Hello, I'm Professor Sridhar Narasimhan of the Scheller College of Business at

Georgia Tech.

I'm teaching this module on calculating treatment effects

in this course on data analytics and business.

There are five lessons in this module.

The first is on correlation versus causality.

The second is on selection bias.

The third is on randomized controlled experiment and the difference estimator.

Lesson D, is a classic example,

the star experiment, which is to study the effect of small class size.

And then, natural experiments and the difference-in-difference estimator.

So correlation, correlation refers to any of a broad

class of statistical relationships involving two variables.

The sample correlation between two variables, X and Y,

is defined as shown below.

And it's a measure of the linear relationship between X and Y.

And it always lies between negative 1 and positive 1.

Note that if Y is equal to X squared and if X ranges say from

negative 10 to positive 10 then the correlation between X and Y is 0.

In other words, they are uncorrelated, even though they are perfectly related.

If A and B are strongly correlated,

there could be several possible relationships.

A causing B.

B causing A.

If you think A causes B, be careful of reverse causality that is B causes A.

For example, the faster that a windmill rotates,

the more the wind speed is observed to be.

A and B are a consequence of some other common cause.

For example, as ice cream sales increase,

the rate of drowning deaths increases sharply.

Therefore, ice cream consumption causes drowning.

Here C could be the summer season.

A causes C which, in turn, causes B.

A and B are correlated by chance.

There's also the notion of post hoc ergo propter hoc.

This is what's referred to as a faulty line of reasoning that

could occur in some situations.

This is from the Latin after this, therefore because of this.

So it's a logical fallacy that if A happened, and then B happened,

then A must have caused B to happen.

Examples, is a common one, the rooster crows just before sunrise,

therefore, the rooster causes the sun to rise.

Or Professor Sri moves into a new office and the college's A/C breaks down,

therefore, Professor Sri caused the A/C to break.

And there may be other examples like this.

Be very careful when you make a conclusion based solely on the order of events.

But rather, you should also consider other factors that could be responsible for

the result that might rule out the order of events connection.

Causation is a complex concept,

established causation in time,

the hypothesized cause must preceded its anticipated effect.

The change in cause must lead to a change in effect.

You must discount all other plausible explanations,

other than the one proposed, that could explain the relationship.

So why do we, as business professionals need to establish causality?

Causality is needed in certain fields.

For example, in medicine.

Causal models are used to build theories, very useful.

Managers need to know how things work, which is provided by theories.

Managers want to make changes for example, the price, or

the product, make some promotions in the market.

So they need to determine causal impacts of their actions.

So if X does not cause Y, then spending effort in X will

yield no results and it's a waste of money.

If X causes Y then we can use a theory

to explain why X causes Y.

[MUSIC]

In this lesson, we talk about selection bias and

its effects when we estimate the parameter of a linear regression model.

Selection bias occurs when individuals are selected for

treatment without proper randomization.

And this could occur due to several reasons.

There's what's called self-selection bias,

this could occur in poorly designed experiments.

Example, a university offers a program to measure and

improve teaching effectiveness that allows faculty to opt in.

It could be the case that those faculty who did this program

were already very effective teachers and wanted to become even more effective.

It could also be voluntary response bias.

For example,

callers to a radio show are folks who are already interested in that topic.

This sample over represents people who are interested in that topic.

This may not be representative of the population.

Another common problem is what's called nonresponse bias.

That is if non-respondents differ from respondents.

This is often a problem in surveys where the response rate can be very low.

So when estimating the slope coefficient in ordinary least squares,

there's some assumptions.

So this is the model.

So when we regress Y on X we are using the ordinary least squares estimated to

estimate b1, the slope, and that's given by the expression there.

And that's works out to b1 plus the term on the right covariance

of the error term subjects or covariance of y times x.

See the orthogonality assumption in OLS is that the edit terms and

the predictors are not related at all.

So when X and E are uncorrelated,

the OLS estimator is a good estimate of the slope B1.

When X and E are correlated,

then the OLS estimator is not a good estimate of b1.

There are assumptions in the linear regression model.

If X is a dummy variable, then we get the OLS

estimated as b1 + e1 bar- e0 bar.

So b1 is called the treatment effect.

And the e1 bar minus e0 bar is termed as the selection bias.

When the difference between e1 bar and e0 bar is 0,

the OLS estimator is a good estimate of slope b1.

If not, it's a bad estimate.

So how do you control selection bias?

We do that through randomized controlled experiments.

By random assignment of test subjects into treatment and

control groups, we prevent any selection bias from occurring.

This is very often used in the sciences, agriculture, medical research, etc.

In economics, this may or may not be easy to do,

because subjects are people and their economic well being could be affected.

In business, nowadays,

thanks to the Internet and other technologies, we are seeing more and

more experiments being conducted, in some situations, using random assignment.

You should have a natural experiment.

Which we are going to study later on.

There are some techniques to control for selection bias, and

then finally you could add control variables.

[MUSIC]

In this lesson, we'll see how a randomized controlled experiment is set up.

And we introduce the difference estimator.

How is a randomized controlled experiment set up?

We take a quick look.

So we have this initial sample of all the folks for

whom we are considering for this experiment.

We draw a random number for each observation.

If the value is less than 0.5, we put that individual in the control group,

which gets no treatment or gets a placebo.

If the random number chosen for that observation is greater than or

equal to 0.5, we put that individual or observation in the test group.

And if you notice, we've introduced the dummy variables.

So essentially for each observation, we set that variable to 0 or 1.

Set to 0 if that individual is in the control group, and

d = 1 if that individual is in the test group.

So here's the regression model.

So to analyze the results of a randomized controlled experiment

via regression model, we've defined that variable d.

And the regression model for

that observation is shown here, y as a function of d.

We need to use this dummy variable to distinguish the subjects in the treatment

and control groups.

So the regression functions are depending on which group and individual is in.

Expected value of y, is it a b0 + b1 or just b0?

It's just b0 if the individual is in the control group.

So the difference estimator is used in these cases so

to calculate the treatment effect.

So the OLS estimator is given by that expression.

And its value equal to y1 bar- y0 bar,

with those two terms defined there.

So y1 bar is for the average value of y for the observations in the treated group,

and y0 bar is the average value of y for the observations in the control group.

And N1 and N2 are defined similarly.

So bOLS is called the difference estimator, because it's the difference

between the sample means of the treated and controlled groups.

The difference estimator can be rewritten as this.

And if you allow individuals to self-select, then the expected

value of e1 bar minus expected value of e0 bar is the selection bias.

By using random assignment of individuals through treatment and control groups,

we have no systematic difference between these two groups, except for

the treatment itself.

By using random assignment, we want to have the difference to be 0 so

that the ordinary least square estimator is unbiased.

[SOUND]

In this lesson, we'll discuss the star experiment that was held in the state of

Tennessee in the late 1980s.

We're going to study the effect of class size

on the academic achievements of students.

The star experiment was done on a cohort of students in Tennessee, and

each student was randomly assigned within each school to either a small class,

regular-sized class, or a regular-sized class with a paid aide.

The teachers were also randomly assigned to one of these three groups.

We're going to focus on small versus regular class size performance,

And the data is present in the star data set in the Ecdat library.

I've created a data frame, mydata, to focus on the small and

regular size classes.

And I have defined a new column, totalscore,

which is some of the math and the reading scores.

So next steps.

I'm gonna create indicator variables called small

class, foreign observation if any observation is for

a student who has been assigned to a small class, zero otherwise.

And boy=1 if the gender of the student is boy and

so on, for whiteother, and for freelunch.

So I've created these four indicator variables.

And now you should look at the mydata, you'll see these variables that I've

created appearing there in this structure command.

So one check for a random assignment is regress small,

which is the class size binary variable 01 on the other factors and

check if there are any significant coefficients.

If there is random assignment there should not be any significant coefficients,

because small is an indicator variable, v is a linear probability model.

And this is what we get.

So small is regressed on boy, whiteother,

total experience of teachers, and freelunch or not,

and we notice that random assignment did take place.

None of the coefficients were boy, whiteother, totexpk, and

freelunch are significant, and we also can't reject b0 being different from 0.5.

So students are allocated to a small class with a coin toss.

Here are some summary stats for the regular size classes where small =0,

and for small size classes where small =1.

And note that the total score values is 917.9 for

the regular size class, and 932.05 for the small class size.

Now, we're going to run a regression model of total score on the W variable small.

We're going to study the effect of class size.

So this is the results that we get.

The coefficient of small is 14.1.

You see the results of the previous slide, so here is the question.

So you have this value 917.94, so

what's the average total score for regular sized classes?

The answer is 917.94.

This is the same answer that we got in the summary stats.

So what's the average total score for small classes, where small = 1?

So we got to add b0 plus b1, we get 932.05,

which is the same value that you got in the summary stats for small classes.

So what's the difference estimator, b1?

So the difference estimator is 14.109, which is the coefficient of small.

So this is the amount on average that is added to a student's total score if he or

she were moved from a regular size class to a small class,

if everything else was constant.

Now I'm going to run a different model, and

I'm going to add teacher experience to the model.

So total school regressed on small and

teacher experience is the totexpk variable.

So I run it and I get these results.

In this case, if you look at the coefficient of teacher experience,

each additional year of a teacher's experience, on average,

adds 1.16 points to the total score attained by a student.

The difference estimator in this model is 14.21.

So the effect of small class size is the same as having

a teacher with 14.21 divided by 1.16,

roughly 12 additional years of teaching experience.

[SOUND]

In this lesson, we look at natural experiments.

These are not intentional randomized control experiments.

And we develop the difference-in-difference estimator

to get the treatment effects.

So natural experiment is a study from real-world conditions.

We're observing what's happened in those situations.

It's used to approximate what would happen in a Randomized Controlled Experiment.

In a natural experiment, the subjects who might be undergoing treatment are not able

to choose if they are in the treatment or control group.

This choice is made by an external agent, or a factor like weather,

a policy change, etc.

So we need to have some subjects or observations in a group that are treated

and others who are not treated in order for the natural experiment to work.

And then researchers compare the average change over time of the Y variable for

the treatment group to the average change over time of the Y variable for

the control group.

This comparison is called difference-in-difference.

And panel data is used to measure these differences.

Here are some examples of natural experiments.

A treatment that just happened, not intentionally designed as an experiment.

So a law that changed the tax rate for some subjects, but not others.

An IT-system that allows online orders to be picked up in some stores, but

not others.

A hurricane that hits a few stores among a large sample of stores in the state.

A mobile carrier implements an unlimited data plan in some cities, but not others.

Minimum wage is changed in one state, but not another.

State Inclusionary Zoning laws are enacted in some cities but

not others within the state.

Another concept that you have to be familiar with

when looking at natural experiments is this concept of counterfactual.

So many scholars have emphasized that when we want to estimate the causal impact

of a treatment, we need to compare the outcome with the intervention

to what would have happened without the intervention, that is, a counterfactual.

The control group needs to be such that it is more or

less similar to the treatment group.

If we can't establish counterfactuals,

it is impossible to estimate treatment effects properly.

Here is an example of a natural experiment.

This example comes from a research paper that my colleague

Professor Jeffrey, who wrote.

My thanks to him to illustrate natural experiments.

Say New York City lower sales tax for local stores.

Neighboring states do not change sales tax for local stores.

We want to estimate the difference in purchase behavior between New York City

and nearby states.

We use the sales in stores in New York City and nearby stores.

Lower sales tax could result in stronger tendency to purchase locally.

Lower Internet sales, if you want more information, look at this citation.

So we introduce the difference-in-difference estimator.

So consider time t1 and t2.

These are the situation where t1 occurs before the treatment and

t2 occurs after the treatment.

Let's call t1, before and t2, after.

We are measuring the average value of the dependent variable, Y.

So A, let that be the average value of the dependent variable for

the control group measured at time t1.

B is the average value of the dependent variable for

the treatment group measured at time t1.

C, average value of Y, for the control group at T2.

And D, average value of Y for the treatment group, measured at time t2.

So we have this table.

For the control group, the difference of

the average Y values at time t2 and time t1 is C- A.

For the treated group, we have the difference of

the average Y values at time t2 and time t1 is D- B.

The difference between these values is called difference-in-difference or

diff-in-diff.

So diff-in-diff in this case is (D-B)- (C-A).

It may be easier to see these in the form of graph.

So I have time on the x axis, Y on the y axis

So I have the values for the control group and

treatment group at before and after time periods.

And then I'm going to draw the B-E line which parallels the A-C line,

which therefore B- A is the same as E- C.

For the control group, the before and after values of Y are A and C.

And similarly for the treatment group, these values are B and D.

What is the treatment effect?

Is that equal to D- C?

Answer is no, the diff-in-diff is (D- B)- (C- A) which we saw in the table.

Now B minus A is the same as E minus C.

So the difference is equal to D minus E, which is the correct

difference-in-difference estimate of the treatment effect.

So how do we estimate the DID, or

difference-in -difference using regression?

For the New York City example, we create two dummy variables.

One is if the observation from a store in New York City,

NYC is equal to 1, otherwise it's set to 0.

After, if the observation of sales is in the after time period,

after set to 1, otherwise set to 0.

And then we have an interaction term which is a product of NYC and after.

We observe sales at stores in New York City.

Treatment and sale at stores in surrounding areas which are the control,

before and after New York City lowered its sales tax.

Here's the model, say simple regression model sales,

regressed on NYCAfter and the interaction term.

We can show these coefficients graphically.

So we have, before and after, and the values of b0,

b1, and then b2 and b3 shown in this chart.

So sales for the control group, a time before = b0,

since after = 0, and NYC term is = 0.

Sales for the control group, at time after = b0 + b2,

since after = 1, and NYC= 0.

Similarly, for the treatment group,

average sales before = b0 + b1, since after = 0, NYC = 1.

And for the treatment group, the average sales at the time after = b0 +

b1 + b2 + b3, since both the dummy variables are set to 1.

So here's the regression model.

We know how the difference values are computed.

And the difference-in-difference estimator,

the difference of the two differences and you get that equal to b3.

So b3 is the coefficient of the interaction term NYCAfter.

Here are some steps in doing natural experiments,

understand the treatment or the event that just happened.

Check if you can theoretically argue that this treatment appears as if it

were randomly assigned.

Check if there's a control group and a treatment group.

Check if the empirical evidence shows that these two groups are roughly the same

before the experiment.

And then you analyze the treatment effect using the diff-in-diff estimator.

In this module, we had several lessons on treatment effects.

We looked at correlation versus causality, we looked at selection bias,

we looked at randomized controlled experiments,

we looked at the star experiment.

And then we looked at natural experiments and

the difference-in-difference estimator.

Thank you.

[MUSIC]

[NOISE]