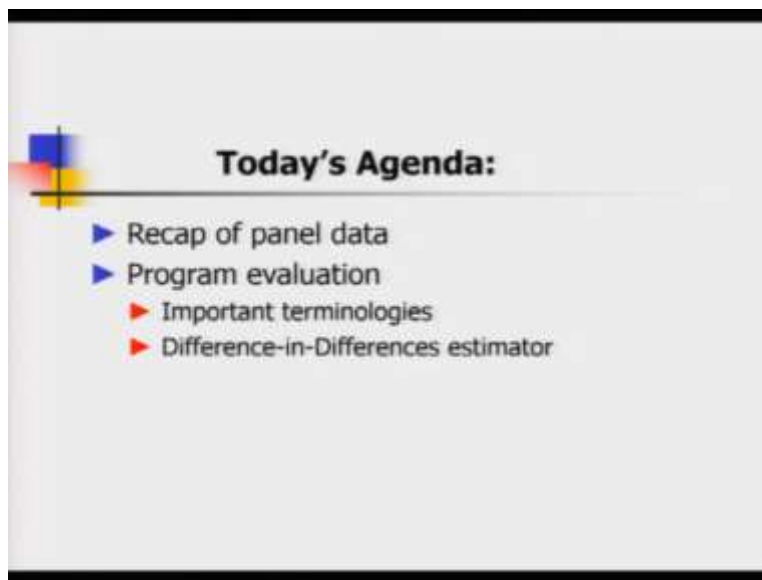


Applied Statistics and Econometrics
Professor. Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture No. 39
Program Evaluation

Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics and today we are going to cover the last topic in the course and that is called Program evaluation. So, in last 15-20 years, this particular topic has become very important in the field of economics but this is not an invention by the economist. So, the basic methodology has come from statistics but it has gained immense popularity in the fields of medical science and socio-economic research of late.

So, although it is very difficult to give you a very good flavor of this particular area because as it is a growing field, tons and tons of literature and methods are developed in last 20 years, so in a 50 minutes lecture it is very difficult to talk about all of those existing methods. So in today's lecture I am going to give you a very brief idea about this field so that at least you are aware that such a field in statistics and econometrics exists. So, before we deal with formal models, let us have a look at today's agenda items.

(Refer Slide Time: 01:39)

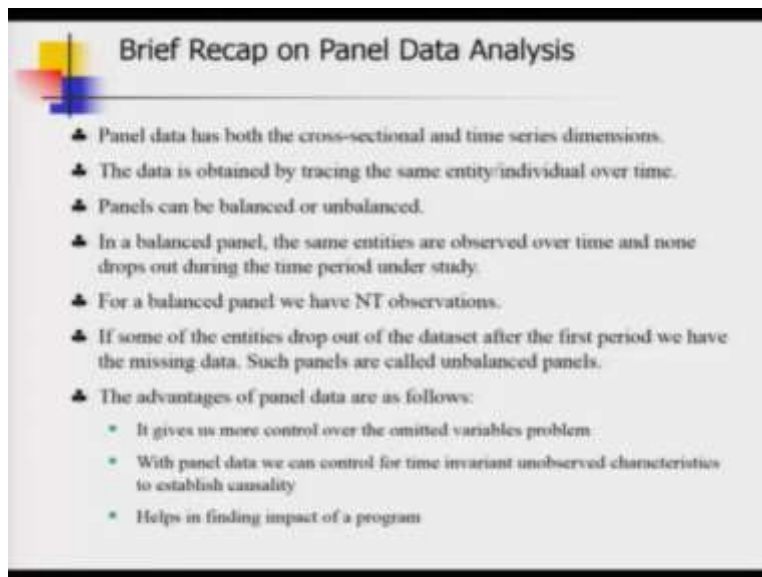


So, at the very beginning we are going to have a very brief 1 or 2 minutes of recap of panel data because now this is going to be used for program evaluation and then we are going to move on to

the basic concept of program evaluation. We will first start with some motivation and then we will cover important terminologies which exists in the field and without knowing them you cannot really appreciate what program evaluation is all about and finally we are going to cover one estimator which is very popular in applied program evaluation and that is called Difference-in-Differences estimator.

So, in the last lecture we already have come across the nature of panel data and we have gone through different estimators for panel data. So, I am not going to give you a recap of the estimators but let us have some discussion on the panel data because then it is going to give you a very good recap and probably will appreciate what panel data is all about.

(Refer Slide Time: 02:50)



Brief Recap on Panel Data Analysis

- Panel data has both the cross-sectional and time series dimensions.
- The data is obtained by tracing the same entity/individual over time.
- Panels can be balanced or unbalanced.
- In a balanced panel, the same entities are observed over time and none drops out during the time period under study.
- For a balanced panel we have NT observations.
- If some of the entities drop out of the dataset after the first period we have the missing data. Such panels are called unbalanced panels.
- The advantages of panel data are as follows:
 - It gives us more control over the omitted variables problem.
 - With panel data we can control for time invariant unobserved characteristics to establish causality.
 - Helps in finding impact of a program.

So, a panel data is obtained by covering the same entity or individual over a period of time to collect data on some prefixed set of variables. Now these panels can be either balanced or unbalanced and in a balanced panel, the same individuals are observed over time. If there are four time periods all individuals are observed for all four time periods and none drops out during the time period under the study.

So, for a balanced panel, we always get capital N times capital T number of observations where capital N is basically the number of individuals and capital T is the number of time periods. But if you have a pooled data set where you see some of the entities or individuals are dropping out of the data set after the first period, then that will lead to missing data problem and such panels

are called unbalanced panels. So, in an unbalanced panel the number of observations are generally less than capital N times T.

Now we are going to revisit some of the advantages of panel data because one of the advantages will lead us to this area of program evaluation. So, of course it gives us control over omitted variable problem and we have already discussed that in the last lecture and you see this is going to haunt us, this particular omitted variables problem is going to haunt us even in this lecture. So, in panel data, we can control for time invariant unobserved characteristics to establish causality and the second bullet point actually leads to the point that these panel data helps us in finding impact of a program.

Now what do we mean by impact of a program? So, there could be either medical experiment where you have in certain lab setup, you have a treatment group and you have a control group and treatment group receives a particular treatment or some kind of drug use whatever and then there is in a control group and then you check what is the impact on the treatment group after you have implemented some treatment on them and in the public policy or socio-economic platform.

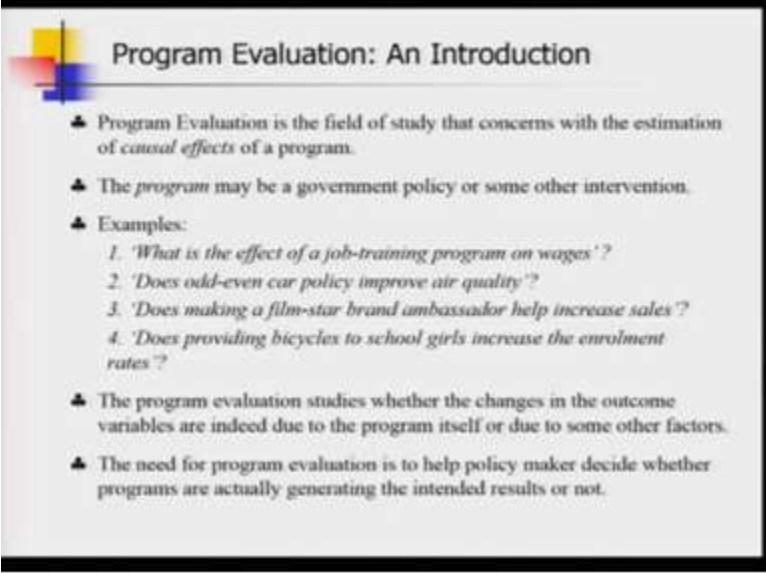
There could be governmental programs or there could be programs led by academicians or NGOs and these are developmental programs, so after you run a particular program the funder of the program or the sponsor of the program may want to know what is the impact of this particular program that they have funded because money is precious. So, everybody wants to see the return of the money. So, if after spending crores and crores of rupees on a particular socio-economic developmental program, if the net benefit of the program is not positive, then the funder will not give money to this kind of projects in the future.

And also it is important from the government's perspective also because if government has some idea some kind of policy or intervention it is thinking about, then before it launches that kind of program or policy at a national level they may want to run a pilot project and then they actually want to see what is the impact on the society from that pilot project and if the results are positive then they can actually roll it out at the national level.

So, you see program evaluation has been very important and effective in decision making and especially in developmental field for last 10-20 years and that is why probably in 2019 the Nobel

committee has awarded Nobel Prize in economic sciences to professors Abhijit Vinayak Banerjee, Esther Duflo and Michael Kramer for their contribution in this particular field.

(Refer Slide Time: 7:27)



Program Evaluation: An Introduction

- ▶ Program Evaluation is the field of study that concerns with the estimation of *causal effects* of a program.
- ▶ The *program* may be a government policy or some other intervention.
- ▶ Examples:
 1. 'What is the effect of a job-training program on wages'?
 2. 'Does odd-even car policy improve air quality'?
 3. 'Does making a film-star brand ambassador help increase sales'?
 4. 'Does providing bicycles to school girls increase the enrolment rates'?
- ▶ The program evaluation studies whether the changes in the outcome variables are indeed due to the program itself or due to some other factors.
- ▶ The need for program evaluation is to help policy maker decide whether programs are actually generating the intended results or not.

So, we have given some kind of motivation to you. Now let us talk about program evaluation more formally. So, this slide you can think about it as an introductory slide. It actually talks about the some examples which can be solved by program evaluation method. So, program evaluation is a general field of you know statistical analysis or econometric analysis that concerns with the estimation of causal effects of a program.

Now note that here we are talking about a very interesting point in this bullet point. So, here in this particular bullet point, we are talking about causality. So, when you want to evaluate the impact of a particular program or intervention, then there has to be a causal effect. So, if you see that there is a benefit accrued to the society after running this particular program, then you should establish causality, so that you can say that this program actually caused that need benefit that we are observing today. But if you run a simple OLS or all regression models that we have gone through so far they are unable to tell you about causality.

So, the main point that differentiates program evaluation from the previous models that we have studied is causality. So, here at least in terms of the regression analysis only we are going to establish causality, atleast that is our objective. Now let us talk about some examples here. So

here I have listed 4 examples from four different fields so that you can see that this idea of program evaluation can be applied to many fields.

So, let us first talk about developmental program, so there suppose government has rolled out a job training program like Pradhan Mantri Kaushal Vikas Yojana and then one may ask that what is the effect of such a job training program on the labor market wages. So, this question can be addressed by program evaluation techniques. Now second example I pick from the field of environment and of course you may know that recently in India, in the Delhi national capital region, government has rolled out an odd event car policy so that it can put a check on the deteriorating air quality in Delhi region and one may ask this question what is the impact of this odd event car policy on air quality in Delhi and for that program evaluation technique can be useful.

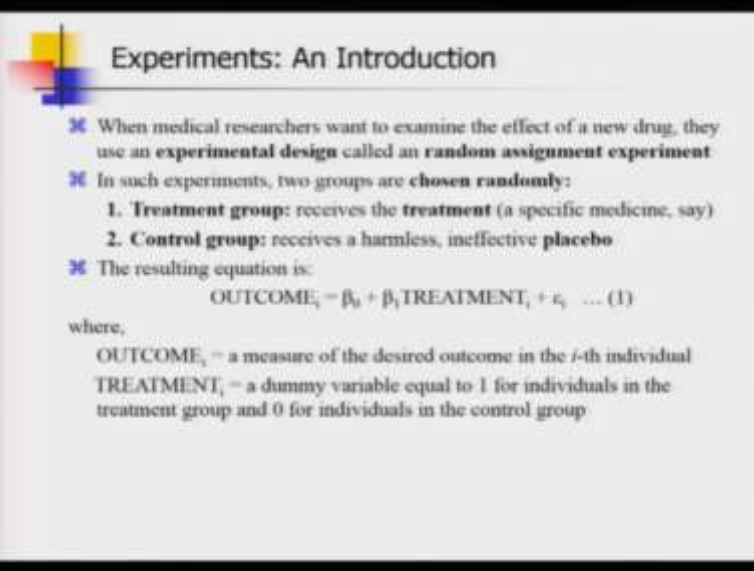
Let us now pick example from corporate or business sector. So, it does not mean that program evaluation techniques are always to be applied by academicians or government folks to calculate the impact of some intervention in the society or in medical sciences when they are trying to come up with a new drug. Program evaluation techniques can also be applied by business folks, so suppose as we see that in India and in many countries film stars or the prominent faces in some sport like cricket or football they are chosen as brand ambassador to endorse certain product and they take a lot of money for that of course.

So, now a company may be interested to know whether employing some film star as a face of their product as a brand ambassador of their product whether there is a statistically significant increase in the sales or not, that kind of research question can also be handled by program evaluation technique.

Now, the fourth question again it comes from developmental projects only rural areas of course there are not very good public transport systems, so to actually help them so that they can go to school every day on time government has gifted them with bicycles and if someone wants to now address this particular policy decision that whether it was effective enough or not so whether gifting bicycles to girl students whether it had any impact on the girl student enrollment or their exam performance or not then you can again adopt program evaluation techniques that are there and of course and as I said we are going to only talk about one particular technique but there are several other techniques.

So, as I said that I can summarize this slide by the last bullet point that there is a need for program evaluation studies to help the policy maker and other stakeholders to decide whether the rolled out programs are actually generating intended results or not so that these programs and policies can be rolled out at a much larger scale. Now, I am going to talk about two different types of program evaluation that exists in the academia and of course in practice. So, the first one I am going to talk about experiments which can also be called natural experiments and the other one I am going to talk about is quasi experiment. So, first let us start with experiments.

(Refer Slide Time: 13:47)



Experiments: An Introduction

- When medical researchers want to examine the effect of a new drug, they use an **experimental design** called a **random assignment experiment**
- In such experiments, two groups are **chosen randomly**:
 - Treatment group**: receives the **treatment** (a specific medicine, say)
 - Control group**: receives a harmless, ineffective **placebo**
- The resulting equation is:
$$\text{OUTCOME}_i = \beta_0 + \beta_1 \text{TREATMENT}_i + \epsilon_i \dots (1)$$

where,
OUTCOME_i – a measure of the desired outcome in the i-th individual
TREATMENT_i – a dummy variable equal to 1 for individuals in the treatment group and 0 for individuals in the control group

So, now in this slide I am going to talk about experiments very briefly. So, let us start with what happens in the medical sciences because this is pandemic times, so all are anxiously waiting for quality vaccines and all. So, let us talk about what happens in labs. So, the medical researchers want to examine the effect of a newly developed drug. So, they actually conduct an experiment and this is called the random assignment experiment and sometimes it is also called randomized controlled trials and in such experiments two groups are chosen randomly.

One is called treatment group and the entities or the individuals in this group be it rat or human being, they receive the treatment say specific dose of the newly developed drug and then there is a control group which receives a harmless ineffective placebo. Now we are talking about a new term treatment and control one can understand that placebo probably is a very new word to you.

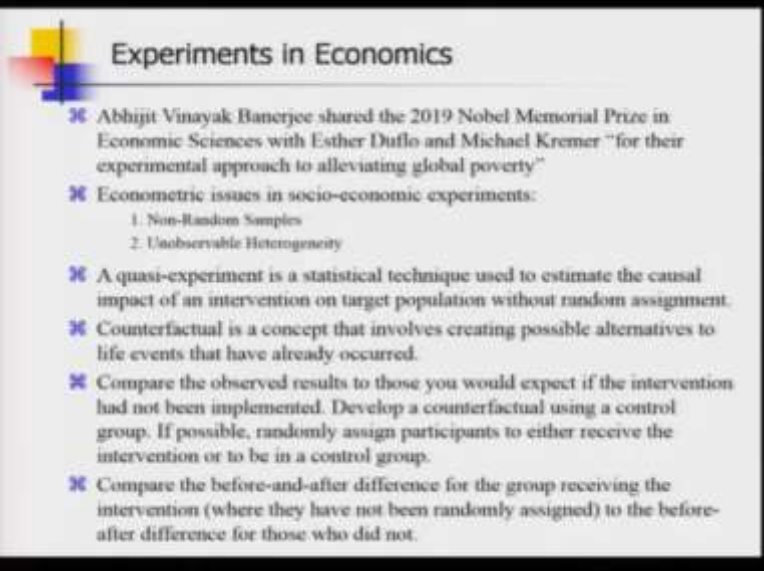
So, what is placebo? So, placebo actually has come from psychology or philosophy, this means that or it is also very popular word in medical science.

So, it is something ineffective, so that is the way in nutshell I can explain it and I just want to tell you from this medical science perspective that what a placebo means. So, sometimes what happens the patients are given some pseudo treatment, some drug which is totally ineffective. It is not going to have any therapeutic value but just for their mental satisfaction or psychological benefit they are given some ineffective dose of drugs or some ineffective chemical which has no role to play.

So, this is basically called the placebo, so when we are talking about placebo effect, we actually say that those subjects in the control group, they actually have not received any kind of intervention, they have not received any kind of treatment that may have some potential effect. So, now let us try to look at the experiments data generated by this subjects in two groups treatment and control, so of course it is going to be a cross section data because all subjects are noted for one time period only.

So, here the resulting regression equation may look like that there is an outcome. So that is a measure of the desired outcome from the i th individual and then there is beta not intercept plus beta 1 times treatment and this treatment i is basically a dummy variable which takes on value 1 for the individuals in the treatment group and takes value 0 for the individuals in the control group and of course there will be a stochastic random error term. Now you may say that okay this regression is good enough, it will give me the expected result, the treatment effect but unfortunately life is not that simple. So, we are going to talk about the complications with this in the next slide.

(Refer Slide Time: 17:25)



Experiments in Economics

- Abhijit Vinayak Banerjee shared the 2019 Nobel Memorial Prize in Economic Sciences with Esther Duflo and Michael Kremer “for their experimental approach to alleviating global poverty”
- Econometric issues in socio-economic experiments:
 1. Non-Random Samples
 2. Unobservable Heterogeneity
- A quasi-experiment is a statistical technique used to estimate the causal impact of an intervention on target population without random assignment.
- Counterfactual is a concept that involves creating possible alternatives to life events that have already occurred.
- Compare the observed results to those you would expect if the intervention had not been implemented. Develop a counterfactual using a control group. If possible, randomly assign participants to either receive the intervention or to be in a control group.
- Compare the before-and-after difference for the group receiving the intervention (where they have not been randomly assigned) to the before-after difference for those who did not.

So, now we are going to talk about the experiments in economics. So, as I started the lecture by saying that professor Abhijit Vinayak Banerjee got the Nobel Prize with two other colleagues of his for their experimental approach to elevating global property. So, of course this price indicates that off late these experimental methods have gained enough importance and popularity in the field of socio-economic research but there could be several problems associated with this kind of randomized control trial or socio-economic experiments in economics.

And I am going to talk about two issues here and they are noted down as the second bullet point and they are non-random samples and unobserved heterogeneity. So, let me explain what are these econometric issues? So, the first one is the example of non-random sample, so what the proponents of randomized control trial say that you should have randomly picked subjects for both the treatment and the control groups like they always do for a medical science lab experiments.

But unfortunately in socio-economic experiments, it is very difficult to assign subjects randomly to treatment and control groups and many times what happens that some people volunteer for the socio-economic experiment then the results that you can generate out of this kind of voluntary participation may not be correct because there could be self-selection bias and all. So, why people do self-select?

Because if they see that there is a benefit coming his or her way by participating in this randomized controlled trial socio-economic experiments, then some people may just opt for it because they want to get some benefit but that is not a very good idea. Actually there has to be random assignment of subjects in these two groups as per the statistical theory. And let me now talk about the second point which is the unobserved heterogeneity. So, now you think about that equation that I have shown you in the last slide.

So, it was a very simple equation, so suppose you have subjects randomly picked and assigned to treatment and control groups but are there any other confounding factors? There could be dozens of confounding factors or other covariates which you are not controlling for but they are impacting the outcome. So, like this job training program that I was talking about, suppose you get a set of subjects assigned to treatment group and the control group but suppose the composition of this group is not homogeneous.

So, what do I mean by that? It may be the case that in one group there is more presence of girls and in one group there could be more presence of some people who have not attended school, so they are not that educated or in one group the average age of the subjects maybe higher than the other group. So, there could be many other things which are not controlled by that simple regression equation. So, you have lots of omitted variable in the regression equation and then that can lead to omitted variable bias.

Now you may ask okay I can get data for these observed variables like age, education, gender these are all observed. Yes, they are observed but even after you have taken care for those observed factors, there could be some unobserved factors like ability of a person, motivation of a person. So, these things you cannot control for. So, even you have a very nice looking regression equation involving a lot of socio-economic demographic variables taking care for this kind of potential confounding factors but you cannot be sure.

So, omitted variable bias actually hurts us. It is a basically kind of obstacle in order to get a causal effect. So, how to take care of that? So, as I have told you in the previous lecture that one nice way to come around or get away with this omitted variable bias problem is collecting panel data and make use of panel data estimation techniques. So, you see we are going to make use of that strategy here also.

So, if we cannot conduct a randomized control trial following statistical theory by assigning subjects randomly to either of these two groups treatment and control, then we have to go for a quasi-experiment. And what is a quasi-experiment? It is a statistical technique that is used to estimate the causal impact of a program intervention on target population without random assignment.

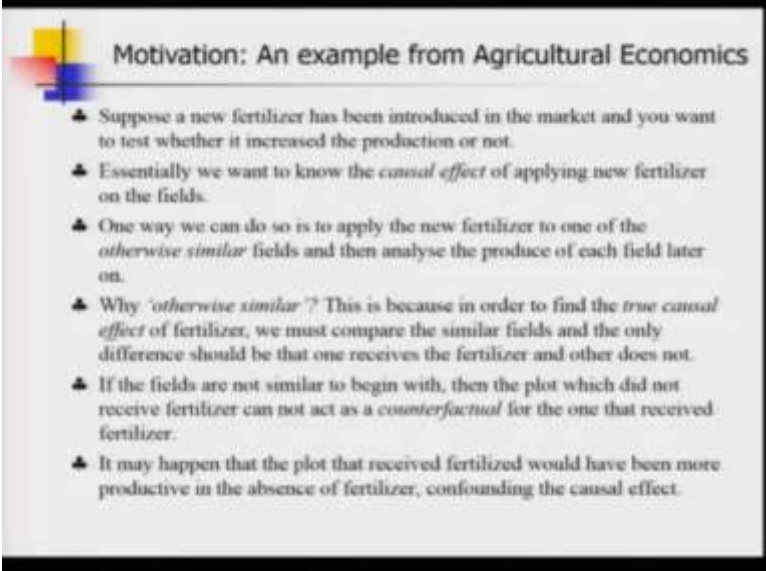
So, that is the most important thing that even if we do not have random assignment, we have some way out and that is what quasi-experiment tells us. So, here in the quasi-experiment method and in the experimental method, there is one terminology that you will come across again and again and that is called counterfactual and I think at this juncture, we must have a very good idea about the counterfactual. So, it is a concept that involves creating possible alternatives to life events that have already occurred. So, it is basically talking about what if scenarios.

So, now you have to make use of this counterfactual strategy and you can compare the observed results to those you would expect if the intervention had not been implemented. So, you have to develop a counterfactual using a control group, now if unfortunately you cannot assign subjects to this control group randomly then actually you can use some other control group where maybe the participants are or the entities are not randomly assigned.

So, in that case basically if the random assignment is not possible then actually you have to go for an econometric method which is popularly called DID or difference-in-differences method and in that method what do we do? So, there you will see that for both these treatment group participants and control group participants we actually compared the before and after scenarios.

So, basically we compare the before and after difference for the outcome variable for the group receiving the intervention even if they are not randomly assigned to the before and after difference in the outcome variable for those who did not receive that treatment, it implies those who are in the control group. Now let me talk about two stories one from agricultural economics and one from health economics to motivate you further and make you understand how we know in reality we can solve these issues or how one can go about this program evaluation, if they face some problems like this.

(Refer Slide Time: 25:57)



Motivation: An example from Agricultural Economics

- Suppose a new fertilizer has been introduced in the market and you want to test whether it increased the production or not.
- Essentially we want to know the *causal effect* of applying new fertilizer on the fields.
- One way we can do so is to apply the new fertilizer to one of the *otherwise similar* fields and then analyse the produce of each field later on.
- Why '*otherwise similar*'? This is because in order to find the *true causal effect* of fertilizer, we must compare the similar fields and the only difference should be that one receives the fertilizer and other does not.
- If the fields are not similar to begin with, then the plot which did not receive fertilizer can not act as a *counterfactual* for the one that received fertilizer.
- It may happen that the plot that received fertilized would have been more productive in the absence of fertilizer, confounding the causal effect.

So, let us first talk about the agricultural economics problem. Suppose there is a new fertilizer in the market and you want to test that whether it increase the production or not. So, basically you want to know the causal effect of applying the new fertilizer on your fields. So, one way we can do this is to apply the new fertilizer to one of the otherwise similar fields and then analyze the produce of each field later on.

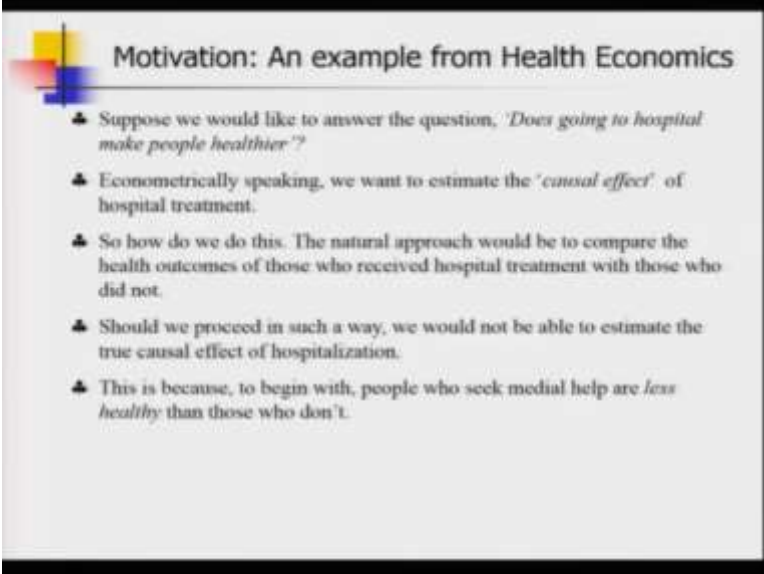
So, basically suppose you have two side by side plots and they are almost equal in the sense that soil and other characteristics of land are equal and you employ same quantity of other inputs like seed irrigation water and monitoring etc. but you just change the fertilizer. So, there will be some difference observed in the total production from these two plots. So, basically now you can compare the yield of these two plots. So, yield is basically total production divided by the cultivated area. So, if there is a difference in yield then you may say that this is due to the difference in the fertilizer.

Now here the fields if they are not similar to begin with then the plot which did not receive fertilizer cannot act as the counter factual for the one that received the fertilizer. So, now let us talk about this problem again. In the first case I said that there are two side by side plots, so you can assume that the soil conditions are almost identical, there will be minute difference but that is negligible but suppose you do not have two side by side plot.

Suppose you have one plot in one village and then you choose another plot in another village, so if there is substantial distance then that distance by that distance actually soil condition may have changed the slope of the land may have changed and of course although to give these two plots same level of monitoring from your end but it may not be possible for you because maybe one plot is nearer to home and you pay more attention there and one the other plot that you have chosen as placebo or control that may be far from your home.

So, you may not be able to pay that much attention there. So, there could be many differences and that differences may contribute to the difference in observed yield. So, you cannot say for sure that the difference that you see in the yield of your crop that is only due to the difference in fertilizer application.

(Refer Slide Time: 28:49)



Motivation: An example from Health Economics

- Suppose we would like to answer the question, 'Does going to hospital make people healthier?'
- Econometrically speaking, we want to estimate the 'causal effect' of hospital treatment.
- So how do we do this. The natural approach would be to compare the health outcomes of those who received hospital treatment with those who did not.
- Should we proceed in such a way, we would not be able to estimate the true causal effect of hospitalization.
- This is because, to begin with, people who seek medial help are *less healthy* than those who don't.

So, now in this slide let me pull one example from health economics. So, suppose we would like to answer the following question. So, does going to hospital make people healthier? So, econometrically speaking, we want to estimate the causal effect of hospital visit. So, how do we do this? So the natural approach would be to compare the health outcomes of those who received hospital treatment with those who did not visit hospital if they feel ill. So, should we proceed in such a way where we would not be able to estimate the true causal effect of hospitalization?

But if we actually follow the procedure that I listed under third bullet point, then we will not be able to estimate the true causal effect of the hospitalization and this is because to begin with

people who seek medical help are less healthy than those who do not. So, here in this point what I try to say that there could be unobserved heterogeneity in individuals that may be detrimental so that you just cannot follow this simple step three and come up with the causal effect. So, you have to now take care of this unobserved individual heterogeneity in order to find the causal effect.

(Refer Slide Time: 30:20)

Hospitalization example continued...

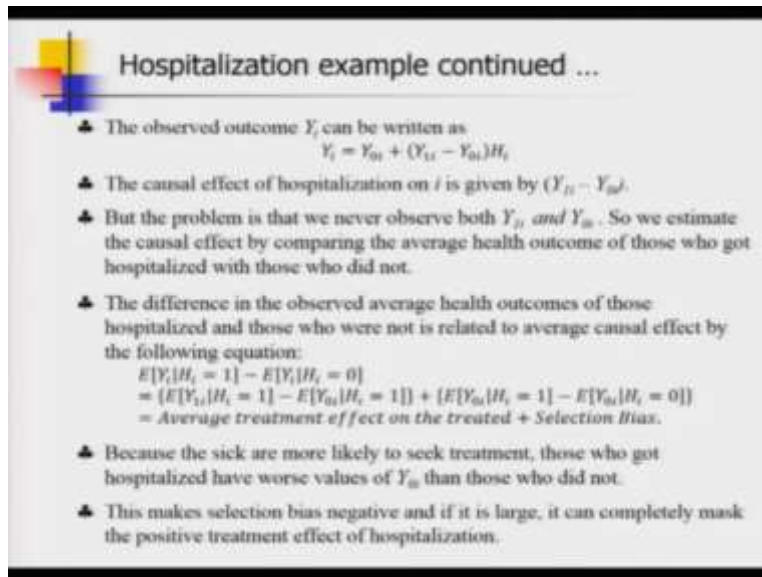
- ▶ In econometric terminology we have the 'selection bias'.
- ▶ People who were hospitalized and sought medical care are not as healthy, on average, as those who were never hospitalized.
- ▶ However, they may be better off than they would have been otherwise.
- ▶ To be more precise about how to estimate causal effect, let's define a binary indicator $H_i \in \{1, 0\}$

$$\text{Potential Outcome} = \begin{cases} Y_{1i} & \text{if } H_i = 1 \\ Y_{0i} & \text{if } H_i = 0 \end{cases}$$
- ▶ Y_{1i} is the health outcome of i if he goes to hospital. Y_{0i} is the health outcome of i had he not gone to the hospital.
- ▶ The causal effect of going to hospital is $(Y_{1i} - Y_{0i})$

So, let us now continue with that hospitalization example. So basically, there could be self-selection bias right. So people who were hospitalized and sought medical care probably were not very healthy on an average compared to those who did not choose to be hospitalized. So, this self-selection bias is coming from the unobserved heterogeneity or the omitted variable. So, now what can we do?

So, to be more precise about how to estimate the causal effect here, let us define a binary indicator or dummy variable say H_i , which will take two values. So, we define potential outcome as Y_{1i} , if H_i is equal to 1 and Y_{0i} if H_i is equal to 0. So, Y_{1i} is basically the health outcome of individual i , if he goes to the hospital and Y_{0i} is the health outcome of individual i who had not gone to the hospital. So, the causal effect of going to the hospital it seems to be Y_{1i} minus Y_{0i} . But is it that simple? How can we get a measure for this?

(Refer Slide Time: 31:44)



Hospitalization example continued ...

- The observed outcome Y_i can be written as
$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})H_i$$
- The causal effect of hospitalization on i is given by $(Y_{1i} - Y_{0i})$.
- But the problem is that we never observe both Y_{1i} and Y_{0i} . So we estimate the causal effect by comparing the average health outcome of those who got hospitalized with those who did not.
- The difference in the observed average health outcomes of those hospitalized and those who were not is related to average causal effect by the following equation:
$$\begin{aligned} E[Y_i|H_i = 1] - E[Y_i|H_i = 0] \\ = \{E[Y_{1i}|H_i = 1] - E[Y_{0i}|H_i = 1]\} + \{E[Y_{0i}|H_i = 1] - E[Y_{0i}|H_i = 0]\} \\ = \text{Average treatment effect on the treated} + \text{Selection Bias.} \end{aligned}$$
- Because the sick are more likely to seek treatment, those who got hospitalized have worse values of Y_{0i} than those who did not.
- This makes selection bias negative and if it is large, it can completely mask the positive treatment effect of hospitalization.

Now note that observed outcome Y_i can be written as $Y_{0i} + Y_{1i} - Y_{0i} \times H_i$. So, we are just making use of that dummy variable that we have created. So, the problem here is that we will never observe both Y_{1i} and Y_{0i} . We can only observe one of them, so we estimate the causal effect by comparing the average health outcome of those who got hospitalized with those who did not.

So, here the difference in the observed average health outcomes of those who were hospitalized and those who were not hospitalized is related to the average causal effect by the following equation. So, here I am showing you these equations, so you see I am taking the difference between two conditional means to calculate my causal effect. So, if you look at the second bullet point in the slide, I have already given you a simple form of causal effect or the treatment effect but I have also explained you in the bullet point 3, why you cannot actually make use of that simple expression to measure your treatment effect.

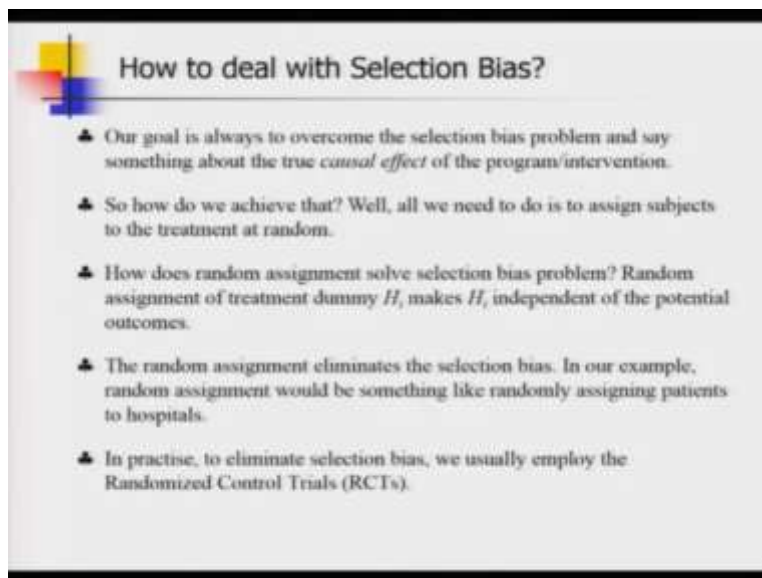
So, basically you have to take the arithmetic mean or average of these two potential cases. So, basically you calculate the mean from the i th individuals, who were hospitalised. So, some health outcome is there, so you take mean of that so that is basically the conditional expectation we are talking about and then you measure the same health outcome or the response, you take mean for those individuals who were not hospitalized.

So, this is basically the case expected value of Y_i given H_i equals to 0. So, now you take the difference, you do some algebra and that is the way you get this average treatment effect on the treated plus the selection bias. So, if you do not understand that expression part and how the causal effect is broken down to the average treatment effect plus the selection bias, do not worry about that.

If you have understood that conducting program evaluation is not very easy because there will be self-selection bias and there could be omitted variable bias issues, then that is good enough for our purpose because this is just an elementary course and in 50 minutes time I cannot explain everything. So, if you are interested, feel free to consult other text books to learn more about this. Now note that because the sick people are more likely to ask for the treatment, those who got hospitalized have worse values of Y_{0i} than those who actually did not.

So, this makes the selection bias negative and if it is large, it can completely mask the positive treatment effect of the hospitalization. So, you see even after bringing in statistics and little bit of algebra and decomposing this causal effect in average treatment effect and the self-selection bias the problem still remains because now the out of these two components, if one becomes a very large negative number it will actually overcome the average treatment effect which could be a positive number and then the total effect will be a negative outcome.

(Refer Slide Time: 35:24)

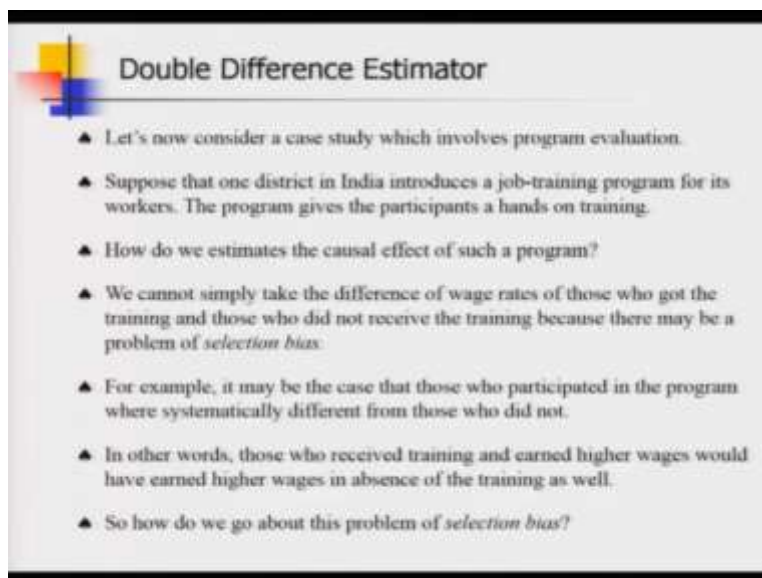


How to deal with Selection Bias?

- ▲ Our goal is always to overcome the selection bias problem and say something about the true *causal effect* of the program/intervention.
- ▲ So how do we achieve that? Well, all we need to do is to assign subjects to the treatment at random.
- ▲ How does random assignment solve selection bias problem? Random assignment of treatment dummy H_i makes H_i independent of the potential outcomes.
- ▲ The random assignment eliminates the selection bias. In our example, random assignment would be something like randomly assigning patients to hospitals.
- ▲ In practise, to eliminate selection bias, we usually employ the Randomized Control Trials (RCTs).

So, major problem is selection bias. How to deal with selection bias? So, our goal should be to overcome the selection bias problem and say something about the true causal effect of the program or the intervention. But how to achieve that? So what we can do of course we can go for random assignment and all but does it solve the self-selection problem? So, suppose random assignment of treatment dummy H_i makes H_i independent of the potential outcome. Well if random assignment is possible then I would also suggest that you go for it to take care of the issue but many times as I said that in socio-economic experiments random assignment is just not possible.

(Refer Slide Time: 36:16)



Double Difference Estimator

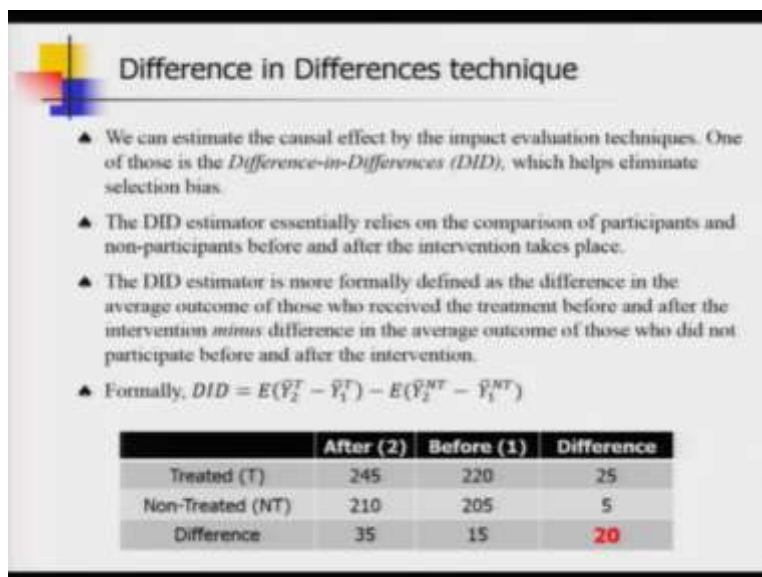
- ▲ Let's now consider a case study which involves program evaluation.
- ▲ Suppose that one district in India introduces a job-training program for its workers. The program gives the participants a hands on training.
- ▲ How do we estimate the causal effect of such a program?
- ▲ We cannot simply take the difference of wage rates of those who got the training and those who did not receive the training because there may be a problem of *selection bias*.
- ▲ For example, it may be the case that those who participated in the program were systematically different from those who did not.
- ▲ In other words, those who received training and earned higher wages would have earned higher wages in absence of the training as well.
- ▲ So how do we go about this problem of *selection bias*?

So, now we are going to talk about double difference or difference-in-differences estimator which is applicable when that your subjects in these two groups control and treatment are not randomly assigned. So, suppose we talk about it through a story or hypothetical example, so suppose in a district government wants to roll out a job training program and you want to evaluate the treatment effect or the program effect for this kind of policy intervention.

Now what to do? So, we cannot simply take the difference of the wage rates of those who got the training and those who did not receive the training because there may be a problem of selection bias and not only that, there could be many other confounding factors. So, for example, it may be the case that those who participated in the program, they were systematically different from

those who did not participate in the program. So, again the problem of self-selection bias is appearing here. Now how to solve that issue?

(Refer Slide Time: 37:34)



Difference in Differences technique

- ▲ We can estimate the causal effect by the impact evaluation techniques. One of those is the *Difference-in-Differences (DID)*, which helps eliminate selection bias.
- ▲ The DID estimator essentially relies on the comparison of participants and non-participants before and after the intervention takes place.
- ▲ The DID estimator is more formally defined as the difference in the average outcome of those who received the treatment before and after the intervention *minus* difference in the average outcome of those who did not participate before and after the intervention.
- ▲ Formally, $DID = E(\bar{Y}_2^T - \bar{Y}_1^T) - E(\bar{Y}_2^{NT} - \bar{Y}_1^{NT})$

	After (2)	Before (1)	Difference
Treated (T)	245	220	25
Non-Treated (NT)	210	205	5
Difference	35	15	20

So, here we propose this DID technique and we can estimate the causal effect by DID technique which helps us to eliminate the selection bias. So, in this slide, I am going to talk about a very simple situation where you see how DID is implemented. So, the DID estimator essentially trusts on the comparison of participants and non-participants before and after the intervention takes place. So, if you remember in the lecture 10-15 minutes before only I have seen spoken about and now we are going to see how the DID will compare the participants in treatment and control group before and after the intervention has taken place. So, how to define a DID estimator formally?

So, it is formally defined as the difference in the average outcome of those who receive the treatment before and after the intervention minus the difference in the average outcome of those who did not participate before and after the intervention. So, formally speaking we have this equation DID, so here let me introduce two new notations. So, T is basically the treatment group subjects and NT is basically the non-treated agents or subjects. So, they belong to the control group and they are denoted by NT.

And now you see that we have to take the difference before and after for both the groups, so let us first talk about the treatment group or the treated subject. So, here suppose we have two time

periods, 1 and 2. So, 2 is basically denoting the after period and one is denoting the before time period. So, basically you have to calculate the expected value of the difference between \bar{Y}_{T2} and \bar{Y}_{T1} .

So, and then you have to now focus on the control group similarly you can calculate the expected value of \bar{Y}_{2NT} minus \bar{Y}_{1NT} and now if you take the grand difference, difference of these 2 expected values then basically you get the average treatment effect by following the difference-in-difference technique. Now we know this is all theory and expressed in terms of words, let us now look at a very simple hypothetical example so that we can understand how we know this technique works.

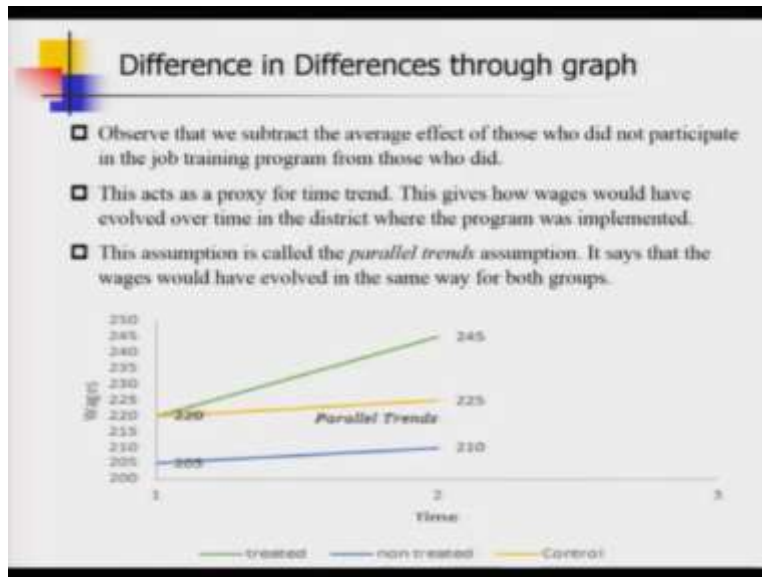
So, now concentrate on the table which is shown at the bottom of the table here and you see that there are 4 cells with some number. So, this is the outcome variable Y on which as an applied researcher you have interest and there is some hypothetical program treatment intervention which has taken place between time period 1 and time period 2 and there were subjects in treatment group who were treated with this intervention and there were some subjects who were not treated with this intervention.

So, here these 4 cells are reporting some numbers. So, now you see that this difference for the treated participants between the before and after scenarios is 25 and for the non-treated participants the difference between the after and before time periods is only 5. Now then you have to take the mean of these two differences that you have computed for the treated participants and the non-treated participants and you get the number 20.

Now, if you look at it from a different point of view, suppose now you want to just compare the numbers by keeping after time period fixed and then you need to compare actually the treated versus non-treated then you see the difference is 35 and if you now look at the column which says time period 1 which is the before scenario and you note down the difference between these two numbers and it is 15.

So again, if you take the difference for two different time periods, you again land up with the same number 20. So, here the 20 actually is the value for DID and that is basically proxy for the average treatment effect. So, now let us have a look at the same problem through a graph so that you can understand what is happening.

(Refer Slide Time: 42:51)



So, now we are going to talk about DID through a graph. Note that we subtract the average effect of those who did not participate in the job training program or in some other program from those who did participate in the program. So, now we know as you can see that with time also there could be some change in the explanatory variable or the response variable or the outcome variable and this is pure trend defect.

So, basically you have to take care of the trend and you have to actually keep it out from the calculation so that if you observe that there is some change, that time trend change or the movement in time shall not contribute to that change. So, you have to take care or control for the time trend also. So, here if you note down in the graph so I have plotted all these numbers that you have seen in the previous slide in that table and here you see that I have three straight lines and they are colored differently to denote different things.

So, there is one treated group, so that is basically given by that green straight line and you see that there is a rise from 220 to 245. So, here you can say that 220 is wedge, if we want to talk about that job training program as the example and now you see that there is a blue straight line which is more or less flat but still you see there is some change in the number over time periods and this blue straight line is referring to the non-treated people.

So, for the non-treated people you see that there is a small jump from 205 to 210 in terms of wage. So, basically, you see those who did not participate in the job training program even for

them with time there is an increase in the salary in the labor market, may be due to increasing demand or inflationary pressure whatever there could be any reason but those who did not participate in the training program even for them there is a 5 rupees increase in the salary. So, that is just time effect and you need to take care of that time effect. So, what to do?

So, now you go back to the graph and then let's see how we are taking care of the trend or the time change effect. So, here let us now look at that intercept for that green color straight line and that is there at 220. So, now what we are going to do, we are going to draw a straight line, another straight line and that is going to be parallel to this blue color straight line and this newly drawn parallel straight line is colored gold or orange and that basically shows the parallel trend.

So, it means that here this golden color straight line is actually talking about a counterfactual. How come? Because you see here, if there is no training program and those who actually participated in the job training program did not get an opportunity to participate in that same program, how their salary could have changed over time and their salary could have changed naturally like for those it increased over time.

So, for others it increased by 5 rupees from time period 1 to 2. So, those who are actually the participants in this job training program, the increment that you see in their salary you have to deduct that 5 rupees from that number so that you can take care of the increment just by changing time. So, now you have a look at this. This is basically the counterfactual, so we are talking about a what-if scenario. So, the what-if in this case is that what if the participants in the training program did not get a chance to participate in the same training program.

So, now if you look at this graph again, we see that we adjust this counterfactual and we say that well in the absence of the training program those who participated in the training program could have anyway experienced 5 rupees increase in the salary and their salary could have gone up to 225 because they started with an initial salary of 220.

So, the gap between 245 and 225, that is basically 20 that is actually the average treatment effect or the impact of the program. So, now note that we are not computing the impact by taking a difference between 245 and 210, so the result would be 35 but that is absolutely wrong because there is this strained thing hidden in that calculated number. So, once you take out the impact of time, the impact of trend then you get the true measure of the treatment effect and as you are

taking the first difference, the unobserved heterogeneity or the omitted variables will all drop out from your analysis. So, your results are also free from omitted variable bias.

So, this is the end of our discussion on program evaluation. Program evaluation is a very important topic, so that is why you see Nobel Prize was awarded in 2019 for that field. So, if you are interested, there are very nice text books and online resources from where you can learn about program evaluation techniques and real life case studies. So, I would at this point suggest that those who find this topic very interesting should make a visit to the website of J Pal.

So, this is basically an organization which was co-founded by Professor Abhijit Banerjee and these works on the developmental issues in third world developing countries, they run a lot of projects amongst the poor people in Africa and Asia and they actually implement this program evaluation techniques to calculate the treatment effect. So, if you visit their website, you will get to know a lot more about not only their activities but you are also going to learn about program evaluation in practice.

So, with this I end and this actually is the last lecture on the theoretical side of statistics and econometrics. So, next week and in next lecture I will come back with some software issues. So, we will not do practical lab sessions but I will show you how you can use R to calculate or compute some of the things that I tried to teach you in this short course. So, see you then. Thank you.