Applied Econometrics

Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology-Kharagpur

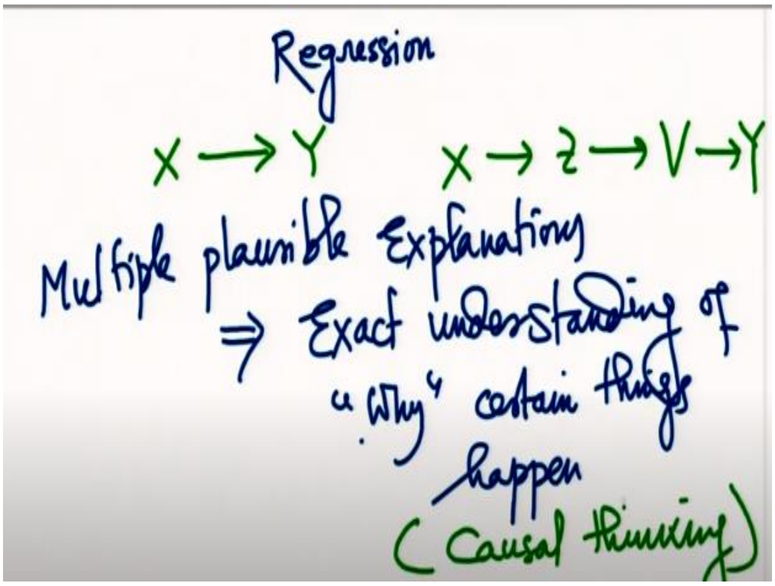
Lecture - 45
Introduction to Regression: Recapitulating Correlation and Causal Thinking

Hello and welcome back to the lecture on Applied Econometrics. In this lecture, we are going to cover a very important topic, actually, this is the beginning of it **is**, which is called regression. And this is essentially going to be the main topic that we are going to cover in our module 2. And when we started this course, we told you that there will be three modules and the module 2, which is essentially regression and its diagnostics, we are essentially going to do a correlational study.

And one particular deliverable of this course is to have a clear understanding between the difference of correlational thinking and causal thinking. So in this module, and the essentially the regression models which we are going to learn, they are going to explain the correlational thinking in detail, right? So before we begin, let me actually recap the concepts of causal thinking vis-à-vis correlational thinking.

Now when I said causal thinking, we said that causal thinking is really a holy grail in any science, any scientific you know pursuit, because we want to understand what causes what? We want to know the why behind certain phenomena, we want to get the exact explanation in the midst of all the plausible explanation.

(Refer Slide Time: 01:42)



Regression

x → Y   x → z → V → Y

Multiple plausible Explanations

⇒ Exact understanding of "Why" certain things happen

(Causal thinking)

So if I have, so what I want to do in when I do a causal thinking is that I want to know the exact x that is causing y or if there is a certain path that x is causing z and z is causing something like maybe v and v is causing y. So essentially, I want to understand the phenomena, the mechanism, that is actually leading one to the other. And why it is so important to understand the phenomena?

I mean, if you look at any scientific discovery of a greater, great magnitude or even a smaller magnitude, you need to know why it is happening. For example, if you see the discovery of take anything like the fire or discovery of the atom bomb or discovery of electricity, any discovery happened because I know exactly, precisely why certain things are happening from where, right?

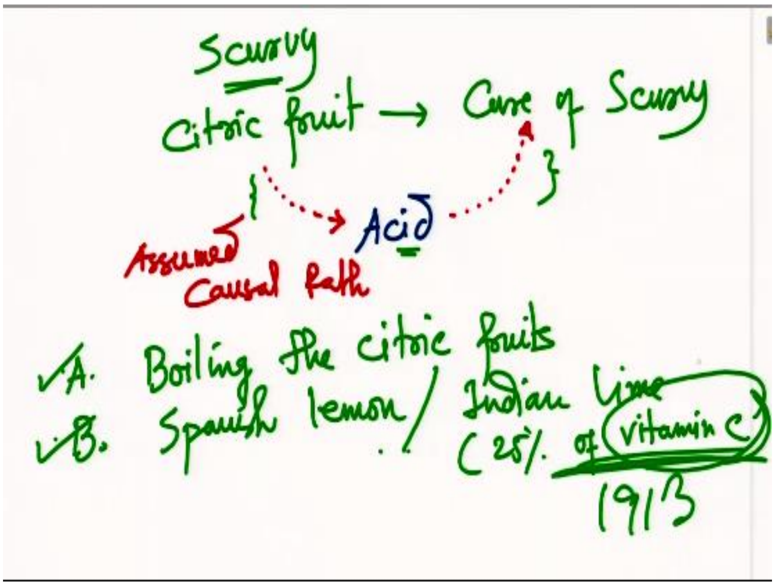
So I know the mechanism. So that is why causality is really important. So it is like a you know like when you have multiple plausible explanations. So there are multiple plausible explanations to every phenomena. But when you do a causal thinking, a causal thinking means exact understanding of why. It is very important why certain thing certain things happen, right?

You need to know exactly why certain things happen. So you need to have this and this is essentially the crux of the causal thinking. Now if we if you actually take examples in you know like, when this the importance of causal thinking is understood in social science as well in economics, social science, sociology, everywhere, and the same question was asked in different context, right?

Particularly, for example, in social science or economics, we say that well, so how the education of a mother really matters to a society. I want to understand the impact and what I do? I actually have this data and I actually try to understand you know the impact of mother education, if that leads to prosperity of the society. Now when I do that, so there are different ways of doing that.

And particularly, when you try to solve the problem using regression, we actually will see that essentially we are thinking from a, you know like a correlational thinking. We are not really able to explain the causal path. We try to do that and there are ways of doing that, and we will actually see that you know. But essentially, when you talk about regression, we are actually, we will actually talk about the correlational path.

(Refer Slide Time: 04:37)


Scurvy
Citric fruit → Cure of Scurvy
Acid
Assumed Causal Path
✓A. Boiling the citric fruits
✓B. Spanish lemon / Indian Lime
(25% of vitamin C
1913

Now again to recall the examples we have given when we distinguished between causal and correlational thinking was the example of scurvy disease, right? Now the example I will briefly recap, and the example was that the scurvy disease it was found that scurvy disease could be, you know cured with the so let us say, with the use of, with the consumption of citric fruit, like lemon or orange, right?

Citric fruit would lead to the cure of scurvy, right? And that was something correlational. People saw that those sailors who consume citric fruit, they actually are, you know they actually do not, they are not likely to have scurvy disease. And James

Lind, Captain James Lind actually published the result of the study. And that become pretty famous.

And people actually started to, you know take a lot of citric fruit when they are going for a sea voyage. And they actually the fatal disease, scurvy was almost like vanished. Now people did not know why exactly the citric fruit is actually helping you to get rid of the scurvy.

So the explanation, the explanation people had is that perhaps citric acid has certain acid and this acid, so essentially what they thought is that citric fruit has certain acid and this acid is actually helping us to cure the scurvy. So that is the sort of assumed causal path. Let us say this is a assumed causal path, right? I am not sure about this mechanism, but I am assuming that this is the causal path.

Now since people did not know the exact causal path, what happened is that many different, you know like, people actually started experimenting, because I do not know the exact route. So I have multiple plausible explanations. So you know I, what I did was that I some sailors actually started, say boiling the citric fruit.

Because when you are going to say, Arctic you know regions, it is really cold, and you might just want to have, you know like a hot beverage that is made out of the citric fruits. So and the assumption is that since the main ingredient we are concerned is the acid in the citric fruit so it is not really spoiled when you boil it.

And then people started consuming that, and you know surprisingly for everyone, the scurvy disease actually came back, you know like after, you know like maybe almost a century the scurvy disease actually came back. And to the surprise of the scientific community, you know like why exactly it has happened. So because the problem was that they did not really understand the actual causal path, it was assumed.

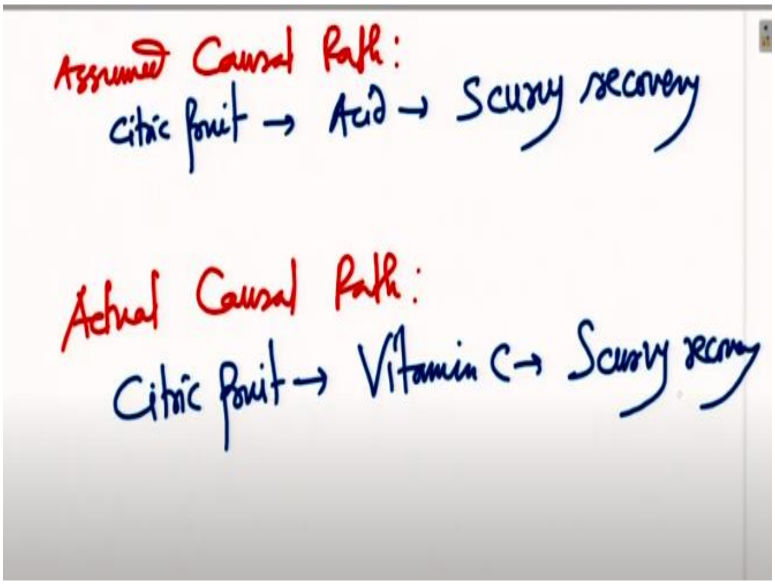
It was only a correlational thinking that led them, you know this vagueness. They did not know exact causal path. So some people actually tried to boil it, some people tried to actually have Spanish lemon or Indian lime, which are actually less in, say I think

25% of actually vitamin C, let us say, which is the actual ingredient, which is later on, you know discovered.

So what happened is because of the substitutes, the substitutes actually were not really functional. And because they did not have the causal understanding, they actually ended up, you know taking all those different ingredients. And because of that, the scurvy disease actually returned back.

Now only when the discovery of vitamin happened in the year I think 1913 people actually started to understand that okay, there is another ingredient called vitamin, which might be there in the citric fruits. And I think later, around after two, three decades, people discovered that it is the vitamin C, which is the main reason that actually helps us to get rid of scurvy. So actual causal path was so the assumed causal path was this, citric fruit acid to scurvy.

(Refer Slide Time: 08:53)



Assumed Causal Path:

citric fruit → Acid → Scurvy recovery

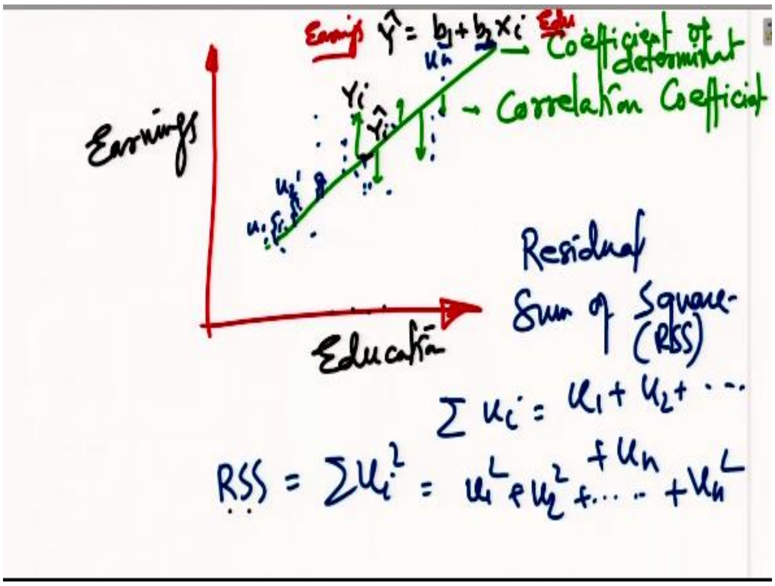
Actual Causal Path:

Citric fruit → Vitamin C → Scurvy recvry

And actual causal path is, so assumed, let us say, let me use a different color. Assumed causal path was and say actual causal path. Let us write down and it is the citric fruit acid to scurvy, scurvy recovery, let us say and the actual causal path, citric fruit, vitamin C, to scurvy recovery. The reason we have given this example previously and again now is that it is really critical to understand this causal path, right?

But we cannot really do that with correlational thinking. So the when we start learning regression, initially we will think we will see it is only correlational. And as we proceed, we will see how we move from correlation to causal thinking by bringing more and more complexity in our regression equation. Now when I talk about correlational thinking, we know that the first thing that we learned is a scatter plot.

(Refer Slide Time: 10:20)


Earnings
Ŷ = b1 + b2 Xi
Edu
Coefficient of determinant
Correlation Coefficient
un
Yi
Ŷi
u2
u1
Earnings
Education
Residual Sum of Squares (RSS)
Σ ui = u1 + u2 + ... + un
RSS = Σ ui² = u1² + u2² + ... + un²

And in a scatter plot what we do, we have say, you know two variables. So let us say, now I take some variables, like let us say earnings and say education, let us say. And what we see is, in general, there is a positive sort of relationship between, sorry let us take different colors, positive relationship between earnings and education, right?

And I know the method, you know like what I do from the scatter plot, we actually draw a regression line, we actually end up drawing a regression line, right? We end up drawing a regression line. Now when I draw this regression line, so when I actually have the scatter plot, I know something called correlation coefficient. And for a regression line, we know something we care about is.

So in general, we calculate correlation coefficient. We will see the, what we mean by correlation coefficient. And we will calculate coefficient determinants, but that we will do when we actually calculate it for that. So but the idea here is that when we draw this regression line, we essentially, you know like, we essentially want to minimize the errors, right? We essentially want to minimize this error terms.

And when you want to minimize this error terms, how we do that? We essentially call something called residual sum of squares. So we could draw this line, we could draw this line, just randomly, we could just, you know just visually see this, visually see these plots, and we can simply draw a line, which is probably somewhat equidistant from all the different points.

But then it is not really, it is a subjective way of doing. We really do not want to do things subjectively. We want to determine the line objectively. And that is why we try to draw a line for which we have basically the errors from all the points, you know the distance between all the points from the line is actually minimum.

And to do that, what we do is, we essentially calculate this residual sum of squares. Now one way to one way we could actually calculate these errors, and we can actually see this, you know minimize the error is actually we can sum total all these errors, right? So what that is, u 1, u 2 … u n, let us say all these different points. So this is, let us say u 1, this is u 2 and this is u n, let us say all these different points.

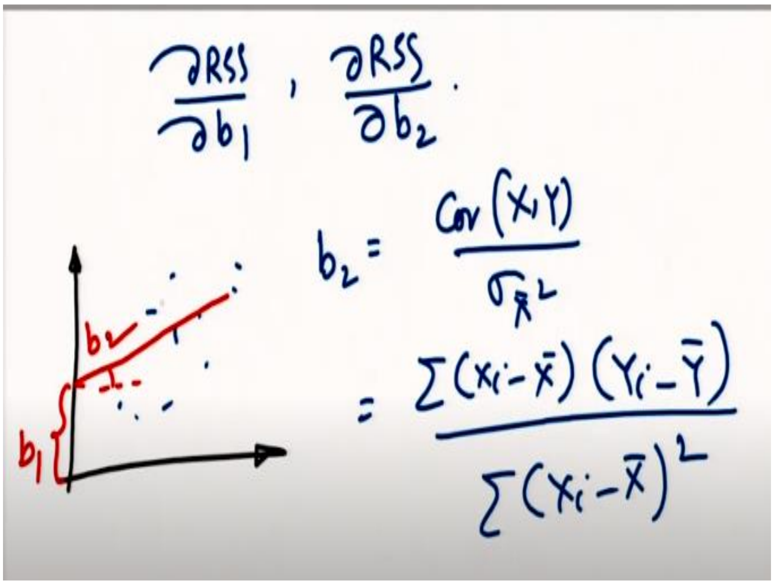
Now but we do not do it because, the problem is if we do that, so what will happen is the errors on the on different sides of the regression line will actually cancel each other. So we do not want that. We will actually want to capture the overall error. And to do that, we actually need to square, we actually need to square this error terms, and then we need to take the sum and that is what we call residual sum of squares or RSS.

Let me write down here or RSS. Now how do we compute RSS? RSS is equal to nothing but summation of u i square, u 2 square, dot u n square. Now how do I sort of minimize RSS? It is quite simple. So if I have this regression line, I will write an equation for this line. And that equation is going to be y, let us say y. So this is I am writing about equation about this line.

So y hat let me, this is going to be y hat. This is estimated y is equal to let us say b 1 + b 2 x i; x i is my all my all the different x points right, all the different points corresponding to education and y is my earnings, okay. Now when I want to, so this is my actual y i, this is my actual y i. And this is my estimated y i hat, right? This y i hat is coming from this line.

Now when I actually do that, so essentially, you know like when I actually calculate this residual sum of squares and when I want to actually minimize this error, so all I need to do

(Refer Slide Time: 15:16)



$$\frac{\partial RSS}{\partial b_1}, \frac{\partial RSS}{\partial b_2}.$$

$$b_2 = \frac{Cov(X,Y)}{\sigma_{\bar{X}}^2} = \frac{\sum (x_i-\bar{x})(Y_i-\bar{Y})}{\sum (x_i-\bar{x})^2}$$

All I need to do is to minimize this residual sum of squares, minimize residual sum of squares. So RSS I have to differentiate with respect to the two coefficients because these are the or parameters in our language. So these are the two parameters which are defining this line. So d beta 1 b 1 and d RSS d beta 2. I am not really going to derive it because this is available in any standard textbook.

So you would rather see that but the result that is going to be helpful for us is that we need to estimate the beta parameter. And this beta parameter that we are going to see basically, or b basically here is that b 2, the value of b 2, the coefficient the or the parameter corresponding to the x variable is covariance of x, y by sigma x bar square. And if I expand it, it is basically x i minus x bar, y i minus y bar by summation of x i minus x bar square.

And this way, what I am going to get finally, is that regression line which I have been looking for. So for all these different points, for all these different points, I am going to have a line which has a, so let us say this estimated, my b 1 and this slope parameter b 2 okay. So this is how I get my simple linear regression line. Here I have only I have taken only one variable that is education.

Here my x is education, y is my earnings, okay. So I have I just have one variable, and that is essentially we conceptualize the idea of regression. Now this is the first part of the, you know lecture on regression. And in the next lecture, we are going to see how we estimate or how we actually conceptualize multiple regression.