

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

Lecture - 47
Degrees of Freedom

Hello and welcome back to the lecture on Applied Econometrics. We are in module 2, and we have been talking about the regression equation and we are trying to explain different terms associated with the regression equation and regression table. So one term you will often come across is degrees of freedom.

And you will see or degrees of freedom when you are explaining F-distribution, when you are talking about F-test, you are talking about t distribution, you are talking about chi-square distribution. In so many places, the concept of degrees of freedom is very important. Of course, when you are explaining a regression equation, you need to understand what are the degrees of freedom.

Now interestingly, the concept degrees of freedom is not very clearly explained. And also you will see that for different types of tasks, for example, different error terms in the regression equation, you will have different degrees of freedom. Sometimes, the degrees of freedom is represented by n , sometimes it is presented by $n - 1$, sometimes it is represented by $n - k - 1$.

Now for the same degrees of freedom, the same term degrees of freedom, why do I have, you know different sort of, you know expression. So we need to kind of understand what this different expressions of degrees of freedoms are? And what exactly is the concept of degrees of freedom? And how do I actually you to sort of see all the different terms associated with the degrees of freedom together. So let us begin with that.

(Refer Slide Time: 01:48)

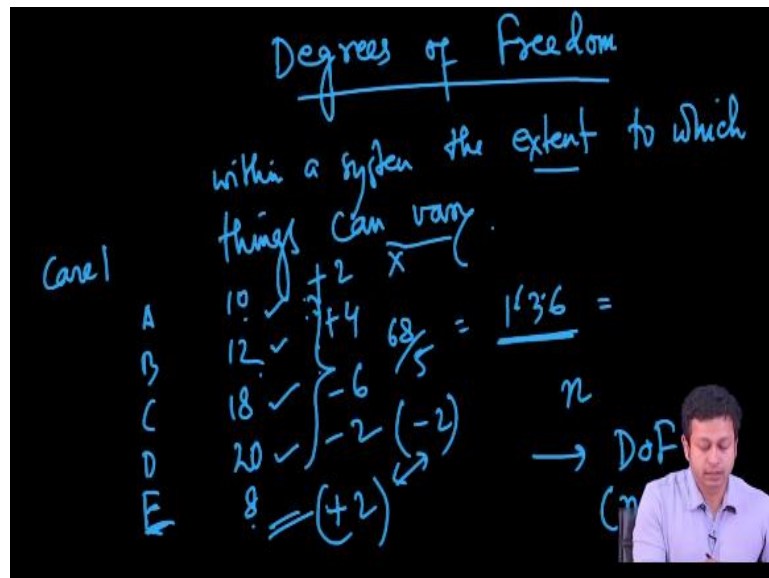
Degrees of Freedom

within a system the extent to which things can vary.

Case 1

A	10 ✓	+2	$\frac{68}{5} = 13.6 = \bar{x}$
B	12 ✓	+4	
C	18 ✓	-6	
D	20 ✓	-2	
E	8 ✓	+2	

$n \rightarrow \text{DoF}$
 (n)



So by degrees of freedom we mean that within a system, the extent to which things can vary, let us start with this generic expression, or generic definition of degrees of freedom. So what do I mean, within a system the extent to which things can vary. So let us say, we take a just a bunch of data, just a you know sample data or a population, sample drawn from a population, let us say of that of students, let us say.

And let us say we want to estimate the mean of the class tests, okay? And we kind of say, let us say we get the score of five students, and we get say A, B, C, D, E and we get the score out of say 20, somebody has gotten different scores, somebody has got 08, let us say. So all these different scores they have got. Now I want to estimate the mean.

And the mean is simply if I just take the mean is 22, 40, and then another 20 is 60, 68 by 5, which is 13.6. Now, 13.6. So now once I have the mean, suppose I want to sort of, you know change the different scores. I want to say, you know like, I want to bring in, you know I want to sort of, you know change the scores.

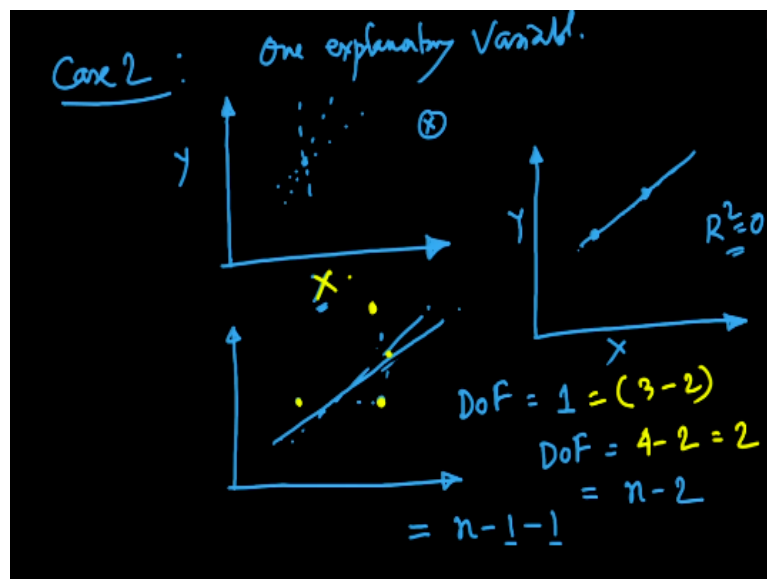
And if I want to change the score suppose, and my mean is fixed, when my mean is fixed, and if I want to change the scores, I can say vary, I can change the score of A and C and D and B and whichever I want. So I can change maximum up to 4 students. I cannot change the score of all the 5 students, because as long as I can change the 4 scores, I have and I keep one, you know like one entity where I cannot change.

So all the changes that happened in the first four scores will be kind of taken into account in the fifth, right? So if I say, do it +2, +4, -6, -2. So that would mean like +6 and -8 means what minus of, so net would be -2, and then I will add here +2 and this we will adjust. So essentially, the mean will remain same by keeping this one observation constant. So it can be any observation.

It can be A, B, C, D, E. So the whole idea is that I have to keep one observation constant. So that is why in this case, if I have say n observation, my degrees of freedom is going to be $n - 1$ because I have to keep one observation constant. Now it is true for not just me, if you want to estimate variance in the same way, you can actually see that your degrees of freedom is going to be $n - 1$.

So just reflect back to the generic idea the within the system, the extent of which the things can vary. So I can vary $n - 1$ things, not all n things as long as I am only concerned about the properties of the distribution. So mean invariance and so on and so forth. So this is how I will understand my first, this is my first understanding about the degrees of freedom. Now let us say, we will try to understand degrees of, let us say I will give a number. So this is my case 1.

(Refer Slide Time: 05:18)



Now let us say we will try to understand other concepts of degrees of freedom and how we will see the different, there are different expressions for different degrees of freedom. So now let us say in the previous example, we had only this observations n .

Now suppose I have apart from n I have k , I have k means I have also the variables included, right?

Now when I use when I will try to understand degrees of freedom with respect to both n and k , how do I do that? Let us say, let us again go back to our basic simple linear regression, okay. So let us say I have only say let us say I have y and x and I have only one data point, just one data point. Now if I ask you to draw a regression line, what will happen? Will you be able to draw a regression line?

Essentially we will not be able to draw a regression line because you know this you know infinite number of lines can pass through this point. So you will never be able to draw any particular regression line. So this is cancelled, this is not a possibility. You cannot basically draw a regression line with one point. So we can actually ask you this question. So what is the minimum number of data point you need to draw a regression line?

So I can have like okay, let us say let us increase the number of points from 1 to 2, okay. And if I have say two data points, with two data points, how many lines we can draw and it is very simple question to answer. We can have only one straight line with two data points. You can never have more than one straight line using two data points. So if you have a fixed straight line, so that basically denies the requirement of any regression equation because it is fixed.

So you have nothing to, you know no stochastic component. You have nothing to minimize, no error. So your line is set and your R-square value is going to be 0, because this line will have no explanatory power. Now so this is again, you are not you know done with two data points, if you want to run a regression line. Suppose you have three data points now.

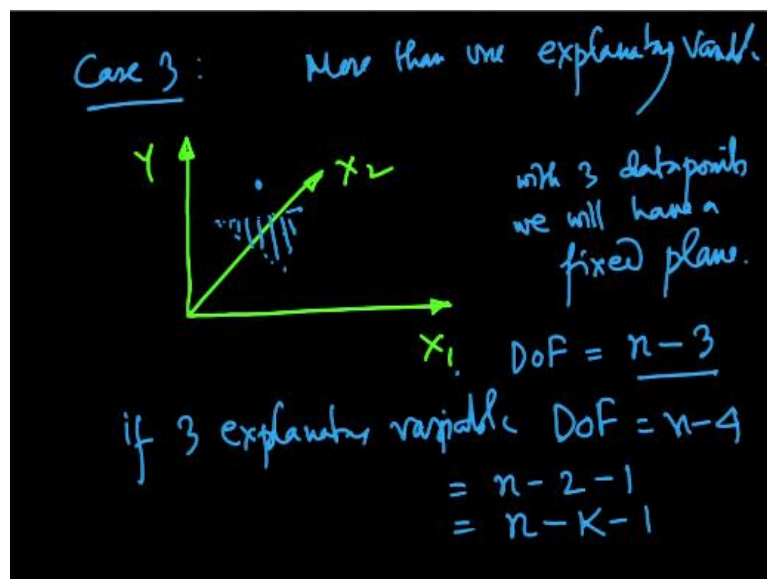
Suppose you have three data points and I can again draw this, if I have three data points. So now I can have something like a regression line, and this line can vary. And I can actually minimize the, you know errors, and I can actually get R-square value, which is between 0 and 1. So when that is the case, when with three data points with three data points, I can have a regression line.

And why is it, why I managed to have a regression line because this one additional data point, it has given me the freedom, it has given me the freedom to sort of draw the line where I can actually minimize the error. So because I have added one data point, my degrees of freedom here is 1 and this is how we will understand the concept of degrees of freedom.

Again, go back to the definition, to which things can vary. Now that I have added one data point, my degrees of freedom has become 1. Suppose I add two data points. Okay, so when I add two data points, I have my degrees of freedom, my degrees of freedom will be 2, sorry. Here sorry two data points means 4 data points. So I had 2 here, let me use a different color. I had 2 here already. And I have one more.

I had three actually. And I have added one more. So the moment I have four data points, my degrees of freedom will be $4 - 2$, which is 2. So initially it was $3 - 2$ is 1, right? So that is, you know from when I have one explanatory variable x is equal to 1 here, I have only one explanatory variable. Now I have more than one explanatory variable.

(Refer Slide Time: 08:59)



Let us say, I use a different page here, case 3. So here I had say, one explanatory variable. And here is more than one explanatory variable. So when I have more than one explanatory variable. Let us say I have now two explanatory variables, let me use

a different color. I have two explanatory variables. And this is let us say one x is this, another x is let us say x_1 , this is x_2 and this is my y , alright?

Now if you think that I want to, what would be the number of data points you need to sort of draw a regression line. Now here you have 3D plane 3D, you know three dimensional space and in a three dimensional space what you can have is a plane, okay, what you can have is a plane. So let us say and to draw any plane you need at least three data points, right?

So as long as there are three data points you have, it will give you a fixed line. With three data points, a fixed plane. With three data points you will have we will have a fixed plane. So to introduce any sort of variation, you need to have at least another data point. You need to have at least another data point. So that is at least you need to have four data points.

So the moment you have four data points you can introduce some variability. You can introduce some variations and that is when you can actually get a regression line. As long as you have only three data points you cannot get a regression line because this plane is fixed, right? So then I can write, from here I can actually write the degrees of freedom is going to be say n data point minus 3 here, okay.

Now let us say you now try to sort of you know think forward from here. So what is happening? Previously we have seen, we actually subtracted degrees of freedom is essentially $n - 2$ we have done. Here what we are doing? Degrees of freedom is $n - 2$, okay. And all that I have changed is the number of variables. So here I have two dimensional space and here I have three dimensional space, right?

So how exactly it is playing a role here, okay. So we have to see that or here I have one explanatory variable. Here I have two explanatory variable. If I have say, another dimension, we can kind of think forward as degrees of freedom for a say, let us say, if three explanatory variable, degrees of freedom is going to be $n - 4$.

And you know it is, you can actually try to sort of, you know like to conceptualize, I will not say visualize. And that is because you have another plane, and you have like,

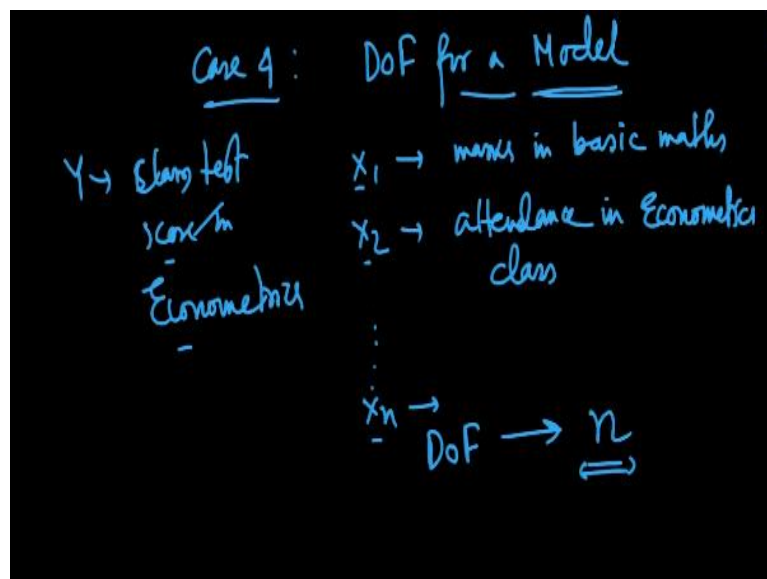
you need at least four data points to have the first you know to construct one plane. And you know like, and then you will have, you need to have at least five data points to get a some sort of freedom for, some sort of freedom to get a regression line.

So essentially, if you actually see these cases, all the cases here, so initially I had $n - 1$ when I had only n observations. Here I have $n - 2$ when I have one explanatory variable, or we can write $n - 1 - 1$. So this is the one constant that has come for x and one constant that has come for the, this constant, for the number of observations, right? And here I have say $n - 2$, 2 for two explanatory variable, minus 1 right, $n - 2 - 1$.

So $n - 2 - 1$. So essentially what I am, you know sort of the generic formula that I am going to get is $n - k - 1$ okay. So the moment I introduce number of different variables, or explanatory variables, I have this k term, I have this k term you know included in my degrees of freedom. So essentially, with the increase of number of variables, explanatory variables, my degrees of freedom is decreasing.

My degrees of freedom is decreasing, right? So that is how we will understand the concept of degrees of freedom when I have more than one explanatory variable or you know when I have explanatory variables, alright. Great.

(Refer Slide Time: 13:37)



So now let us say, DoF degrees of freedom for a model. So when I am creating a model, what happens, when I am creating a model? So I have say, some data. So

again, let us go back to the first example where I have all the student data of their scores. And I try to understand the variability, the variability of the scores with respect to different explanatory variables.

Now when I include these different explanatory variables, and I am talking about the model, what I am doing is I am actually including more and more dimensions, which will help me to explain the variability in that data, try to understand the difference very carefully. Here what I am trying to do is I am trying to get the degrees of freedom for the model.

And the moment say I include one, my say, let us say my x_1 , let us say your, you know marks in basic math. And my y variable here is the say score, class test score in econometrics, test score in econometrics. x_1 marks in basic maths and let us say x_2 is your attendance in econometrics class. And you are trying to understand the variability to the extent you can explain the class test score.

Now when you have these different explanatory variables, each explanatory variable is actually giving you the freedom to explain the variability. So with n number of variables, you are explaining the variability in n number of ways. So that means for when you are considering the degrees of freedom for a model your degrees of freedom DoF for a model is going to be n , okay.

So we actually essentially we see that, again refer back to the definition. Within a system the extent to which things can vary. So within your, you know the system of the class test score, these are the different ways you can actually vary the score or you can actually see the score, okay. So this is why it is degrees of freedom is equal to n . And we will actually see this at least this three different cases of degrees of freedom you know when we do the regression table.

(Refer Slide Time: 16:06)

DoF
i) $n-1$
ii) $n-k-1$
iii) n

So let us say the DoF, so three expression we got. One is $n - 1$. Second is $n - k - 1$. And third is n . And we will see we have already explained this situations where these three things three expressions of degrees of freedom is used. And we are actually going to see in the regression table how these are relevant. So with this we end the lecture on degrees of freedom. Thank you.