Pages 1-2

In this video, I will be introducing the key concepts that will bring you into the field of causal inference. I will first discuss why causal inference has become a crucial topic in quantitative social research. And then I will explain the key challenges and the techniques we have for us to evaluate causality. I will conclude this video by discussing different types of causal inference and some of the key practical principles.

Page 3

Social researchers attempt to address different varieties of research questions. Some research questions aim to address the cause-and-effect in the social world. And causal questions are very common, and you will see some examples in the next slide. Causal inference has become one of the most important topics in quantitative social research following the emphasis on evidence-based policy making, as government agencies seek to establish solid causal evidence before they decide to invest a lot of time and money on different policy initiatives.

Causal inference, as I will explain later, is also important because the ignorance of the distinction between correlation and causality can be a matter of life and death. The conventional quantitative social research using multiple linear or non-linear regression cannot make the distinction between the two, and often if we use correlation to infer causality, it is very likely that we can derive misleading and biased conclusion about how the social world works.

Causal inference is also challenging given the complexity of the real world – I am going to explain why the real world and conventional multiple regression often prevent us from deriving valid causality and what we can do about it.

Page 4

These are some of the classical causal questions in political science. As you can see, many of these questions are specific and straightforward. Many of these questions also have very strong policy implications. More broadly speaking, the rise or introduction of causal inference in social research also changes the way we think about our research questions.

Page 5

But just what is causal inference? In a nutshell, causal inference is a field where researchers seek to evaluate the causality of a cause on a specified outcome.

Researchers often use capital X to represent a causal or a causal variable and capital Y as the dependent and outcome variable. The arrow from X to here means that X causes Y.

When it comes to causal inference, more specifically, researchers usually try to address one of the following three analytical objectives. First, we will use different techniques to establish there is indeed causality between X and Y. In other words, before we consider other aspects of causality, we must confirm is X is indeed a cause of Y. This is what we mean by “existence” here. Next, we will also need to show X is an important cause of Y – the quantitative techniques for causal inference can allow us to estimate the size of effect, and the size will allow us to determine whether X is a crucial or trivial cause of Y. This is what we meant by “importance.”

Finally, one of the frontiers of causal inference is to look at the causal “mechanisms.” Very often knowing X is a major or primary cause of Y is not good enough, and in recent years it is expected that researchers should try to address the question of mechanism by discussing how the causality really works. For instance, suppose now we are interested in the causal relationship between corruption and development. Once we have used different quantitative techniques to evaluate and estimate the size and the sign of the effect of corruption on development, as researchers we will be responsible for explaining how corruption may undermine development. And since the real world is very complicated, often it would be good enough if we could carry out additional research to demonstrate at least one theoretically plausible mechanism. In terms of the causal effect of corruption on development, one possible mechanism is political corruption might deter firms from investing in the country to generate job opportunities and tax revenues to support economic development. The focus on mechanism also means that often you can extend one causal inference study that has addressed the questions of existence and importance by considering different alternative mechanisms.

The study of causal mechanisms is one of the frontiers in the field of causal inference. And there are several methods we can consider here. One is to use heterogeneous causal effect to demonstrate the mechanism. Heterogeneous effect basically means we are trying to show the causal effect only exists under some conditions, and we can use the conditions to get a sense of how the cause influences the outcome. Another possibility is to use mediation analysis, which requires to identify one additional variable between X and Y to show how the third variable can “mediation” the causal effect of X on Y. The final common approach is to use qualitative case studies and process tracing to demonstrate how the cause really contributes to the occurrence of the outcome.

Page 6

Now you know (1) what causal inference involves and (2) the main analytical objectives of causal inference, I would like to spend some time talking about why causal inference is difficult. Despite its importance, causal inference is tricky for at least two reasons. The first reason has to do with real life and means we will never be able to observe perfect causality. The second reason will help you understand why the conventional multiple regression analysis that can provide correlation between the predictors and the outcome variable cannot give us causality.

Page 7

If we think more carefully, in the most ideal situation, we will need “counterfactual” to observe perfect causality.

What is counterfactual? Consider the following example: Suppose we want to evaluate the effect of Covid-19 vaccination – for everyone, to get a sense of the causality between vaccination and prevention of Covid-19, we will need to know the outcomes of a person receiving and not receiving the vaccine and compare the difference between these two outcomes to know whether the vaccine is effective. In real life, there is no way we can observe both outcomes – we can only observe one of these two outcomes. The outcome that we cannot observe is the counterfactual. If I receive the vaccine, then we will never know what would happen if I did not receive the vaccine. The outcome caused by me not receiving the vaccine is the counterfactual. Vice versa, If I did not receive the vaccine, then we will never know the corresponding consequence of me being vaccinated. We cannot simply use the fact that me being vaccinated and not getting Covid-19 as the evidence to show the vaccine is effective. To show the vaccine is effective we will also need to know the outcome of me not being vaccinated because it is possible that I would not get Covid-19 even if I am not vaccinated.

Now you show see the conundrum here: We will never be able to observe perfect causality at individual level. For each person, we will have two potential outcomes, one is the outcome when the person receives the treatment (like the vaccine) and the other is the outcome when the person does not receive the treatment. This conundrum is also understood as the impossibility of causal inference.

To get around this conundrum, we will have to go to the group level. Say, we can create two groups of participants -- one group would receive the treatment and the other would not receive the treatment. We usually call the group getting the treatment as the treatment group and the group not getting the treatment as the control group. In the case of vaccination, we can compare the average rate of getting Covid-19 between the treatment and control groups to see if the vaccine is effective.

Page 8

Here we can use some mathematics to model the potential outcomes.

Suppose we have a binary treatment D, which can take the values of 0 or 1, with 1 showing the situation such that the participant receiving the treatment. In this case, 1 means the participant i is in the treatment group and 0 means the participant i is in the control group.

And when the participant is in the treatment group, we know the corresponding potential outcome is Y one lower i. In contrast, when the participant is in the control group, let us define the corresponding potential outcome is Y zero lower i.

Then which potential outcome we will observe in real will depend on whether or not the participant is in the treatment group. This is what the last line in Equation (1) tries to tell us. If the participant is in the treatment group, we know D i is 1, then Y i, the observed outcome will be Y 1 i. In contrast, if the participant is now in the control group, which means D i is 0, then then observed outcome is Y 0 i. Since the observed outcome depends on which group the participant belongs to and we can only observe one of these two potential outcomes, this equation tells us it is impossible to observe causality at individual level.

Page 9

If you find it a bit difficult to understand the concept of counterfactual, then the movie Everything Everywhere All at Once might be a good example, as we can only observe counterfactual scenarios in a world where we can access multiuniverse.

Page 10

In addition to the lack of valid counterfactuals at individual level, perfect causal inference is nearly impossible in real life because the treatment and control groups may not be comparable. That is, the treatment and control groups may be different other than the treatment status and making the inference of causality difficult.

To put it differently, the issue of incomparability between treatment and control groups has to do with the treatment assignment mechanism. If we do not introduce any interference, in real life it is very likely that there is a systematic process for some respondents or participants in the population to receive the treatment of interest, such as vaccination. Right? Very often, people decide whether they want to be vaccinated by themselves, and as a result, people who are vaccinated and who chose not to be vaccinated can be systematically different other than the treatment or vaccination status. If we do not take the factors that can determine individual’s decision to take the vaccine into consideration, we may get biased results, especially when the factors that can determine people’s decision to receive the vaccine and the factors that can determine whether people will contract Covid-19 are identical very similar. Technically speaking, the factors or variables that can determine the cause (or the treatment status) and the outcome at the same time “confounders.”

Confounders prevent us from producing valid causal inference. For one thing, confounders can create correlation between X and Y without having X being the cause of Y. For instance, saying a person’s age can determine whether they will be willing to take the vaccine and whether they will get Covid-19. In some countries, it is very likely that older people are more likely to get Covid-19 or even die from it. If in the example older people given their choices are less likely to get vaccinated and are more likely to get Covid-19, then we cannot draw any valid causal conclusion between vaccination and the prevention of Covid-19. To put it differently, since age is correlated with individual’s decision to get vaccinated, people who are in the treatment and control groups are not comparable, as other than the vaccination status they are also different in terms of their average age. If older people are less likely to get vaccinated, then on average people in the treatment group will be younger than people in the control group. As we cannot use the control group to be a valid counterfactual to the treatment group, we cannot be certain that the statistical relationship we observe between X and Y or between vaccination and Covid-19 is correlation or causation.

Pages 11-13

Here are some examples to see how confusing correlation with causation will lead to misleading or biased conclusion about the social world.

First, when we look at government data in the US, often we will see a positive correlation between the amount of police officers being deployed to a neighborhood and the amount of homicide and other severe crimes. Can we use the correlation to conclude that the presence of more police forces causes more crimes? Not really when we start thinking about possible confounders. In other words, without any intervention, there may be a variable systematically predicting the amount of police officers being deployed and the frequency of criminal activities in a neighborhood. One possibility, for instance, is the presence of local criminal syndicates – the presence of local gangs might push the local authorities to deploy more policy forces and increases the intensity of criminal activities in a neighborhood.

Likewise, while often regression analysis cannot establish a statistically significant correlation between the deployment of international peacekeeping troops and the number of citizens being killed when one or more countries got involved in some serious armed conflicts, this cannot conclusively tell us the lack of causality between international peacekeeping and conflict resolution or prevention. The reason is very often international organizations will selectively decide where and when they should send international peacekeeping troops, and it is very often the case that international organizations such as the United Nations are more likely to send international peacekeeping troops to conflicts where a large number of citizens have been killed. In other words, the dispatch of international peacekeeping forces and whether these international keeping forces can be effective for conflict resolution may be jointly determined by the underlying intensity of the conflict.

Finally, without taking confounders into consideration, we are very often seeing a positive correlation between the number of hospital visits and the probability of someone passing away. In other words, people who go to hospitals more often are more likely to die or pass away. There is definitely no way we can use this correlation to infer that hospitals are ineffective in terms of treating their patients, because very often how often a person has to visit the hospital and how likely a person will die can be jointly predicted by the person’s underlying health situations.

But this is certainly not the end of the world. We can take the effect of age on both X and Y into consideration by controlling for age in our regression analysis. In other words, age will be our control variable. However, based on our previous discussion on multiple linear regression, you perhaps will see two issues very quickly: First, it is nearly impossible we can come up with a complete list of confounders to be included in our regression. Next, even if we can come up with a complete list of confounders, some confounders may not be measurable. The impossibility to include all possible confounders in our multiple regression also tells us why the traditional kitchen sink approach to regression analysis is very problematic. Using the traditional kitchen-sink approach, researchers usually only include all possible predictors or explanatory variables of the dependent variable in the multiple regression model, but based on our discussion on confounders, you will see including different predictors in the model will never give us causality. That being said, kitchen-sink regression may still be a good choice for exploratory research.

Page 14

~~Here we can use the following equation to model why the presence of confounders may prevent valid causal inference. Keep in mind that when we ignore confounders, we will run into the situation that the control group cannot serve as a valid counterfactual of the treatment group – we often call this as selection bias, as we may get biased conclusion by ignoring how the confounder will determine the selection process of individuals into the treatment group.~~

Now we know it is impossible to use control variables to completely remove the influence of confounders and tease out the causality between our X and Y. What can we do then? The solution that is widely considered more effectively is randomization. By definition, by randomizing our treatment assignment, we are trying to break the causal arrow from our confounders to the causal variable, X. In this sense, we are trying to create the situation such that there are no variables that can predict whether or not a respondent or a participant will be assigned into the treatment or control groups. By getting rid of the effect of confounders on our causal variable, we can create a situation such that the treatment and control groups will be comparable on average, as the only difference between the treatment and control groups is whether or not the participants in the group receive the treatment or not.

Page 15

Getting rid of the confounders in our analysis is the core idea in the field of causal inference – or put it more technically, our goal is to identify a causal effect. To identify a valid causal effect, we will need to make sure the control group can be a valid counterfactual to the treatment group. To do so, we will need to create or find a situation such that the treatment assignment is a random process so we can remove the selection bias.

In this case, causal inference is mostly a design endeavor and has less to do with the selection of regression models or regression estimation strategies.

Page 16-17

And depending on how we can achieve the goal that the treatment assignment is a random process, we will engage in different types of research.

For a randomized controlled trials, we have a well-defined treatment for our research and randomization is carried out by the researchers themselves. If the treatment is carried out by others and not designed specifically for our research, we may find ourselves in the territory of natural or quasi-experiment. However, in either case, it will require the researchers to have a clear sense of the treatment assignment process, which usually demands a solid understanding of the research context.

Page 18

We will try to cover some of the most common identification strategies in this module.

Page 19

Here is a list of practical reminders for any quantitative social research involving causal inference. First, state a simple and yet thoughtful causal research question. Be clear of the treatment. Next, depending on the treatment assignment mechanism, decide whether this is an experimental causal inference research or a non-experimental or observational causal inference research. Once you decide the type of causal inference, test the assumptions you need to maximize the credibility of your chosen design. Once you state and evaluate the assumptions you need for different identification strategy, the modeling choice or the choice of estimation strategy is often a secondary matter. Finally, once you establish the existence and size of the causal effect, think through the questions of mechanisms.

Keep in mind that we do not mean to say you should only do causal inference in your quantitative research (although some researchers do think so). Rather, we would like to provide you with a solid analytical framework to think through the difference between correlation and causation, and make sure you do not overclaim your quantitative findings.

Page 15

Page 15

Here we can use the potential outcome model again. If we can define the average treatment effect as the observed mean difference between the treatment and control groups, we can rewrite the difference between the average outcome for the treatment group and the average outcome for the control group as the average difference between the potential outcomes for the treatment group and the potential outcomes for the control group. The key question here is we cannot observe the potential outcome of treatment group when they do not receive the treatment and the potential outcome of control group when they receive the treatment. These two are the formal definition of selection bias.