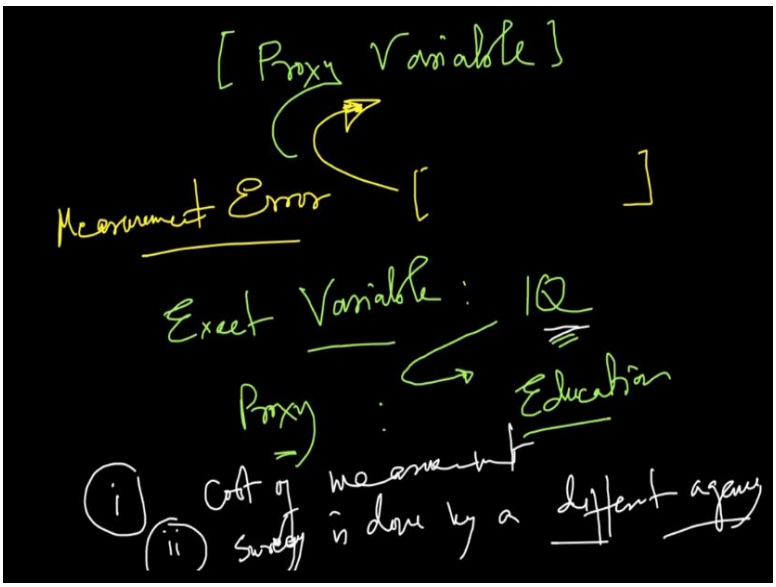
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Lecture – 90
Model Specification - Continued

(Refer Slide Time: 00:50)



Hello, and welcome back to the lecture on applied econometrics. We have been talking about model specification and in model specification, we spoke about the problem of exclusion or inclusion of a variable. And we also talked about measurement error. So in this particular lecture, we are actually going to talk about the problem that is associated with the measurement error.

So let me write it down, measurement error occurs because we do not have the values for the exact variable that we are looking for. And because we do not have values for the exact variable, what we use is a proxy variable. And because we use a proxy variable, we leave out some of the part that we actually want to measure from that exact variable. And let me explain that with an example so that becomes easier for us to get a hold of what is the exact variable, what is a proxy variable.

So let me say that an exact variable, I use the term for let us say measuring wage is IQ of talent of an individual, but we really do not have the data on IQ, rather what we have is some value for education, some data for education, which is basically a proxy. It is a proxy for IQ.

So when I want to measure IQ, but I am actually measuring education, so I am basically using a proxy varibale and it is something like saying like it goes a one eyed uncle is better than no uncle.

So, I have something instead of having nothing. So, that is why there are plenty of uses of proxy variable we will see in the literature of economics where we basically use survey data. And the reason is that it is very difficult to collect data on specific items, let us say IQ. Now, a surveyor will not be able to collect IQ. He cannot go to each and every household or the sample households and actually go and measure IQ of people.

The other part is that like let us say I am a researcher and I am using a survey data, but the survey data is actually collected by someone else. So, the first problem I say is that the cost of measurement. And the second problem that is talking about is that it is a survey data that is collected by; let us say in the case of India it is by the Ministry of Statistics and Programme Implementation and it has a large number of wings across the states and districts and these people are all working to collect this data.

Now, this data is collected based on a standard format and that format is prescribed by again a group of statistician and they decide on what format the data should be collected on and there are different rounds and in different rounds they collect data on different topics let say migration, education, unemployment, health and so forth. Now, I as a researcher may not have the exact requirement whatever data they are collecting.

But rather I would have some requirements for which there is no data or there are some partial data, which is basically collected by some proxy variable. So, that is a second problem. So, it is basically survey is done by a different agency, not by the researcher, so that is a problem. So, the agency does not know or here the ministry does not know what exactly me as a researcher sitting in a corner of a university would be requiring.

And there are so many other researchers, they are requiring so many different aspects of data, even if the ministry knows that these are the data required it is very difficult to collect data for everyone. So essentially, the collect the data in a standard format which I am going to show you in a moment and that format is essentially, that data is used by everyone and people actually use proximity. So let me actually show you before I actually talk further about it.

(Video Starts: 04:53) Let me actually show you some sample of data, which is collected by basically Ministry of statistics in the form of national sample survey data. What I am going to show you is a report. So this is 66th round national sample survey data collected on employment unemployment back in 2019. And let me just show you what are the different tables for which data is collected and what are the different variables for which data is collected.

And this is as I said, it is employment unemployment as well as consumption data that was collected. So let us say I want actually to understand let us say income the people are making and wages and salaries people receive and what are the variables which are responsible for the income or the wages and salaries. So, let us see what are the variables which are covered. So you can actually go to visit, the reason I have kept it is, you can actually go and just type down NSS 56th round and you will actually go to this links.

And if you click this mospi one, mospi dot nic dot in you will actually end up getting the report and basically the same thing you will get, I have already downloded here. So, let us look at the table. So, here, this table 1, table 2 are not much of use, table 3 also we can use it, but let us say if I am interested, I am coming back to the table 3. So let us say I am interested about the education part. So, let us say I want to understand how I can measure the IQ part.

You see, this is the block, block 4 that provides, basically details on education and other things, skills and education and so forth. So, if you see that we have education level general is provided and we have let us say not literate, literate without formal schooling like and then you have below primary, primary, middle, secondary, higher secondary, diploma, certificate course, graduate, postgraduate and above. So that is all you have about if you want to capture information on general education.

Now, if you want to sort of go and see further about people who joined some sort of technical education let us say engineering or medical and things like that, so we have data for that. So no technical education, technical degree in engineering, agriculture, medicine, diploma or certificate below graduate level in agriculture, engineering, and so forth and above, diploma or certificate above graduation level in agriculture, engineering, medicine, craft, and other subjects.

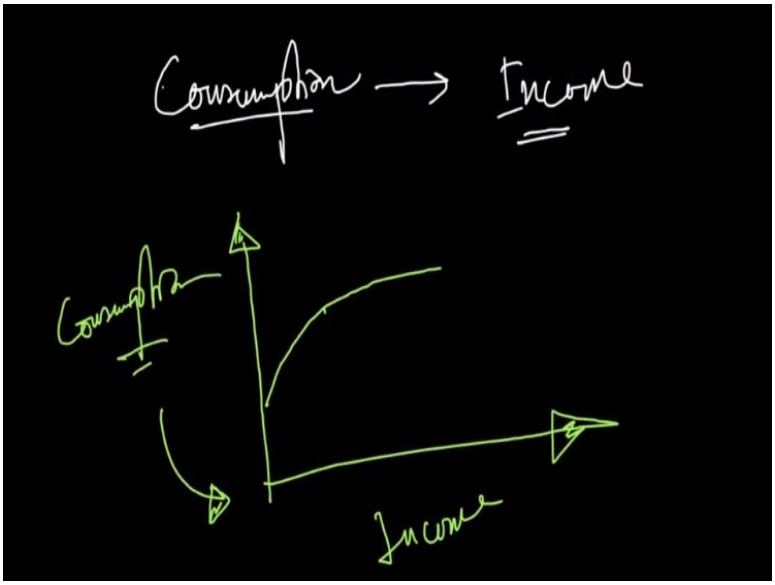
So all that is talking about is the level of education; whether it is technical or it is basically engineering or medical and whether you had a degree or you had a diploma or certificate that is all you get. So you have no idea about how they performed, what is the score or whether they were excellent in some topics, whereas not so excellent in some other topics, whether they had a good communication skill, whether they had a good networking skills, you do not have any idea about all this.

You can also get about if they are currently attending school or they have dropped out so that details you get. And there is another table that provides you the details about vocational training. Whether you have received formal vocational training, non-formal vocational training, learning on the job and all the details, about your field of training and so let us say mechanical engineering and then what are the sources of degree.

If you received this degree from Industrial Training Institute or ITI or polytechnics and so forth. So, it is talking about your trade, sources of your certification or basically degree and basically like this is all the information that you get, but it has no; so a person who has to first in let us say mechanical engineering from certain ITI, you would not be able to distinguish that from a person who has stood last perhaps in the same discipline.

And there are so many other variations which are not provided, it is okay. Now, this is all the data that you can get about education and quite obviously we do not have any variable for my IQ. So, all I can do is basically figure out a way to use data from these different tables. So, this is something that can be explaining as a proxy variable issue. **(Video Ends: 09:29)**

(Refer Slide Time:09:48)



So, let me give you another example which is pretty common and that example is of income and consumption and I previously told you a story of how people actually report wrongly about their income data and that is why often we use informations on consumption as a measure for income. I know that consumption cannot here give you a clearer picture on the income, but it is somewhat telling the true picture.

It is some more telling the true picture because we know there is a relationship that we can see and that is if I have consumption on the Y axis and income on the X axis, so what happens is it is kind of seen everywhere that his consumption is something appears like that. So, basically as your income increases your consumption keeps on increasing but at a decreasing rate, so it does not increase in the same rate.

So, because of this relationship, I can actually use consumption to get some idea about the income. And that is why in NSS we actually can see different consumption information being collected, whereas if I want to measure the income. **(Video Starts: 10:49)** so, this particular round, this one is employment unemployment round and still we we have a separate consumption round, so that data is collected separately.

But even here you will see that income data is not directly collected, some wage data is collected. But if you see there are a lot of consumption data being collected. So let us say what kind of foods are being consumed? So it is milk product, cereals and pulses and vegetables, edible oils and fruits and nuts. So people actually tend to give the right information about what are the things they consumed over the past one week.

And then as an economist I should be able to see for what would be the income level from the existing relationship that you can already see. So that is the idea of proxy variable and this is just a couple of examples of proxy variable. And I will end this lecture with one more example that we often use in social science and that is the social status. The social status, it is indeed difficult to measure how, who is standing where in terms of the social status and all the data that you can get here is here.

In this item 6, you can see the social group and the social group if you can go to the code in item 6 you will have social groups, scheduled tribe, scheduled caste, backward classes and others. So these are the four classes; SC, ST, OBC and others. Now, it does not tell you the entire picture of social status because with someone might have a lot of networks and someone might have a lot of information available to them.

Someone's parents might have a lot of lands irrespective of whether they are SC or ST or general caste, but you cannot really get all this information. So all you can use is to have some idea that schedule tribes are really lower in terms of the social status and perhaps the SC's are little better and OBC's are little better and then on the top of everyone is general caste. So this way, we sort of try to get a sense of the data that we have and usually we see the results that we are getting usually are not wrong. **(Video Ends: 13:09)**

I mean we can actually intuitively explain the results. So, with this I will end the lecture on the introduction to the proxy variable. And in the next lecture, we are going to actually see what are the impacts of actually having a proxy variable in terms of what are the errors, what are bias and what are the standard deviations that we have while using a proxy variable. Thank you.