

module_stat4_lecture14

📅 Thu, 2/17 4:18PM ⌚ 15:30

SUMMARY KEYWORDS

p value, gdp, beta, crime rate, regression, connection, data, unemployment rate, scatterplot, relationship, significant, co2 emissions, interested, coefficient, equal, check, sufficiently, reject, t statistic, positive

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

The other type of statistical analysis that we commonly do in the social sciences is look at whether there is a connection between two variables. Right. I've talked earlier about whether there is a connection between the GDP of a country and its co2 emissions. Right? Does as a country produce more in terms of GDP? Does it also produce result in more co2 emissions? or other environmental impacts? Is there a connection between unemployment rate and crime? Does as unemployment rate go up? Does it? Do you also see more crime? So these are questions that social scientists are very often interested in. And we have seen so far how to look at connections like this, right. And we have, when I was talking about visual representations of the data, one thing that we did was a scatterplot. In the previous module, Professor Makran, also talk to you about looking at such connections by using regressions. So in this lecture, I'll quickly connect you the concept of regressions with a hypothesis testing. So, so far, in Excel, when I showing you scatter plot, this is the scatterplot between GDP and CO2 emissions for India between 1970 and 2018. Right, so the blue dots, those represent the scatterplot. And this line, the dotted line, this is the regression line between CO2 emissions and GDP per capita. Similarly, if one wants to look at the connection between unemployment rate crime, so what I've done is that I've taken data from the 50 US states in 2018, and looked on on the x axis, I've plotted their own employment rate, and on the y axis, I've plotted the crime rate. So this is crime rate per million residents. And the orange dots like these, right? Did they show you the scatterplot, between the two. So if we look at the orange dots, there seems to be a kind of a positive relationship between the two, right, because mostly, you see an upward kind of connection between the two. And if you were to do a regression line, which is this dotted line, it, you see that that slopes upward, which does seem to suggest a positive relationship between unemployment and crime, in the sense that where there is states with higher unemployment, they also seem to have usually higher levels of crime. Now, the question is that is this relationship? Is this a robust, strong, significant relationship? Or does it? Or is it just happened in that one year 2018. And if we were to look at other years, maybe the relationship may not be that strong. So how much confidence do we have that this relationship here is a positive relationship? So I'll connect this question, because this is an important question, right? And I'll connect it with hypothesis testing that we have been doing so far in this module. So the simplest form of regression is that you have Y equal to α plus βx . So

this is your independent variable. And this is your dependent variable. And what you're really interested in is this coefficient beta, because that gives you the connection between y and x. And what you are interested in is that is this positive is this negative or is there no connection? Is this zero. Now, once you have this regression, what do you can if you want to test for this right, you can check whether this beta is zero. That means there is no relationship versus if there is This relationship, that means is beta naught equal to zero. So this is what we have been doing in hypothesis testing, we'll be checking whether this beta is equal to zero VS beta is not equal to zero. So how do we use the tools that we have developed so far in this module to check for this? So what we'll be looking at is what's called What's the P values. Remember, we talked about p values thing a couple of lectures ago. And what you're really interested in is that if the p value is very small, if it's less than 0.05, then we know that beta cannot be equal to zero, it has to be beta naught equal to zero. In that case, we'll reject the null hypothesis. Similarly, if you look at the Zed statistic, and check whether it lies within this interval, negative 1.96 to 1.96. Right? If it does not, if it is outside this interval, right? In both of these cases, we're going to reject sorry, reject the h naught. And again, the basic idea was that, that if the null hypothesis is true, right, if beta is actually equal to zero, then we should not see observations very, very far away from it. The probability of seeing observations there, or there, right, that's what the p value is. And if that is less than 5%, then we reject this hypothesis that beta is equal to 0. So let's see what we need to do in Excel to check for this. Remember, again, what we are the question that we are trying to determine is that is this relationship between y and x? So CO2 and GDP? Are crime and unemployment? Is that relationship strong? Is that relationship significant? Right? So what we are checking for is whether that beta coefficient that you've got, is that close to zero, or is it not equal to zero. So now I'm in my Excel file. And this sheet here gives the data on Indian GDP per capita, and emissions from 1970 to 2018. So in particular, this column B gives the GDP per capita, and the column C gives the CO2 emissions, which is kg per capita. And here is the scatter plot using this data that you've also seen before. And now what I want to do is I want to run a regression using these two data. So to run regression in in Excel, that I go to data, I go here on the right, extreme right to data analysis, and I click on that, it gives me lots of different options. And choose regression. And I do okay. And what it will ask me is for the input range, input y range, and the input x range, remember y is the dependent variable. Right? So here, I'm interested in co2 as the dependent variable, right, so I select this. So this is my CO2. Now, let me go to the input x range. So that's my independent variable, which is GDP. And again, I input all of these. So that's my input y range and put x range, and I hit the confidence intervals. And I'm going to choose a 95% confidence interval, because that's the most common, right, you could choose other levels of confidence as well. And secondly, it asks you where do you want the output, so I would like to put the output in an empty place. So I've chosen P two, right, and I hit OK. Now if I go up here, so this is my P two, and this is where it gives the results of your regression. So what we're really interested in is this bit here, see, you have the intercept, you have the coefficient on the x variable, that's the beta, right? So in this case, the beta is point seven, six. So that means the slope of this line of this regression line is point seven, six. Now, if we're interested in looking at, at is this beta different from zero or not, right after taking into account its standard errors. So what we're going to look at is the p value. See, the p value here is 2.1, e to the power negative 31. So what that means is that it's 2.1 times 10 to the power negative 31. So what that means there are 31 zeros before this decimal point, right. So that means this p value is really, really, really small. Similarly, if you look at the, what they call the t statistic, this is similar to the Z-statistic, right? So this seat is 28.64. So this is way bigger than 1.96. So in this case, we're going to reject the hypothesis that this beta is close to zero. So in other words, this is a significant relationship between GDP and CO2 emissions for India in this period. So this is how you do how you check whether this relationship is a significant one. So let me do one more example. This is the relationship between crime and unemployment. In the US states, right? Again, I have here the

data on unemployment rate by state and the crime rate per 1 million residents. So again, let me go to data and let's do data analysis. We're going to choose regression. Okay? It's going to ask me for the input y range. So what I want here is y to be the crime rate. So this is my y, the dependent variable, it's going to ask me for the x, which is the independent variable, which is my unemployment rate. Again, I'm going to choose the 95% confidence interval, and choose where it's empty, that's p two, okay. Now, again, if I come to here and look at the coefficient on the x variable, see, it's positive, it's 11.77416. So that means this line here, the trend line here has a positive slope. So what it does seem to suggest is a positive connection between unemployment rate and crime rate. Now, is this significant is this beta sufficiently different from zero? I look at the t statistic. Remember, I said this is the same as the Z-statistic for this sort of sample, right? And it's way bigger than 1.96. If you look at the p value, it's 1.79, e to the power minus five. So that's this multiplied by 10 to the power minus five. So it's like 0.0000179, right? So it's sufficiently smaller than 5%. So both of these suggests that this coefficient beta coefficient here is positive and significant, right? So it seems to suggest a positive significant relationship between unemployment rate and crime rate. So this is how you connect the estimates or test whether the estimates that you got in a regression equation whether they are significant or not. In both our examples, it was positive that you would do the same test if it was negative, only in that case, you would compare the Zed value with negative 1.96. If it's sufficiently smaller than 1.96, then you reject the hypothesis that beta is equal to zero, okay, or the P value is less than 0.05. In both cases, then your beta coefficient is significant. But if it is, if the p value is greater than 0.05 then you cannot reject the hypothesis that the beta is actually close to zero and there is no real relationship. So this is how hypothesis testing is connected with the regression estimates that you