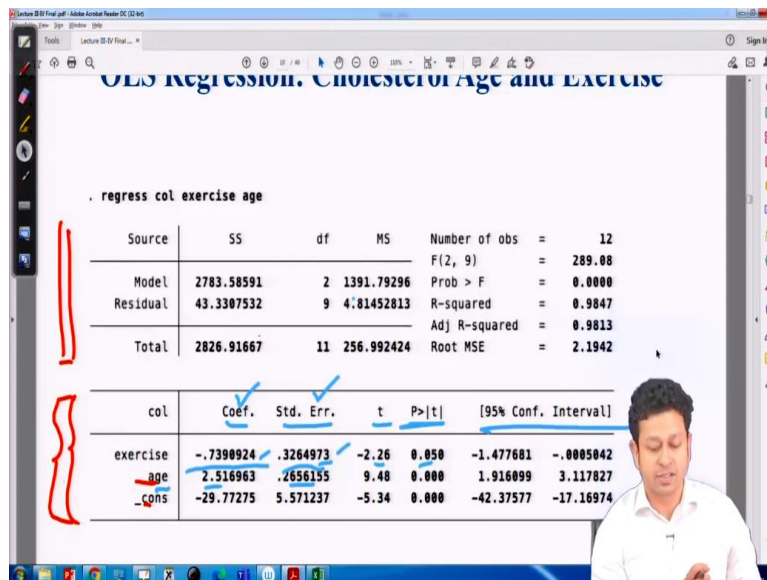


**Applied Econometrics**  
**Prof. Tutan Ahmed**  
**Vinod Gupta School of Management**  
**Indian Institute of Technology - Kharagpur**

**Module - 6**  
**Lecture - 51**  
**Regression Table (Contd.)**

Hello and welcome back to the lecture on Applied Econometrics. So, we have been talking about very interesting topic called the regression table, and we are trying to understand the different components of regression table.

(Refer Slide Time: 00:35)



OLS Regression: Cholesterol Age and Exercise

```
. regress col exercise age
```

Source	SS	df	MS	Number of obs =	12
Model	2783.58591	2	1391.79296	F(2, 9) =	289.08
Residual	43.3307532	9	4.81452813	Prob > F =	0.0000
Total	2826.91667	11	256.992424	R-squared =	0.9847
				Adj R-squared =	0.9813
				Root MSE =	2.1942

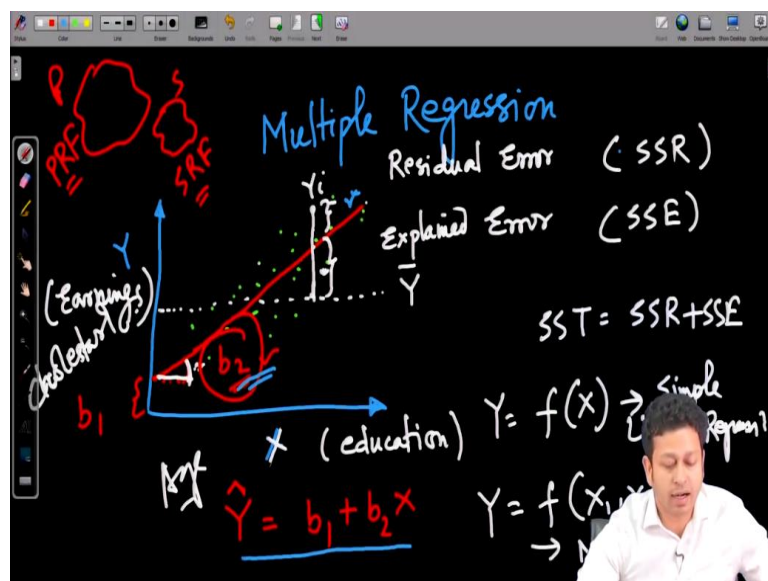
  

col	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exercise	-.7398924	.3264973	-2.26	0.050	-1.477681 -.0005042
age	2.516963	.2656155	9.48	0.000	1.916099 3.117827
_cons	-29.77275	5.571237	-5.34	0.000	-42.37577 -17.16974

So, in the previous lecture, we explained the first part of the table, which is the ANOVA part of the table, and we said that in the next lecture, we are going to explain the hypothesis testing part of this table. And this is the second part that we are going to talk about. And here we are actually, what we are going to do is basically hypothesis testing. Now, hypothesis testing of what basically? That is what we are going to see in this lecture.

So, we have all these different explanatory variable like exercise, age, and we also have the constant term. Now, what we are trying to do here? So, let us try to actually look back, what we did when we actually plotted the;

(Refer Slide Time: 01:12)



You remember that; let us say we are regressing with 1 explanatory variable X. And we get some line which will somewhat approximate all the points that we have on the plane. And we are not talking about multiple regression now, let us say it is just the simple regression with 1 explanatory variable. And we get a estimate of this beta 2 coefficient, that which is basically the slope of that line.

Now, remember, what we are doing here is that, it is a bunch of data that you got. You took a sample from a population. You never get the entire population; that we already know. So, there is a population here. And all that you can do is, you can take a small sample out of that population. And when you are running a regression equation, you again have to depend on the same mechanism where you have a small amount of data taken from the population, which is called sample.

And you run a regression line or you basically get a regression equation using that sample data. So, this particular regression equation that we see here is essentially, is a regression line that you obtain using those sample data set. So, you can also call it; sometimes it is called sample regression function. So, sample regression function, SRF, this. And whereas, the actual regression equation that you might have or that you actually have, which you will never know is a population regression function.

We can just use these terms sometimes. So, here, what you get is that; so, here you are essentially getting the regression equation from the sample. So, the beta 2, we normally denote small b or b; whereas, for population, we actually denote as a beta. So, for sample, is

small  $b$ ; whereas, for population, is  $\beta$ . So, here, what we get is, this  $\beta$  coefficient which is an estimate; so, the  $b_2$  is an estimator of the  $\beta$ .

So, you never know  $\beta$ , but you can know  $b$ , right? And if you can know  $b$ , you can estimate  $\beta$ . So, that is the mechanism that we are trying to follow here. Now, what we are trying to do here is, we are trying to see if that  $b$ , the coefficient that I obtained from my sample, whether that is able to explain the population parameter which is  $\beta$ . So, that is why we need to do a hypothesis testing whether my  $b$  is able to sufficiently explain  $\beta$ . So, how do I actually do that? So, essentially, we create some statistic.

(Refer Slide Time: 03:55)

The image shows handwritten notes on a blackboard, divided into two columns by a vertical red line. The left column is titled 'Known Population Parameter' and the right column is titled 'Unknown Population Parameter'. Both columns show the general formula for the test statistic and the specific formula for the null hypothesis  $H_0: \beta_2 = 0$ .

Known Population Parameter	Unknown Population Parameter
$Z = \frac{b_2 - \beta_2}{SD(\text{known})}$	$t = \frac{b_2 - \beta_2}{SE(\text{unknown})}$
$H_0: \beta_2 = 0 \implies Z = \frac{b_2}{SD}$	$H_0: \beta_2 = 0 \implies t = \frac{b_2}{SE_{\text{unknown}}}$

Let us say it is a Z statistic. So, Z statistic here, if I have, say, if I am using a normal distribution where my population parameters are known, I normally write  $b_2$ . So, this is my slope, and this is the population. If I would have known the population parameters, so, the  $\beta_2$  would have been the population parameter; and you have the standard deviation. So, that is how you would have gotten your Z value.

Now, your null hypothesis is what? Your null hypothesis is saying that the  $\beta_2$ , actually it does not have any significance. So, your  $\beta_2$  is actually;  $H_0$  is saying, your  $\beta_2$  is actually 0, because that is what you are assuming; your model does not have any, your variable or that explanatory variable does not have any power to explain the variations, right? That is your null hypothesis.

And if your  $\beta_2$  is 0, so, then your Z is going to be  $\beta_2$  by SD. So, that is what your null hypothesis is saying. Now, you actually create your confidence interval, you can take like whatever alpha value you want to take, 5%, 1%; usually we take 5%. For 5% confidence interval, you get your Z value, and you create that confidence interval. And you see, if your Z value is going beyond the confidence interval, you reject your null hypothesis.

So, that is how you actually do, basically the hypothesis testing here. Now, this is for known population parameter. But this is an ideal case, I mean, actually, what happens in reality, you never know the population parameter. And then, basically, what we have is that unknown population parameter. So, when we have unknown population parameter, you cannot really use a Z statistic. Let me use a different colour for this part.

So, for unknown population parameter, what we use, a t-statistic. So, t-statistic, where you actually use this; again, the same; idea remains the same, the only thing that differs here is the standard error term. So, where you have the standard error, which you actually estimate from the; basically you can write the unknown, you essentially estimate it from the sample. Here also it is a standard error, but it is a known. We will talk about standard error.

So, here,  $H_0$ ; once your  $H_0$  is  $\beta_2$  is equal to 0, your t-statistic is going to be  $\beta_2$  by SE unknown. Now, we will essentially, this is what we are going to do. So, we have this corresponding t-statistic for each of these different  $\beta$ . And we will try to see whether they are significant or not. So, we will do the significance test for each of these different  $\beta$ . So, let us again go back to our regression equation or the table.

So, what we see here is this. So, we have different items. So, one, we have our coefficients; we have our standard error; we have this t-statistic; we have this P-value; and then we have the confidence interval. So, first, let us talk about the coefficient. So, coefficient, by now, we have some fair idea about what a coefficient is, and we have actually explained this coefficient.

We have these coefficients here, where we have; so, here my coefficient is  $\beta_2$ , which is like explaining how much my, this line, this fitted line is actually able to explain my Y variable, right? Now, so, all these different values, essentially, what they are saying is that; so, if I take

exercise here, so, what I will understand is, so par unit, whatever unit I have taken, par unit increase in exercise is going to decrease the cholesterol level by 0.73 unit.

So, that is what it explains, like any straight line equation may explain. So, it is essentially the same thing. To what extent the variation of exercise is going to impact the variation of cholesterol level. Similarly, for age also, it says that par unit increase in age will increase the cholesterol level by 2.5. So, essentially, it is again like the straight line where your par unit change in X is, you are basically measuring how much it is going to change the Y.

So, it is 2.5 times. So, that is the coefficient thing that talks about. So, with multiple regression explanatory variables, you have like, on multiple dimensions you can think, all these different explanatory variables are increasing. The Y, or basically, influencing the Y in different dimensions. Now, that is fine, but, so, just that you have some value, does that mean that you have to end there? Of course not.

You have to see whether these impacts are really significant, if they really are meaningful here. So, how you really understand that? And for that, you actually need to consider all these different standard error, t-statistic and P-value. Let us look at the other terms. Let us first look at what is the standard error. So, standard error is something that is a little tricky here. And we have to keep in mind what these standard errors are.

So, standard error, it sounds pretty much like standard deviation, but we say it is standard error and not standard deviation. So, why is that? You know, is there any specific reason why we say that? We will just talk about this. And how we really measure the standard error? Here we have actually explained this case of unknown thing and known thing; known standard deviation, unknown standard deviation, and how they are; so, we will try to see how they are related to the concept of standard error.

**(Refer Slide Time: 10:18)**

Standard Error

Standard Deviation  $\sigma_x = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$   
(population)

$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$       SEM :  $s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$

So, what happens? Let us say, standard error: So, what is happening here? Let us try to explain. So, standard deviation is something is very straightforward. We know like how we actually get the sigma, right? Sigma is nothing but 1 by n summation  $X_i$  minus  $\bar{X}$  square. Now, that is fine. That is the standard deviation you got. But then, you have to get; this is the standard deviation; let me actually write down; deviation for; so, when you already have all the values  $X$ , you know all the  $X$ 's, you basically calculate the standard deviation.

We usually denote population standard deviation by sigma. So, let us say we are talking about the population. Now, when we actually talked about this concept of estimation from sample, usually, we know that it is not that easy to get the population standard deviation. So, what we do here? We actually have to estimate from sample. And when we use the sample, we actually denote it as  $S_X$ .

And when we denote it as  $S_X$ , what we do is, we actually use 1 by  $n - 1$ , for sample, because there are some approximations involved. And we explained where these approximations are coming from. So, we write this. And we use this approximated value to the population standard deviation, because that is how we conceive the whole idea, how  $S_X$  could represent the population standard deviation now. There is another trick here.

So, when you are doing this estimation thing, so, you actually want to estimate, you actually have your estimator; the mean is fixed, right? So, you actually want to draw different samples and you want to draw, you want to get their standard deviations and you want to actually get the standard deviation of the means, not the individual item. And to do that, to get the

standard deviation of means, you actually have to draw a different samples time and again from the observation.

But if you just remember how we have progressed in this development of getting this regression equation or the minimisation of error term, we really have not done something where we actually draw the sample time and again. So, this is an approximation that we use. And the moment we do not do that, like taking the observations time and again and actually calculating the standard deviation of the means, we actually take all the observations in one go.

And what we do is, we essentially; I mean, it is a simulation that statisticians have come up with; is that you just divide it by square root of  $n$ . And if you do that, so, then, that will actually give you the standard deviation; so, the standard error for mean. So, is what we write it as  $S_{\bar{X}}$  is equal to say  $\frac{1}{\sqrt{n}}$  into  $\frac{1}{n} \sum (X_i - \bar{X})^2$ , or let me;  $\frac{1}{n-1} \sum (X_i - \bar{X})^2$ .

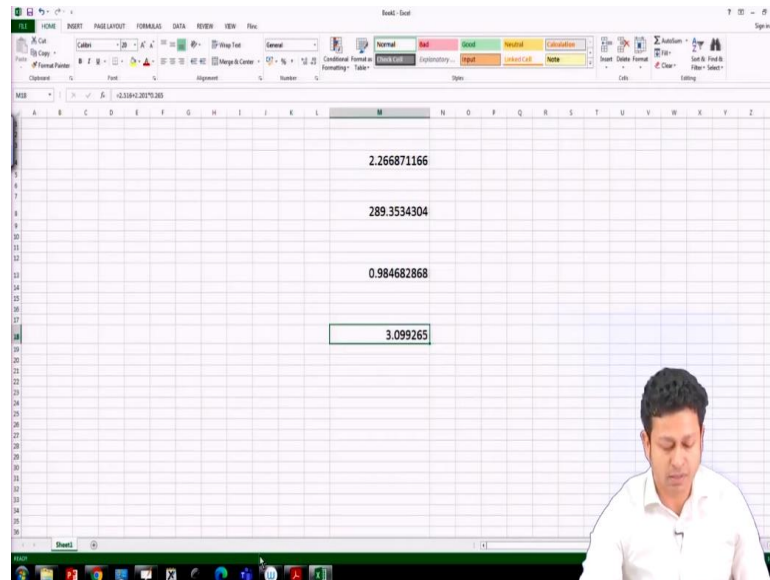
So, this is what we call as standard deviation or a standard error of mean. So, we are calculating the standard error for the mean. So, that is what we are more interested in, standard error of mean; and this is how we actually estimate. So, when you are estimating from a sample, we simply divide it by square root of  $n$ . And similarly, for here also, this is nothing but, is  $\sigma_X$  by square root of  $n$ .

So, we essentially divide it by square root of  $n$  to come to this approximation; just remember this. So, what we are actually calculating there in the regression table is that, we are actually getting this standard error of mean. And since we are calculating from samples, we do not have the population parameters readily available with us. So, we divide the whole thing with  $\frac{1}{\sqrt{n}}$ .

And all this expression that you have seen there are essentially this, where we have this  $\frac{1}{\sqrt{n}}$  into square root of  $\frac{1}{n-1} \sum (X_i - \bar{X})^2$ . So, that is how we have got the standard errors. So, these are the two; so, we have explained these two answers, coefficient and standard error. Now comes our  $t$ . Now,  $t$ , we have already explained;  $t$  is essentially, given our hypothesis testing,  $t$  is essentially  $b_2$  or the coefficient by standard error.

So, if we do that, if we go back there and if we actually do that, so, what we will see is; so, t will be ratio of this coefficient and standard error; so, ratio of these two. So, let us actually do that. So, if the coefficient for exercise is 0.739, 0.739 by 0.326 it should be.

**(Refer Slide Time: 16:19)**

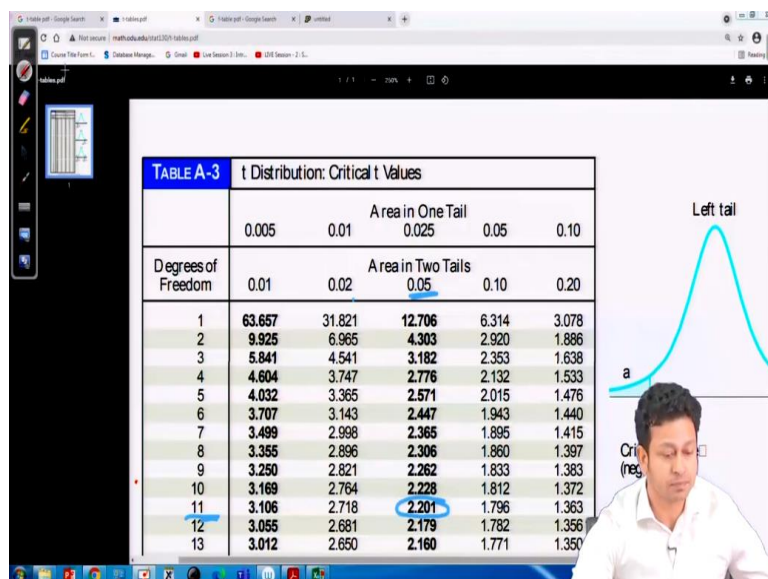


Here, 0.739 by 0.326. So, 2.26, something like that. And since I have a negative sign here, for exercise, you will also have a negative sign for your t-statistic, right? So, that is how we get all our different t-statistic. So, you get for different coefficient as well as the constant term. So, you basically, we just take the ratio of coefficient and the standard error. So, that is how we got the t-statistic.

So, you have got that. Now you got all the 3 terms you need. Now, we come to the fourth item, which is the P-value. And that is very important. Like in the previous case, we have seen how to get the P-value for F-statistic, we will get in the same manner, the P-value for the t-statistic as well. Only thing that will differ is that, in this case, we will see a t-table instead of a f-table. So, let us do that. So, let us actually try to find how a t-table looks like. And I actually have done that for you, I actually got the t-table here. So, we have to see how to see a t-table.

**(Refer Slide Time: 17:39)**





**TABLE A-3 t Distribution: Critical t Values**

Degrees of Freedom	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350

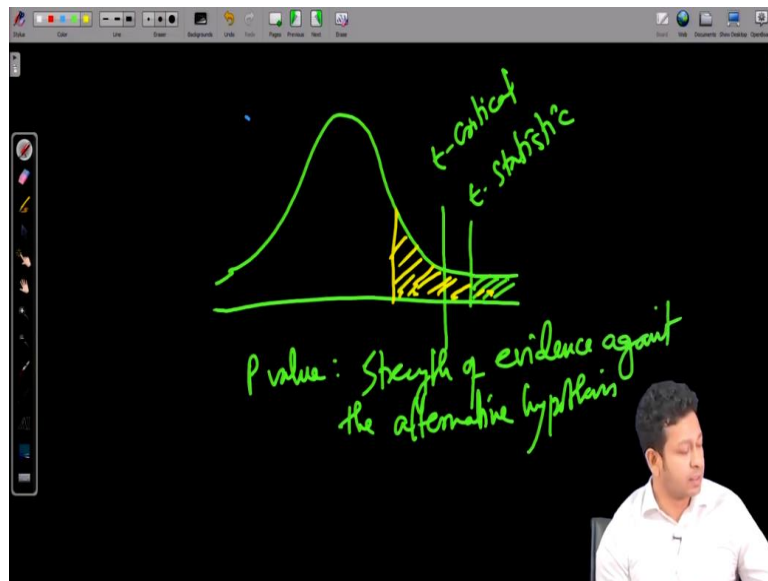
And for t-table, you will see, here we are talking about only 1 degree of freedom; like 1, the degree of freedom in the column, not in the row like we had for the F-distribution. And that is because, in t-distribution, we are only talking about; so, we are not taking a ratio of variance or anything; we are only thinking about one sort of distribution here; and we are just taking into account the total number of observations, we are not fitting any model, we are not actually calculating residuals.

So, you do not have to get into  $k$  or  $n - k - 1$ . All you have, the  $n$  number of observations. So, we know that it is an estimation problem, so, we have to have 1 thing fixed, 1 parameter fixed, which is because of that, we will have degrees of freedom equal to  $n - 1$ . So, here, in our case, we have how many observations? We have 12 observations. So, my  $n - 1$  would be 11, and the degrees of freedom.

So, it is a 2-tailed test, because we are just trying to check if the equality holds; we are not talking about greater than, less than anything; so, we are basically doing a 2-tailed test. So, when you are doing 2-tailed test, we are; and alpha is 5%, so, we have this. And this is 11. So, what we get here is, the corresponding t-value is essentially 2.201, corresponding t-value is this.

Now, what I will do is, I will draw a t-distribution and I will actually try to see where my observed t-value falls vis-a-vis the critical t-statistic value, okay? So, let us do that.

**(Refer Slide Time: 19:32)**



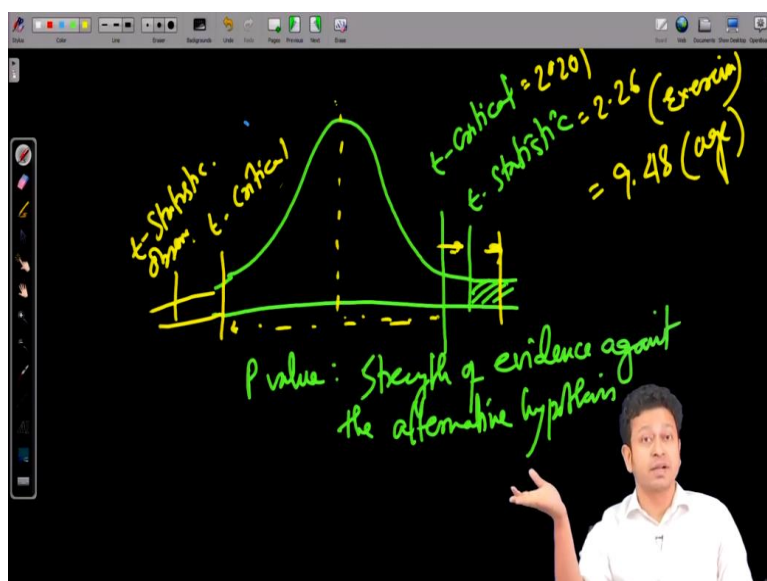
So, I will draw the t-distribution. We know that t-distribution is somewhat similar to, it looks somewhat similar to normal distribution though it is not a normal distribution. And I have a t-critical here. And the moment my t-statistic is here, if it is depending on where it falls, if it falls right to the t-critical, so, the P-value here would be very low. And when the P-value is very low, we have explained how we can explain the P-value.

So, if it is very low, then we say that the null hypotheses is rejected. And the reason is, like we can explain in different ways actually. We will again recap the P-value. It is very critical in all the things that we are doing. So, the moment I have a low P-value, it means that P-value actually defines the strength; I can actually use this definition; strength of evidence against as the alternative hypothesis.

So, if my P-value is very high, let us say I have my P-value, like my t-statistic here, and my P-value, like, area right to that t-statistic is very high. So, that means, I have lots of evidences that are going against the alternative hypothesis. It means that my model is actually not explained by my alternative hypothesis, but it is rather explained by other randomness; so, which is essentially kind of telling that your alternative hypothesis, you can reject.

So, essentially, your strength of evidence against alternative hypothesis is very high. And that is why you do not reject the null hypothesis. You say that, your null hypothesis, whatever it is, you are saying that it is, you cannot reject that. So, this is something is the P-value or other way you can think that, like, if you do not; this is one explanation, this is one way of looking at it.

(Refer Slide Time: 21:49)



Other way of looking at P-value is that, well, if you have, like the evidence that you have got is like, which is telling about the alternative hypothesis, it is so extreme here, it has come so extreme that, say from your null hypothesis value that even after giving so much of, like the width of the interval, your evidence is even falling beyond that. So, the moment your evidence is falling even beyond that, you say that this is not something that you can say that it is close to null hypothesis.

So, your alternative hypothesis is far away. So, you are essentially, you are bound to reject the null hypothesis. So, that is where you actually, this is different ways of looking. This is kind of feeling how the P-value actually talks about. So, we basically calculate all this, this t-critical, and we see the t-statistics. So, now, we found our t-critical to be 2.201. So, our t-critical is 2.201. And in my regression table, what I found my t-statistic to be 2.26.

So, it is 2.26 for exercise. For exercise, you have 2.26. So, it means that it is actually falling somewhat right to the t-critical. And the moment it is falling somewhat right to the t-critical, we will say that I will reject my null hypothesis. And how do we actually get the P-value? Here you see the P-value is 0.05. We can actually go to the t-statistic table again. And to make some sense of this, so, for 11 degrees of freedom, we are talking about area in 2 tails, so, it is 2.201.

So, we do not have value for all the different areas. We have like area for 2%, 5%. So, we have 2.201, 2.7 if it is 2%. And it is 1.79 if it is 10%. So, essentially, if you see that what

your P-value or the t-statistic what you observed is 2.26, which is coming very close to this one. And that is why you kind of say; I mean, it could have been a little more here. So, that is why you can say that; your P-value is like approximately 0.05; it could be like 0.046 or something, 7 or 8; but we do not have all these values, so, we approximate it as 0.05 in this.

Now, for age, your t-statistic is 9.48. Now, you already have calculated your t-critical. So, when you have your t-critical and you have to, your calculated t-statistic is 9.48 for this term is your age. So, it is again falling far right to your t-critical. And that is why we can definitely reject that. Even for the constant term, you have your t-critical value coming out to be 5.34. And it is again like on the other side of the t-distribution, you can actually show.

So, if you have like t-critical here and your observed t-statistic is coming here. So, this is how we actually calculate the P-values and we take a call on whether to reject or not, the coefficients, whether they are significant or not. So, we are done with all these different 4 terms. So, here, now we have done the coefficient.

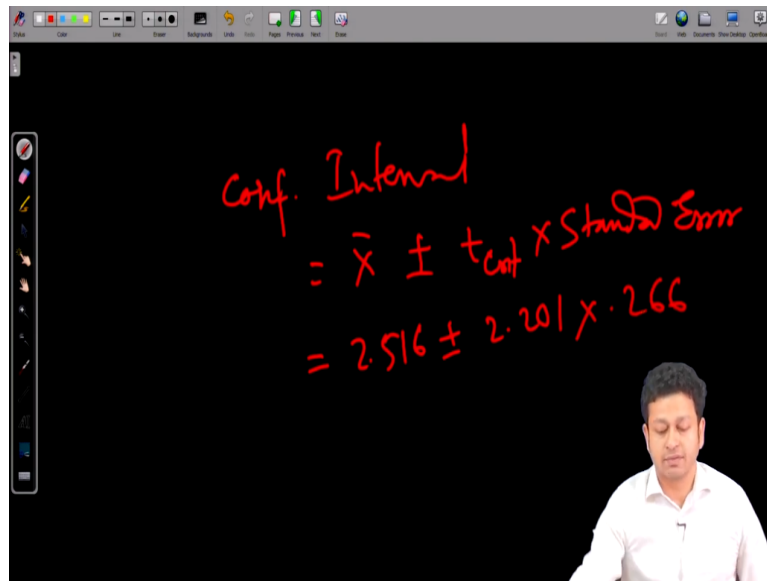
**(Refer Slide Time: 25:54)**

Degrees of Freedom	Area in One Tail					Area in Two Tails				
	0.005	0.01	0.025	0.05	0.10	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078					
2	9.925	6.965	4.303	2.920	1.886					
3	5.841	4.541	3.182	2.353	1.638					
4	4.604	3.747	2.776	2.132	1.533					
5	4.032	3.365	2.571	2.015	1.476					
6	3.707	3.143	2.447	1.943	1.440					
7	3.499	2.998	2.365	1.895	1.415					
8	3.355	2.896	2.306	1.860	1.397					
9	3.250	2.821	2.262	1.833	1.383					
10	3.169	2.764	2.228	1.812	1.372					
11	3.106	2.718	2.201	1.796	1.363					
12	3.055	2.681	2.177	1.782	1.356					
13	3.012	2.650	2.160	1.771	1.350					
14	2.977	2.624	2.145	1.761	1.345					
15	2.947	2.602	2.131	1.753	1.341					
16	2.921	2.583	2.120	1.746	1.337					
17	2.898	2.567	2.110	1.740	1.333					
18	2.878	2.552	2.101	1.734	1.330					
19	2.861	2.539	2.093	1.729	1.328					

So, we have done the coefficient part; we have done the standard error part; we have understood t; we have understood the P-value. Now, last part is the confidence interval. So, once we do this, we are through with all this regression table. So, how do I calculate the confidence interval? And that is what we have to see. So, here, usually we take 95% confidence interval or alpha is equal to 0.05 or 5%.

So, when I am creating a 95% confidence interval, so, what I do is, essentially, I get a t-critical value for 95% confidence interval. So, we have already seen that. What is the t-critical we had? We had this t-critical is equal to 2.201. So, we have already seen that. The t-critical for 95% confidence interval, we will have, for my 11 degrees of freedom, so, this one, 2.201. So, what is the formula for my confidence interval? My formula for confidence interval is this.

**(Refer Slide Time: 27:03)**

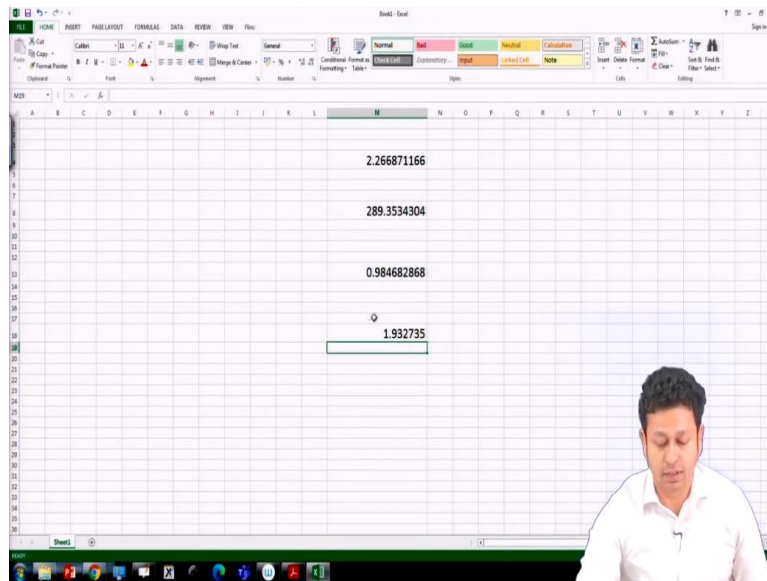


$$\begin{aligned} \text{Conf. Interval} &= \bar{X} \pm t_{\text{crit}} \times \text{Standard Error} \\ &= 2.516 \pm 2.201 \times 0.266 \end{aligned}$$

So, the way I create the confidence interval is simply X bar; t-critical into standard error. So, when I am saying standard error, you have in your regression table, is already calculated, you do not have to do anything about dividing by root n or anything. You just take the standard error and you get the t-critical, and you basically compute the confidence interval. So, what it will be?

So, your X bar, let us say we do it for age; so, 2.516; t-critical, 2.201; and the standard error that we had for this is 0.265 or 0.266 let us say; 0.266, okay? And if I compute this,  $2.56 + 2.201$  into; I have to use a standard error; is that 0.326. And it will give me; 2.516, 2.201, 0.265; oh, I have actually taken a wrong one; 0.265. That is going to give me something around 3.09 or 3.1, okay? So, let me just increase the font here. So, it is approximately 3.1, and that is precisely what you see, 3.1.

**(Refer Slide Time: 29:07)**



And if I just, subtract this, I will get the other side of the confidence interval which is 1.93; and it is 1.91 approximately. So, essentially, that is how we calculate the confidence interval. Now, I will ask you a question. So, what I am trying to do here? So, suppose I calculate the confidence interval and it is showing that the range is from, say for the first one it is -1.47 to some -0.0005; and for the second one 1.912; say, 3.11.

So, what it is really talking about? So, if you look at it carefully, what it talks about; say for example, for age; so, what it is talking about is that the confidence interval is actually on the right side of the 0; so, it is not containing 0. Even for the first one, it is not containing 0; or even for the third one, it is not containing 0. So, everywhere, it is excluding 0. So, what it means is, your null hypothesis was that the mean value of all this, the different coefficients that you are estimating, in null hypothesis was 0.

So, the moment your confidence interval is not including 0, it is excluding 0; so, that means, your estimate, the thing that you are estimating, that is actually not containing 0, but it is falling somewhere on the right or left. So, 0 is not an option for it. So, when that is the case, you call it that your alternative hypothesis is actually explaining your models; you kind of reject the null hypothesis.

So, in all the cases, you see that your confidence interval is not containing 0. So, that is what you see here. And that is what you need to remember. So, you will see that, of course, they will have a correspondence between the P-value and the confidence intervals. So, if your

confidence interval is not containing 0, your P-value is also going to be very low. So, that is about these 4 items.

So, one more thing I would ask you is that, why do we use a t-statistic and not a Z-statistic? So, we use a t-statistic and not a Z-statistic. And if you recall, what we started this lecture with is that, well, your population parameters are not known. So, you have to basically you have to estimate your population parameter from the sample observations, right? So, we use  $S_X$  instead of  $\sigma_X$ , right?

And since we always do that, because we never know the population parameters, we use a t-statistic instead of Z-statistic. So, in all regression table, does not matter what is the number of observations you have, you will always use a t-statistic for testing the significance of all these different explanatory variables. So, with this, I think we can end this lecture on regression table. So, that is that is the end of this lecture.