

Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-16

Relationship Between Qualitative Variables

(Refer Slide Time: 00:28)



Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So today, we are going to start a discussion on 1 of the major themes in statistics and econometrics namely regression analysis. So, we will start with a brief discussion about covariance and then, we will introduce the concept of Pearson correlation coefficient and then, I will show you a hypothesis testing concerning the Pearson correlation coefficient statistical significance. Then I will introduce the concept of regression, I will talk about very briefly about least squares principle and coefficient of determination.

(Refer Slide Time: 00:56)

Covariance

- ◆ **Covariance** is a measure of the joint variability of two random variables, i.e. it determines the relationship between the movement of two random variables
- ◆ Population Covariance expression:
$$\text{COV}(X,Y) = \sigma_{XY} = E [(X - E(X))(Y - E(Y))]$$
$$= E [XY - X E(Y) - Y E(X) + E(X) E(Y)]$$
$$= E(XY) - E(X) E(Y) - E(Y) E(X) + E(X) E(Y)$$
$$= E(XY) - 2 E(X) E(Y) + E(X) E(Y)$$
$$= E(XY) - E(X) E(Y)$$
- ◆ Formula for calculation of sample covariance:
$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

So, now, we are going to talk about the concept of covariance, what is covariance? So, if you have two quantitative variables say x and y, so, they can vary with time or they can vary from one individual to another individual. So, for different values of x, you see different values are why are there when you see a data set, it can be a time series data or it can be a cross section data.

So, now a question automatically arises in mind. So, is there any association or relationship between these two variables x and y, on which I have a set of data? So, expressed in other words, what if my x changes whether y will change? And if at it changes then if x increases is why is also going to increase? Or y is going to decrease?

So, these questions could be answered through the concept of covariance. So, covariance is basically a measure of the joint variability of two random variables. So, in other words, each determines the relationship between the movement of two random variables, say x and y.

So, now, I am going to show you the mathematical or statistical expression for the covariance concept. So, it is given by this symbols sigma xy. So, otherwise it is also called COV X, Y in textbooks. And you see, this is basically a random variable concept. So here, you have two random variables x and y.

So, first you have to subtract the mean of random variable X and from the values of the variable x . And then, you also have to subtract the expected value or the mean value of y from the random variable y . Then you have to multiply them and as these two are random variables, these deviations from mean, respectively of variables x and y , so, the product is also a random variable. So, there will be probability associated with each of these deviance variable values. So, you multiply these deviance observations with their probability, so that you can get the expected value of the total.

So, the first line of this $\text{COV } X, Y$ expression actually gives you that. So, x times y , so when you multiply two random variables, the product is also a random variable. So, you can definitely take expectation of that variable. So, that is why I can write expectation of x times y . And then, expectation of y is basically once calculated, that is a constant. So, when you are taking expectation of x times y , then you can write expectation of x times expectation of y . So, expectation y will be treated as a constant.

So, remember we have discussed this issue how to handle the transformation of random variables and how to take mean and variance when a variable is transformed, either by location or by scale. So, that same concept you can apply here. So, here, expectation of x times y is there in the third bracket. So, basically that is also a constant when calculated. So, y is the only random variable, so you have to take the expectation of y and e of x will actually come out from that expectation and it will stay as a constant, but still multiplied with that expectation of y .

And last term in the second line is expectation of x times expectation of y . So, it is basically multiplication of two constants. So, expected value of a constant is that constant only. So hence, you get back the same item at the end of the third line, right.

So, then you simplify further to get this expression expectation of $x y$ minus expectation of x times expectation of y . So that is basically the simplified form of the covariance and its population covariance, because we actually are dealing with the population here when we are defining the concept. But there is also a sample analogue of covariance. So, now we are going to look at the expression for sample covariance.

So, now you look at the bottom part of the slide and I am showing you the formula here. So, you have a sample of size n and you have observations on these two variables x and y . So basically, you have to calculate the sample mean \bar{x} and \bar{y} for the variables x and y , respectively.

Then you have to calculate the differences from these sample means and then into multiply. And then, you need to sum these products over n number of data points. And then finally, you divide that he knows some product by $n - 1$ to get the value of the covariance between x and y .

Now, we are going to talk about correlation coefficient, which is also a related concept to covariance. But just wait a minute, before we start discussion about correlation coefficient, what do we learn from these covariance number that we have just calculated by using that sample formula version?

So, by looking at the sign of the number, we can actually infer the nature of the relation or the association between two variables x and y . So, if we get the sample covariance value to be negative, then we can say that there is an inverse relationship between x and y . So, if x increases, then y decreases or if y decreases, then x increases.

And then if the covariance value is positive, then we can say that there is a positive relationship between these two variables, x and y . So, if x increases, y is supposed to increase as well. And if x falls then correspondingly, y will also fall. And if the value is 0, then we can say that they are independent of each other.

So, any movement in x is not reflected in any kind of movement of y or any movement in y , is not reflected in the movements of x . So then, you can say that x and y two variables are independent of each other, if you observed covariance value to be 0.

Now, the problem is that covariance number is not standardized or not normalized. So, when you have a data set and you see the covariance value, and if you get another data set when probably the person is talking about same variables, but maybe in different units, then that covariance value from their data set these two numbers you may not be able to match well.

So, for comparison purpose, the covariance has to be standardized and normalized and correlation coefficient exactly does so.

(Refer Slide Time: 08:14)

Pearson's Correlation Coefficient

- ◆ **Correlation** measures strength of the **linear** association between two quantitative variables
 - Population Corr: $\rho_{XY} = \text{COV}(X, Y) / \sqrt{V(X)V(Y)} = \sigma_{XY} / \sigma_X \sigma_Y$
 - Normalized measure of covariance with range: $-1 \leq \rho_{XY} \leq +1$

Sample Corr. Coefficient

$$\hat{r} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Interesting Property:

- ▶ Pearson correl. coefficient is invariant under separate changes in location and scale in the two variables.
- ▶ $\rho_{XY} = \rho_{a+bX, c+dY}$

So here, in the slide, I am going to show you the formula for Pearson correlation coefficient. So, note that correlation analysis is basically linked to the idea of covariance. So, by covariance, we mean that this tries to measure the direction of movements in two linked variables x and y. And now, well, how these two variables x and y are linked to each other. So, there could be a linear relationship, there could be a non-linear relationship. So, when I talk about correlation, it is just a special kind of relationship between two variables.

So, a correlation analysis actually is executed when you want to test whether there is a kind of linear relationship between two variables x and y or 0. So, basically put in other words, you want to see the strength of linear association or linear relation between two variables x and y. So, as the name says that it was again Karl Pearson, who actually proposed a formula how to find a measure of this associations with linear association between two variables x and y. And we are going to study the formula he proposed.

So, at the population level, when you deal with two random variables x and y, the population correlation coefficient is given by this symbol rho, Greek letter rho. And you

see at the subscript, I am showing you these x and y , two random variable and these can be expressed as this mathematical formula covariance between x and y divided by the standard deviation of x times standard deviation of y .

So, normalized measure of covariance, which is this correlation coefficient always falls between minus 1 and plus 1. So, as per the formula the lower bound of product moment correlation coefficient is minus 1 and plus 1 is the upper bound for the measure. Now, this minus 1 and plus 1 means that the data is talking about a perfect linear relationship between two variables, but the nature of the association or the relationship is different. So, I will request you to wait for maybe a minute, I will show you some diagrams, which will talk about different types of correlation possible in a data set.

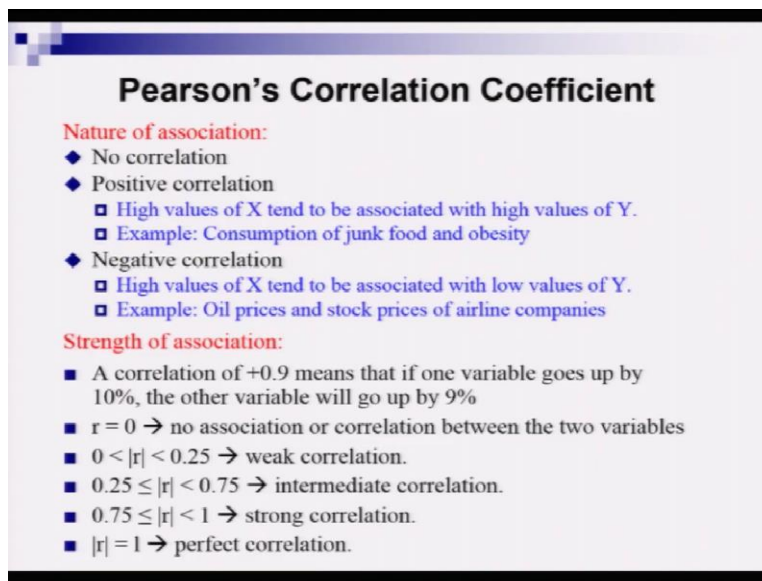
But first, let us have a look at the sample analogue of this correlation coefficient formula. So here, at the bottom part of the slide, I am showing you the sample correlation coefficient formula, which is given as r hat because you need to estimate from a sample you actually do a value of ρ , which is the population parameter.

Now, please concentrate on the numerator let us do numerator and denominator one by one because at one go this looks extremely messy. So, probably this is the most messy expression that I have shown so far in the course. So, let us look at one by one. So, let us concentrate at the numerator first, note that, this is nothing but the sample covariance. And if you now look at the denominator, then you see two square roots. So, if you closely look at the first one or in which is placed first in the denominator that is basically the sample standard deviation.

And the second one in the denominator is the again, another sample standard deviation, but the first one is for the variable x and the second one is for the variable y . So, now, we are going to look at two very interesting properties of Pearson product moment correlation coefficient. So, let us focus on the red box at the right-hand side of the slide. Here, I am talking about two interesting properties. And property number 1 says that the Pearson correlation coefficient, both the population and sample version, they are invariant under separate changes in location and scale in two variables x and y .

So, you define two new variables here say m and n , where m is defined as a plus b times x and n is defined as c plus d times y . So, this capital M and capital N , they two are also random variables because they involve other two random variables x and y . So, A and C are basically the location parameter changes that is happening and the B and D are basically talking about the change in the shape parameter. So, now, basically you see what the Pearson has found or the formula tells us is that the correlation between x and y is equal to the correlation between these newly created variables m and n .

(Refer Slide Time: 13:56)



Pearson's Correlation Coefficient

Nature of association:

- ◆ No correlation
- ◆ Positive correlation
 - High values of X tend to be associated with high values of Y .
 - Example: Consumption of junk food and obesity
- ◆ Negative correlation
 - High values of X tend to be associated with low values of Y .
 - Example: Oil prices and stock prices of airline companies

Strength of association:

- A correlation of $+0.9$ means that if one variable goes up by 10%, the other variable will go up by 9%
- $r = 0 \rightarrow$ no association or correlation between the two variables
- $0 < |r| < 0.25 \rightarrow$ weak correlation.
- $0.25 \leq |r| < 0.75 \rightarrow$ intermediate correlation.
- $0.75 \leq |r| < 1 \rightarrow$ strong correlation.
- $|r| = 1 \rightarrow$ perfect correlation.

So, now, we are going to talk about few more properties of Pearson correlation coefficient, we will talk about the nature and the strength of the linear relationships that one can find from this formula. So, we are going to consider various possible cases. So, in this slide, I will first talk about the nature of the association and I have already given you some idea in this talk only. So, there could be 3 different types of association possible. One is of course, no correlation and that is indicated by the covariance value 0 or the Pearson correlation coefficient value 0.

But there could be two other types and they are actually, mostly found in real life data and that is they are positive correlation and negative correlation. So, what does positive correlation means? So, that means that high values of x tend to be associated with the high values of y and vice versa.

Now, I can talk about the negative correlation. So, that means that high values of x tend to be associated with the low values of y or vice versa. And here, I can give you one example say oil prices and the stock prices of the airline companies. Now, although the correlation and covariance these concepts cannot be used for causality exercise. You cannot actually say that, x is causing y you are y is causing x . But sometimes there is causality embedded, like, let us consider the last example that I spoke about.

So, there is a negative correlation between the oil price and the stock price of airline companies, why? Because oil is the key input for the airline company's business. So, if oil price goes up, that will have a direct impact on the profitability of the airline companies. So, the future projected profits are going to be less for the airline companies and that will have an heavy impact on the stock prices in the share market for this company. So, that is why you see that they are linked. So, sometimes you can think about this causality from outside statistical theory, causality cannot be established by a statistical theory itself.

(Refer Slide Time: 16:09)

Pearson's Correlation Coefficient

Nature of association:

- ◆ No correlation
- ◆ Positive correlation
 - ▣ High values of X tend to be associated with high values of Y .
 - ▣ Example: Consumption of junk food and obesity
- ◆ Negative correlation
 - ▣ High values of X tend to be associated with low values of Y .
 - ▣ Example: Oil prices and stock prices of airline companies

Strength of association:

- A correlation of $+0.9$ means that if one variable goes up by 10%, the other variable will go up by 9%
- $r = 0 \rightarrow$ no association or correlation between the two variables
- $0 < |r| < 0.25 \rightarrow$ weak correlation.
- $0.25 \leq |r| < 0.75 \rightarrow$ intermediate correlation.
- $0.75 \leq |r| < 1 \rightarrow$ strong correlation.
- $|r| = 1 \rightarrow$ perfect correlation.

So, next we are going to talk about strength of association. Now, what do we mean by strength of association? Well, by looking at the sign of the covariance or the correlation coefficient formula, we can say that in which direction y is going to behave, if x is going to behave in a particular manner.

So, it is either going to be no in positive manner or it is going to be negative manner. But how strong this linear relationship is between these two variables, only the sign of covariance or the correlation coefficient is not going to talk about, much about the strength of the relationship that one can observe from the data. So, we compute the score and then, the absolute magnitude of the score, the Pearson correlation coefficient score will tell about the strength of the linear relationship that exists between two variables.

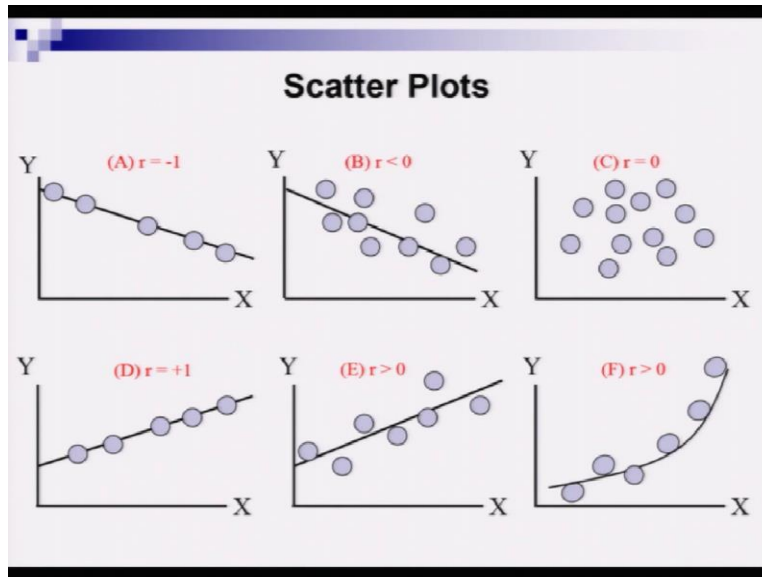
So, I will now talk about the strength of the association and let me spend a couple of seconds on interpreting a particular value of say r that sample correlation coefficient. So, if you find a correlation of say 0.9 then that means that if 1 variable goes up by 10 percent, the other variable is expected to go up by 9 percent.

Now, there are different sort of values possible for r . So, I will start with an extreme case where you find r to be 0. So, basically there is no association or correlation between the two variables. And if you compute the absolute value of r and that is somewhere between 0 and 0.25, then you say that you have a weak correlation between two variables, whether it is a negative or positive relation, that does not matter, because you are talking about the absolute value of r , but the relationship is weak.

Now, if the absolute value of correlation coefficient falls between 0.25 and 0.75, then you can say that the correlation is of intermediate in nature, it is medium you can say, medium level of strength is found from the data. And if the absolute value of r falls between 0.75 and 1 then you can say that, you are seeing strong correlation in the data set. And finally, again, I end with another extreme case possible that is the absolute value of r is equal to 1, so that gives you perfect correlation.

Now, after going through these ranges of possible r values, I am going to show you a slide where I am going to show you an array of diagrams, which will actually explain what you are expected to see in plots or graphs, when there are different possible values of r .

(Refer Slide Time: 19:23)



So, here in this slide, concentrate on the diagram A and here, I am showing you the case of r equal to minus 1. So here, you see you have 5 data points. So, all these circles are basically data points and these are paired data of course. So, for one individual you observe say $x_1 y_1$; for individual 2, you have x_2, y_2 or if you are dealing with a time series data. So then basically it is, it can be say generally as x_t, y_t where the t subscript will come for the time element or time unit.

And you see that all these data points that you have in your data set, they actually fall on a straight line. If the straight line is downward sloping. So hence, you see that r is minus 1. So, r equal to minus 1 says that it is a perfect relationship between 2 variables, but the relationship is negative in nature. So, basically, they lie on a straight line, because they are perfectly linearly related.

Now, I move on to the diagram panel B, and you see that there is a scatter of data points and some 2 data points are lying on the downward sloping straight line. But there are 5, 6 more, which are not lying on the negatively sloped straight line, and they are just here and they are in the, but there is overall some kind of a feeling that there is some negative relation available in the data between y and x .

And that is why you also find the computed value r to be less than 0. So, you say that okay, there is negative relationship between two variables, but they are not perfectly negatively related. Now, we move on to case C, there you see that the scatter diagram or the data points in the x and y plane, they are not showing any kind of pattern and that is why there is no association or linear association between x and y and possibly for this kind of case, you will get a value of r equal to 0.

Now, I move on to the panel D, where you see that, I have 5 data points and all of these data points are lying on a positively sloped straight line. So, in this case, you can say that x and y they are perfectly linearly associated or related and as they lie on one single straight line, the value of r is expected to be plus 1.

Move on to the panel E, where you see another case where you see that there is overall positive relationship between y and x as found from the scatter plot. But the data points are not lying on the fitted line that you see this black line, but as there is a positive overall relationship between two variables, you are expected to see a value of r to be positive here.

And now, the last in the slide the panel Fiscal, here, you see that two variables y and x are showing you some kind of a non-linear trend in the data. But although the relationship is mostly nonlinear between x and y , still if you try to compute r value you are expected to see a value of r being positive.

Now, we are going to talk about a very important concept and that is the hypothesis testing related to the idea of correlation coefficient. Well, I have started by differentiating between population correlation coefficient and the sample correlation coefficient, right.

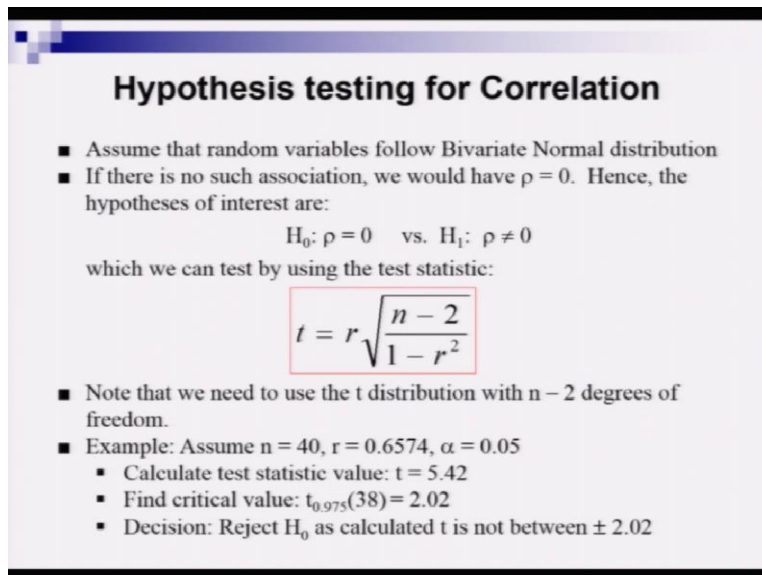
And population correlation coefficient parameter ρ is unknown. So, what we do we just collect a small sample from the universe and then, we try to make a guess about the population parameter ρ . As usual, we have done this thing, we have discussed this thing in the statistical inference part of the course.

So here, if you take one sample you are expected to get one particular value of r hat, the sample analogue of the population correlation coefficient ρ . And then, if you take another sample, it will give you a different value of ρ , ρ , r hat not ρ , r hat.

So then basically, if you keep on generating samples from the population and calculating different values of r hat then you will find a distribution finally. And that is basically the sampling distribution for r hat, right.

So, from that sampling distribution, what can you infer about the population parameter ρ ? So, that is basically the subject matter of hypothesis testing. Here, the target is to statistically check whether in the population at the population level two random variables x and y are linearly related or not.

(Refer Slide Time: 24:49)



Hypothesis testing for Correlation

- Assume that random variables follow Bivariate Normal distribution
- If there is no such association, we would have $\rho = 0$. Hence, the hypotheses of interest are:
 $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$
which we can test by using the test statistic:
$$t = r \sqrt{\frac{n-2}{1-r^2}}$$
- Note that we need to use the t distribution with $n-2$ degrees of freedom.
- Example: Assume $n = 40$, $r = 0.6574$, $\alpha = 0.05$
 - Calculate test statistic value: $t = 5.42$
 - Find critical value: $t_{0.975}(38) = 2.02$
 - Decision: Reject H_0 as calculated t is not between ± 2.02

So, we start with an assumption that two random variables x and y , they follow bivariate normal distribution and I also assume that if there is no association between these two variables, then we expect ρ to be equal to 0. Of course, then you write your null hypothesis as ρ equal to 0. And suppose you are interested to check whether it is not equal to 0 because a priori, you do not know whether ρ is to be positive or ρ is to be negative. So, to be on the safer side, you just set your alternative hypothesis as ρ not equal to 0, so that will lead to a two-tailed test.

So now the next step should be to define your test statistic. So, here in this case, the test statistic formula I am showing you here in this red box and note that that this particular test statistic t follows a t distribution with n minus 2 degrees of freedom. Now, why you have to deduct 2 from n the sample size? It is because, if you remember the formula of the sample correlation coefficient r that there we make use of \bar{x} and \bar{y} . So, these are the sample means for variables x and y . So, as you are computing \bar{x} and \bar{y} to finally get the expression working, you have to deduct these two data points from the data set.

So here, final degrees of freedom will be n minus 2, okay. So here, I am going to show you the working of these test in terms of an example. So, here I assume a large sample you need, not because you are dealing with t here. So, even if you have a sample size of 20, you can perfectly apply this test.

But let us assume n equal to 40 and then, somehow a data set tells me the value of r is 0.6574. Suppose, I have calculated it beforehand and then, I set α at the standard norm 0.05, 5 percent level of significance. So, I have to calculate the test statistic value first and the t calculated value is 5.4.

So now, if I want to follow a traditional method of hypothesis testing, then I have to find the critical values, right? So, as it is two-tailed test, α needs to be split into two parts, in two equal parts, one for the right-tail and one for the left-tail. So, you actually divide α by 2 and so that you save the probability density or mass 0.025 at two different tails.

So, you have to now locate two different critical values from that t table. But t is an amazing distribution, because it is a symmetric distribution. So, you need to just find one critical value, the other critical value will be exactly mirror image of the first critical value.

So, here for the degrees of freedom 38 you find the critical value for the upper-tail $t_{0.975}$ and that is 2.02. So, now you set your decision rule and what to do. So, as our test

statistic value calculated t is not falling in between this range plus or minus 2.02 you reject the null hypothesis.

So, now, I am going to start talking about regression. So, this is the first time we are going to look at a regression at some, elementary level in the course. I have very briefly introduced the concept of regression, when I introduced the course. But now, let me differentiate between the correlation and regression because we are going to talk about these two concepts in the same lecture.

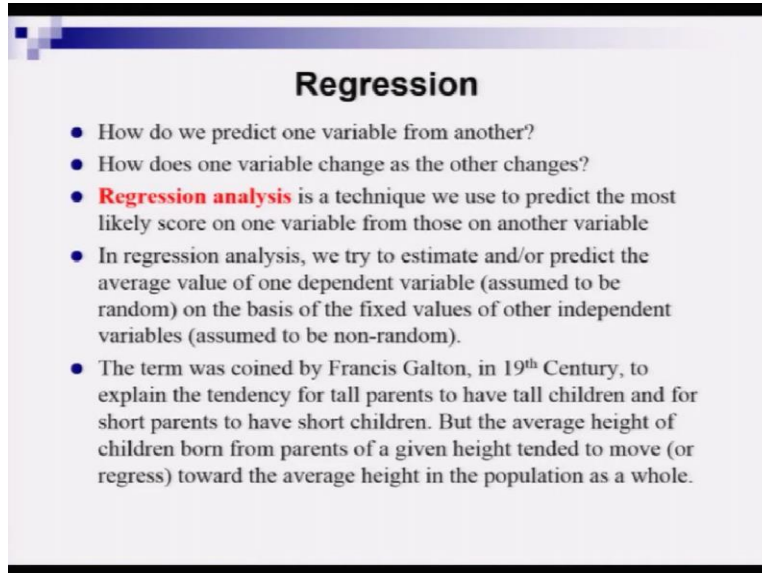
So, in a simple layman's term, why we need two different techniques or why we need two different terms? So, correlation probably by now, you have understood that it is going to measure the strength of linear relationships between two variables x and y . But suppose, I want to create some kind of a model and I want to have some kind of a mathematical or statistical model so, that I can predict the values of y for certain given values of x in future.

So, can I do that by using the concepts of covariance and correlation? The answer is no. And note that, this simple formula of correlation, coefficient and covariance that I have shown that is no simple to collect when you are dealing with only two variables. But what if I have, a model in mind where I have a dependent variable yes, but dependent variable y depends on a set of other variables and there could be five explanatory variables, there could be 15 explanatory variables. So, once you have a complicated model that you want to estimate for decision making, then correlation and covariance will not take you farther.

So, you need a different tool from where actually you can estimate some kind of a mathematical equation, which will help you to predict values for y for given values of x . And regression analyst actually does that. So, in a nutshell in correlation when you are computing the correlation coefficient, you treat two variables y and x as equally. So, there is no distinction between dependent variable or response and the x variable independent or exogenous.

But here, in the regression analyses, you will see that we will constantly talk about these kinds of terms like dependent variable. independent variables. So, that is the major distinction between correlation and regression.

(Refer Slide Time: 31:16)



Regression

- How do we predict one variable from another?
- How does one variable change as the other changes?
- **Regression analysis** is a technique we use to predict the most likely score on one variable from those on another variable
- In regression analysis, we try to estimate and/or predict the average value of one dependent variable (assumed to be random) on the basis of the fixed values of other independent variables (assumed to be non-random).
- The term was coined by Francis Galton, in 19th Century, to explain the tendency for tall parents to have tall children and for short parents to have short children. But the average height of children born from parents of a given height tended to move (or regress) toward the average height in the population as a whole.

So, I have already covered bullet points one to four when I was talking to you. So now I am going to talk about the bullet point five. So, everything has a start and where this regression business started. So, again, I am going back to history stuff. And there was a biostatistician named Sir Francis Galton in UK in the 19th century. And actually, it is him who invented or introduced this term regression analysis.

And he was dealing with some demography data. And accidentally he found out a very interesting feature from the data. And what did he find? So, he had observed the height of individuals and for each individual, he also had data on his or her parents height. So, he observed that on an average, the taller parents used to have taller kids and the shorter parents on an average used to have short height kids.

But what he noted from the data is that that average height of the children born from parents of a given height tended to move towards the average height of the population as a whole. So, what does that mean? It means that the average height of children from taller

parents used to be of less height compared to the average height of the parents and of course, that is going downward, so, that is basically regressing.

So, basically, there is some kind of a grand mean or the grand population average for the height and everybody is basically converging towards that average height. So that is what Galton found. So, here, I am not going to cover more history stuff. So, let us talk about the crux of regression analysis, linear regression analysis, because we are not going to talk about how to capture non linearity through regression models as of now. So, focus on the linear relationships.

(Refer Slide Time: 33:36)

Fitting a Regression Equation with data

- Fitting a straight line to a set of paired observations: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- Regression equation: $y = a_0 + a_1x + e$ where
 - a_1 : slope parameter
 - a_0 : intercept parameter
 - e : error or residual
- How to find a_0 and a_1 so that the error would be minimum?

$$\min S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1x_i)^2$$
- Ordinary Least Squares (OLS) estimator for regression parameters by solving F.O.C. equations

$$\begin{cases} \frac{\partial S_r}{\partial a_0} = 0 \\ \frac{\partial S_r}{\partial a_1} = 0 \end{cases}$$

So here, we are concerned with fitting a straight line to a set of paired observations $x_1, y_1, x_2, y_2, \dots, x_n, y_n$. So, what does that mean? So, I have n data points in my sample and I have data on two variables x and y , these are paired observations, because for one individual one or individual two, I observe both the values of x and y , okay?

So, now, I say that, I have some idea from the theory or I can make some rough hypothesis that my regression equation should be of particular form y equal to a_0 plus a_1 times x plus ϵ or E . So, here y is my dependent variable. So, that depends on the values of x , x is independent variable or explanatory variable. And let me also introduce the other notations in the equation and a_1 is basically the slope parameter, a

naught is the intercept parameter. So, if you for a moment ignore the term e that equation that you are seeing above actually is a straight-line equation that you are familiar with from your class 10 level or 12 level maths.

So, the only interesting part that is added here is that e . So, that is the error or the residual. Why do we need an error or residual? We require this additional term residual or error in the regression model because how do you know or how do I know that the variation in y will be solely explained by the variation in the x .

So, if y is changing, it may be partly because of no a change in x , but there could be some other variable z or w that we are not capturing in this linear regression model and these w and x can also have some impact on y . So, when you see that well x is not changing, but y is changing, maybe that change in y is for a change in w and z . So, you never know. So, that is why as you are not keeping all possible explanatory variables in your model to explain the variation in y , you keep these variable e which is the error variable.

Now, the challenge is to find the a_0 and a_1 and these are the population parameters. So, when we write a regression equation, we are talking about a population relation. So, as usual we do not know the values of a_0 and a_1 's we have to find. How to find a_0 and a_1 ? So, here is a method called ordinary least squares or in other books sometimes it is called least squares principles also.

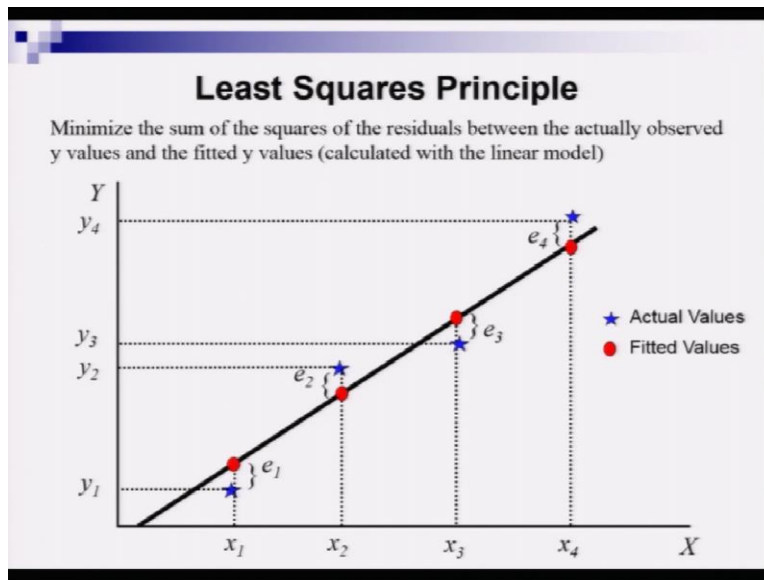
So, you basically concentrate on the error. So, the error is defined as the difference between the observed value y_i and the fitted value that is basically coming from the regression equation. So, suppose, somehow you have an idea about a_0 and a_1 and you have some proxy values for a_0 and a_1 . So, if you plug a no the values of a_0 and a_1 , then for each value of x , actually you can predict the value of y . So, that is the predicted or the fitted value coming from the regression equation.

So, you actually have two data sets on y , right? So, actually you have one y that is observed and one on the other y that is coming from a regression model, or it is the fitted or the predicted value from the model. So, you take the difference and some will be positive some will be negative. So, you square it so, that you can deal with only the

positive numbers. So, the idea is to minimize the sum of error squares. And as it is a minimization problem know for any kind of mathematical optimization, you have to solve by looking at the first order conditions only.

So, you take derivative, partial derivative with respect to the unknown population parameters β_0 and β_1 because these are the variables. So, β_0 and β_1 actually can take many values, so that is why it is the variable. So, you take the differentiation with respect to these variables and you take differentiation of the error sum of squares and set them equal to 0. So, as you have two unknown population parameters, you write two first order conditions and these two first order conditions I am showing you here in the red box.

(Refer Slide Time: 38:24)



Now, in the following slide, I am going to show you this idea of least squares principle through a simple graph. So, now, you look at this diagram in this slide. So here, the dependent variable y is measured along the vertical axis and the independent or the explanatory variable x is measured along the x axis and I am showing you the observed data set, suppose we have only four data points, simple case. So basically, there are paired observations $x_1, y_1, x_2, y_2, x_3, y_3,$ and $x_4, y_4,$ and you see these actual values are now given are shown in this quadrant by the blue colored asterisk.

And now, the idea is that you have to somehow fit a straight line that will pass through these scatter of four asterisk points. But you have to be very clever while you are fitting this straight line, so that you can minimize the distance between the fitted line and the observed data points. So, here how to do it? So, that is what least squares principle actually tells us. So, it suggests that okay, you minimize the sum of the squares of the residuals between the actual observed y values and the fitted y values.

So here, you see many straight lines are possible, which you can draw that will pass through the scatter of four asterisk points. Now, which one to choose? So, here suppose you choose one particular straight line and that is shown here and then, that will give you some values for the intercept parameter and the slope parameter, right.

So, then from this equation, this hypothetical equation if you plug the values of different x 's, x_1, x_2, x_3, x_4 , then you will get what? You will get the predicted or the fitted values. So, for x_1 you will get y_1 hat and for x_2 , you will get y_2 hat and so on, so forth. So, these are the fitted values and these fitted values I am showing here as the red circles that all lie on the fitted black line.

So, that is basically the fitted value and you see the difference. So, you fix one particular value of x and then you see that there is a gap between the red circle and the blue star and this gap is basically your error or residual that is coming from fitting the straight line, because your fit is not perfect as your actual values are not lying on the fitted straight line.

(Refer Slide Time: 41:37)

Fitting a Regression Equation with data

$$\min S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2$$

↓

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i) = 0$$
$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i] = 0$$

→

$$n a_0 + (\sum x_i) a_1 = \sum y_i$$
$$\sum y_i x_i = \sum a_0 x_i + \sum a_1 x_i^2$$

↓

► **Note 1:** Slope coefficient can be expressed as $\text{Cov}(x, y)/V(x)$

► **Note 2:** Regression line always goes through the centre of the data (\bar{x}, \bar{y})

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$
$$a_0 = \bar{y} - a_1 \bar{x}$$

So, we have spoken about the first order conditions, right. So, now, let us see how we can get the proxies the best fit estimates for a naught and a1 by solving this mathematical optimization problem, we have a looked at the diagram, but from the diagram now, let us look at some math.

So, you start from the red box at the top of the slide. So, you minimize this sum problem, so you take partial derivatives with respect to the unknown population parameters a naught and a1. So, you get some expression and then, if you simplify then you get another two set of equations. So, these two set of equations are called the normal equations and they involve two unknown variables that need to be solved.

So, finally, you can solve and then, the last red box in the slide actually shows you a formula, which actually talks about the values or the best fitted or the best proxy values for the unknown population parameters a1 and a naught. So, now concentrate on the box that is placed in the southwest corner of the slide.

So, here, note one says that the slope coefficient that is basically a1 this clumsy formula that you are seeing in the adjacent box that can be actually simplified and that can be expressed in a very simple mathematical expression; and that is basically the ratio of covariance between x and y and the variance of x.

And the second note actually corresponds to the estimate for the unknown intercept parameter a_0 . And by looking at the expression for the estimated intercept parameter, you can very well see that this regression line the fitted regression line always goes through the center of the data, you may not have one particular observation corresponding to the center of the data. But if you calculate the sample mean for your variables x and y , and if you plot that in the quadrant, then the fitted regression line shall pass through that point.

Now, we are going to very briefly mentioned about the goodness of fit criteria for our model. So, once you fit a particular model to our data set, naturally, one question comes to mind that, okay, how good my model is? Can I take my model for future predictions? And you should have some measure by looking at which you can decide whether the model is trustworthy or not? Whether the regression model is doing a good job or not? So, and that is basically done by consulting concept called coefficient of determination.

(Refer Slide Time: 44:20)

Postestimation: Goodness of fit


- Total sum of the squares around the mean for the dependent variable Y is

$$S_t = \sum (y_i - \bar{y})^2$$
- Sum of the squares of residuals around the regression line is

$$S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2$$
- Measure $S_t - S_r$ quantifies error reduction due to describing data in terms of a fitted straight line rather than as an average value.
- **Coefficient of determination** ($0 \leq r^2 \leq 1$) measures the proportion or percentage of the total variation in Y explained by the regression model

$$r^2 = \frac{S_t - S_r}{S_t}$$

Variance explained by X


- For perfect fit, $S_r = 0$ and $r = r^2 = 1$

So, we start with the total sum of squares around the mean for the variable y , the dependent variable and that formula I am showing you here in the first box. So, you see it is looking more or less like the variance of that variable, but the difference is that we are not dividing this some of deviance squares by the degrees of freedom. So, that is the difference.

Next, we talk about the sum of squares of residuals around the regression line and that you have already seen because that is the expression that we minimize to get our best estimates for the unknown population slope parameter and intercept parameter. And now, you measure the difference between st and sr that are shown in these two boxes. And from these difference actually, we can get the coefficient of determination. So, that is defined as st minus sr and this difference now has to be divided by st .

So, this is basically expressed with a notation called r square, why it is called r square? because if you remember that r symbol was used to denote the sample correlation coefficient between two random variables x and y . So, this r square that you get by applying this formula in this box actually is the square of the Pearson product moment correlation coefficient between these two random variables x and y .

So here, of course by looking at this expression, one can say that, this coefficient of determination is going to be always bounded between 0 and 1. So, r can be negative, but r square cannot be. So, r square is bounded by 0 and 1. So, for a perfect fit, of course, sr has to be 0, the sum of squares of the residual terms has to be 0 and only in that case r and r square both will take value 1 from the formula that I am showing you here and this diagram at the southeast corner of the slide is also going to give you a pictorial depiction of the coefficient of determination.

So, note that, we are talking about variation in 3 things. So, one is basically the variation in y and there is variation in the x itself. And finally, when you fit the regression, you generate the variable error the observations for the error variable and then, there is variation in that also.

So, a good model will have less proportion of variability in y , it will, the most of the variability in y should be captured by the variability in x itself. So here, you see that, here I am showing you two circles in two different colors. So, the variance in y that is explained by x is given by the blue circle; and variance, which is not explained by x that is basically the error variance, that is given by the red circle.

So, actually what you want? You want this blue circle encroaching the area of red circle as much as possible so, that there is overlap and this overlap between these two circles as I am showing you here by some markings, shades, so this shaded area as is as the shaded area expands. So, you can say that the fit of your model is becoming better and better. Because, you can then say that the variation in error is getting lesser and lesser as it is going to be explained more and more by the variation in x itself.

So, we stop our discussion at correlation and regression analysis at this moment and we are going to again come back to regression with greater details in the econometrics part of the course. So, next item is going to be analysis of variance. So, please come back and see you then, thank you.