix we were trying to correlate the different technical requirements.

Now when I know that there is a relationship and if I know what is the strength of the relationship then when I am making the product or the phenomena which is causing a product to be defective making it defective. Then I can appropriately take the action depending upon the strength of the relationship. So, this has lot of value in making the system sound and robust and taking the appropriate action understanding by understanding the relationship between two different variables or the phenomena.

(Refer Slide Time: 14:35)



Regression Analysis

- Regression analysis is used to:
 - Predict the value of a dependent variable based on the value of at least one independent variable
 - Explain the impact of changes in an independent variable on the dependent variable
- **Dependent variable:**
 - The variable we wish to predict or explain
- **Independent variable:**
 - The variable used to predict or explain the dependent variable

Logos at the bottom: IIT Bombay, Swayam (Free Online Education), and a circular logo with a gear and a person.

So, basically this is what we had seen about the correlation analysis. Now if you see the regression analysis it is about predicting the value then regression analysis is used to predict the value. And typically it explains the impact of change in an independent

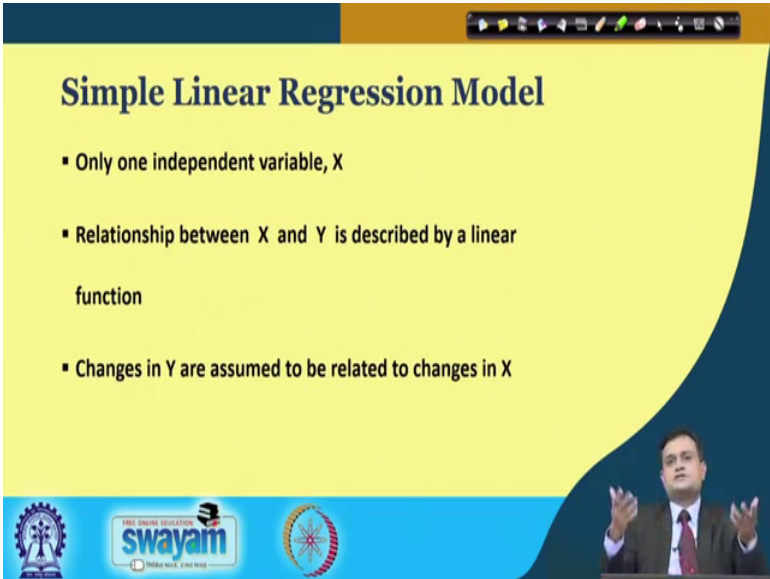
variable on the dependent variable. So, I gave you the example that your production system it is producing some number of defective components.

And this may be because of many independent factors. So, you would like to predict that what would be the contribution of each independent factor because of their coefficient in causing the number of defective items. Likewise you can think about the productivity of the system and many other factors which are contribution. For example, state of the art technology process capability morale of the worker training and skills and many other factors.

So, dependent variable is basically the variable we wish to predict or explain from my example it is very clear an independent variable is the variable used to predict or explain the dependent variable I can take another example say you want to predict the price of a house you want to purchase a house now.

There are many many factors you can think of one is the square feet one is the quality of construction another may be the location closeness to the hospital nearness to the airport and city locality many factors that contribute towards the price. So, price of the house here would be the dependent variable, which depends on many independent factors contributing. Finally, towards the price of the house so, likewise you can just appreciate the difference between dependent and independent variable.

(Refer Slide Time: 16:37)



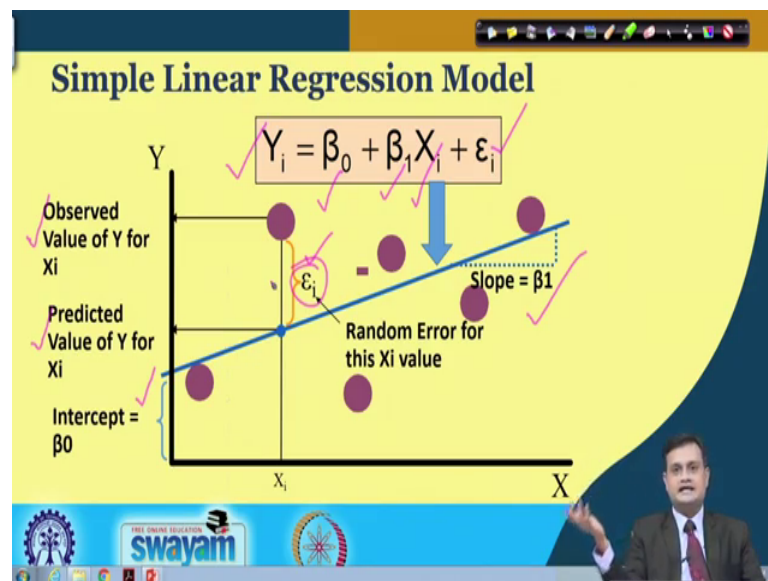
Simple Linear Regression Model

- Only one independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X

swayam
INDIA RISE, CHINA RISE

Now, if I just talked about basics of some simple linear regression model. Then there is only one dependent variable Y sorry independent variable X. And I want to let us say predict the relationship between X and Y which is described by a linear function. And changes in Y assume to be related to change in X. So, this is a very very simple model I am just trying to explain as a basics, but this can even be extended by including more number of independent variable as I gave you the example for price of the house.

(Refer Slide Time: 17:19)



So, this is typically it looks like that you have the Y_i which is your dependent variable. You have β_0 this is the intercept of your regression model you have β_1 this is typically the slope of your regression line. You have the X_i so here the factors as I said location, locality, quality of housing, nearness to airport, closeness to hospital, you can put as many X_i as they are relevant to they say your dependent factor here it is Y_i .

And you have something called epsilon i . In the coming lecture we will see that epsilon i is basically the error which has lot of importance in analyzing my quality of my regression model. So, we will see the validity of regression model in the next lecture. But remember that this epsilon i has lot of value and predicting also the error associated with the regression model can give me lot of insight into the quality of my regression model. So, this is the observed value of X Y for X_i this is the predicted value of Y for a different X_i .

And typically you can see that when I see the difference between observed value and the predicted value. So, I have the actual observation I have the predicted value when I find the difference epsilon, I just think that you are predicting the demand let us say you predicted a demand of 1000 jackets in the month of November.

And suppose when actually you are selling or you are checking the actual demand it is 1200. So, or it is 800 both the weight is true. So, when you take the difference then this will be the error in predicting forecasting typically the number of jackets to be produced for a given season and this is where epsilon i comes in picture.

(Refer Slide Time: 19:44)

Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

$\hat{Y}_i = b_0 + b_1 X_i$

Value of X for observation i

$\beta_0 = b_0$
 $\beta_1 = b_1$

swamyam

So, the model is very simple and interesting that linear regression you have estimated or predicted Y value which is \hat{Y}_i estimated of the rigorous and intercept that is beta 0 or b_0 .

So, please do not get confused beta 0 and b_0 beta 1 and b_1 . I can use interchangeably you have b_1 an estimate of the regression slope and this is my value of the X for observation i. So, this is my simple basic linear regression model which is used to predict the Y value of Y based on the given value of X_i .

(Refer Slide Time: 20:27)

Obtaining b_0 and b_1

➤ **The Least Squares Method**

▪ b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and \hat{Y} :

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Now, here you just try to have little understanding on least square method on your own. And least square method basically I would like to apply here to minimize the error component. And you can see here that error component is explained here as $Y_i - \hat{Y}_i$ whole square. So, I want to minimize the sum of square of error. So, that my regression line can have a best fit when I see that it should pass through the scattered plot it should pass through the data point collected.

So, I want to minimize basically my objective is to minimize and I can just plug in the values. So, Y_i remains Y_i and from the previous one \hat{Y}_i that is the predicted value can be written like this $b_0 + b_1 X_i$ whole square. So, this is my say function which is to be minimized and if I can minimize my error component; obviously, the quality of my regression or base fit regression I can have for the given data set.

(Refer Slide Time: 21:56)

Finding the Least Squares Equation

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

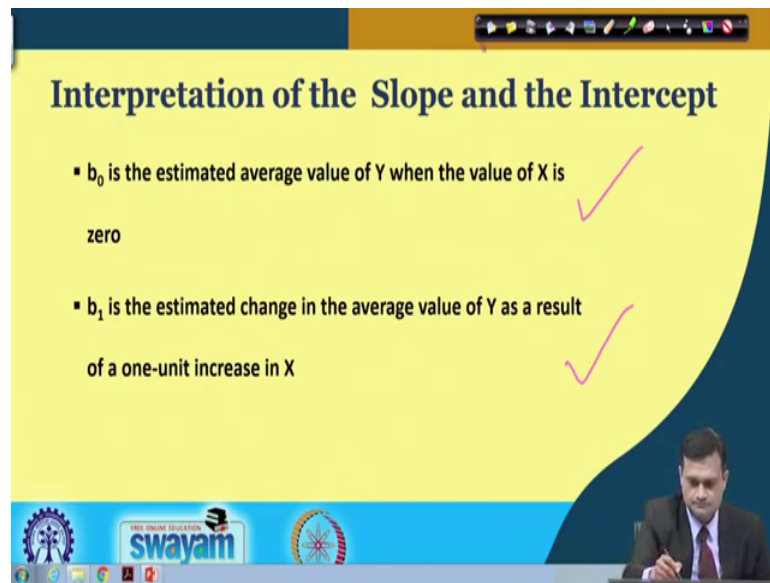
$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$$

Sxx, Sxy and Syy are the sum of squares

So, here I am just trying to minimize this function and what I am doing that you have already studied the calculus. And when you take the differentiation of a particular function equated with the 0. And you try to find the value of constants this process you have already learnt in 12th standard. So, when I do it for a function which is to be minimize, then as a result I will get beta 0 hat or b 0 which is y bar minus beta 1 hat x and beta 1 hat or b 1 these are the constants of my regression model.

So, this would be S xy divided by S xx. So, now you have S xx S xy S yy. All these are basically the sum of squares and I am trying to minimize the sum of square. So, you have S xx which can be easily determined by using this expression sigma i is equal to 1 to n x i square minus sigma i is equal to 1 to n x i whole square divided by n similar way S xy and S yy you can determine.

(Refer Slide Time: 23:15)



Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one-unit increase in X

So, we have the basic equations to find the β_0 β_1 for my regression model. And for that I also have the S_{xx} S_{xy} S_{yy} basically the sum of squares. So, b_0 is typically estimated average value of the Y when the value of X is 0 and β_1 is the estimated change in the average value of Y as a result of one unit increase in x. So, if you increase your X by delta what is that you need change in slope and this is typically reflected by my b_1 or β_1 .

So, now I just want to end this session with some think it. And this lesson basically has provided you some basic idea created interest in learning the correlation and regression analysis. And some basic idea on what is correlation what is regression what we are trying to do in correlation and regression analysis. We will say study some example as well as validation of the regression model in the next lecture. But before we end just let me try to float couple of think it for your introspection and self say revising.

(Refer Slide Time: 24:41)

1. Why we perform Regression analysis?

2. What is the difference between Regression Correlation analysis?

3. What is the significance of b_0 and b_1

UGC swayam Ministry of Education

So, why we perform regression analysis? Similar way you can think why we perform correlation analysis. What is that something which forces us to go for correlation analysis when we can easily find the covariance? What is the difference between regression correlation analysis.

So, this is what I explained and you just try to recollect. What is the significance of b_0 or β_0 and b_1 β_1 and also how this constants are determined using the least square function.

(Refer Slide Time: 25:25)

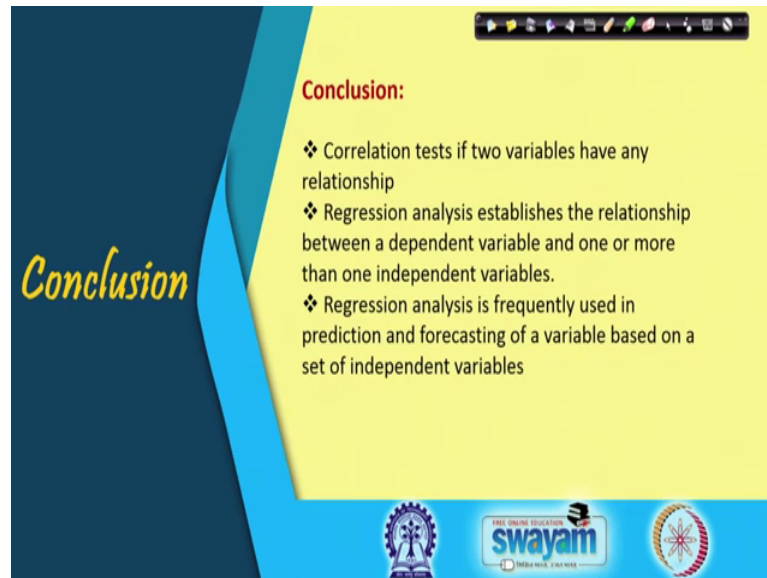
References:

- Aczel, A., Sounderpandian, J. and Saravanan, P. , Complete Business Statistics, McGraw Hill Publication.
- David M. Levine, Timothy C. Krehbiel, Mark L. Berenson and P. K. Vishwanathan, Business Statistics, Pearson Publication.
- T. M. Kubiak, Donald W. Benbow, The Certified Six Sigma Black Belt Handbook, Pearson Publication.
- Forrest W. Breyfogle III, Implementing Six Sigma, John Wiley & Sons, INC.

UGC swayam Ministry of Education

So, this are the references you can use to appreciate the concept of correlation and regression analysis.

(Refer Slide Time: 25:26)



And you can strengthen your understanding. So, correlation basically test if two variables have any relationship. Regression is established to predict the value of independent variable based on the number of as a dependent variable based on the number of independent variables that are contributing towards the dependent variable. So, regression analysis is frequently used in prediction and forecasting of a variable based on a set of independent variables. So, thank you very much for your interest in learning correlation and regression analysis some basics concepts interesting examples.

And I hope this would have strengthen your understanding on fundamentals of the correlation and regression that will help you subsequently in the deeper understanding developing deeper understanding on the model validation regression model and that we will do next time. So, till that time please keep revising think about various examples where you would like to apply the correlation or regression analysis. And just try to justify that to what extent your analysis is really useful to the organization company or a function. And what kind of benefit or result it can give for the improvement in the processes or the function to the organization.

So, once again I would like to remind you that we are discussing basically the analyzed phase of our DMAIC cycle. And as a part of analyzed phase we have gone through the

detailed discussion on hypothesis testing for one population, for two population Minitab application. And now we are doing the correlation regression analysis and we will go further into the details. So, please be with me keep revising enjoy.