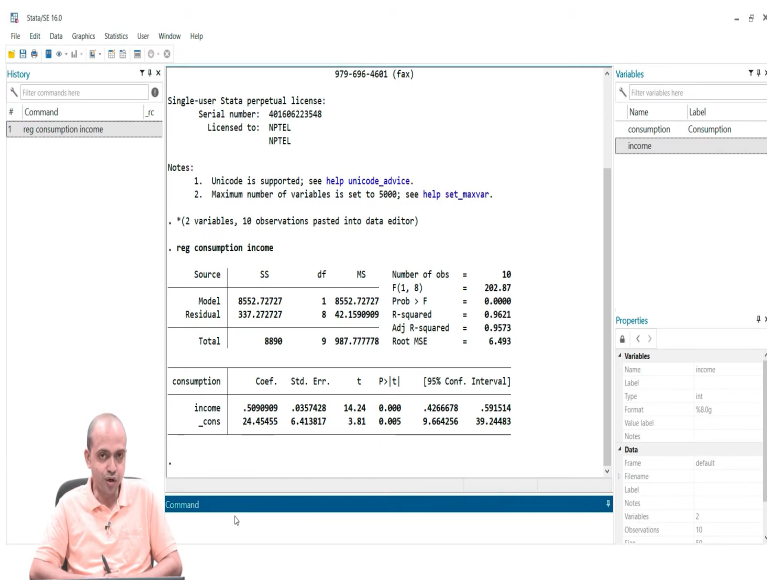


Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras

Application of STATA for hypothesis testing and introduction to multiple linear regression model Part - 1

Welcome to our discussion of hypothesis testing. So, yesterday we were discussing the theory behind the hypothesis testing, what is the actual procedure, what is the statistical mechanism behind the hypothesis testing that we are discussing yesterday. Now, today we will re-estimate our income consumption model and then we will see whether we can apply our knowledge of hypothesis testing to interpret the results that Stata is giving to us. So, we will re-estimate the model once again.

(Refer Slide Time: 01:09)



Stata/SE 16.0

979-696-4601 (fax)

Single-user Stata perpetual license:
Serial number: 40168623548
Licensed to: NPTEL

Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.
*(2 variables, 10 observations pasted into data editor)

. reg consumption income

Source	SS	df	MS	Number of obs	F(1, 8)
Model	8552.72727	1	8552.72727	10	202.87
Residual	337.272727	8	42.1590909		0.0000
Total	8890	9	987.777778		0.9621

Adj R-squared = 0.9573
Root MSE = 6.493

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
consumption						
income	-.5090909	.0357428	14.24	0.000	-0.4266678	-.591514
._cons	24.45455	6.413817	3.81	0.005	9.664256	39.24483

Command

Variables

Name	Label
consumption	Consumption
income	

Properties

Variables

Name	income
Type	int
Format	%10g
Value label	
Notes	

Data

Frame	default
Filename	
Label	
Notes	
Variables	2
Observations	10
Free	on

Stata/SE 16.0

File Edit Data Graphics Statistics User Window Help

979-696-4681 (fax)

Single-user Stata perpetual license:
Serial number: 48166623548
Licensed to: NPTEL
NPTEL

Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5800; see help set_maxvar.

*(2 variables, 10 observations pasted into data editor)

. reg consumption income

Source	SS	df	MS	Number of obs	=	10
Model	8552.72727	1	8552.72727	F(1, 8)	=	282.87
Residual	337.272727	8	42.1590909	Prob > F	=	0.0000
Total	8890	9	987.777778	R-squared	=	0.9621
				Adj R-squared	=	0.9573
				Root MSE	=	6.493

consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
income	.5090909	.037428	14.24	0.000	.4266678 .591514
_cons	24.45455	6.413817	3.81	0.005	9.664256 39.24483

command

Variables

Name	Label
consumption	Consumption
income	

Properties

Variables

Name	Label	Type	Format	Var label	Notes
income		int	%10.0g		

Data

Filename	Label	Notes	Variables	Observations	File
default			2	10	en

Notepad - Windows Journal

File Edit View Insert Actions Tools Help

Page Work

Note Title

04-11-2020

Hypothesis testing

$$t = \frac{\hat{\beta}}{s.e(\hat{\beta})}$$

$$= \frac{0.5090}{0.0357}$$

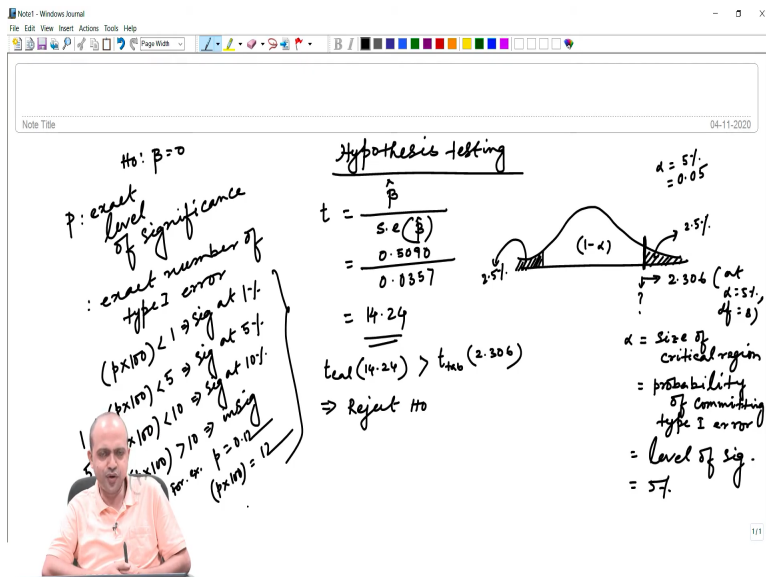
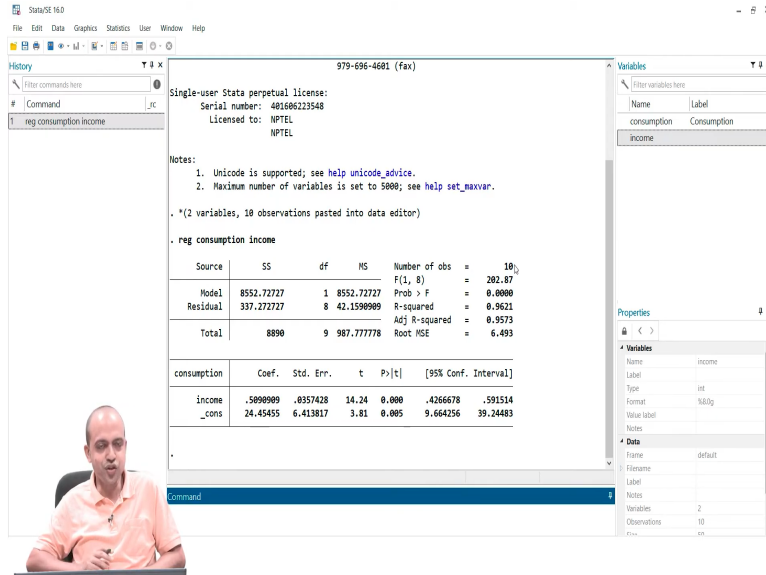
$$= \underline{\underline{14.24}}$$

$\alpha = 5\% = 0.05$

2.5%

$?$

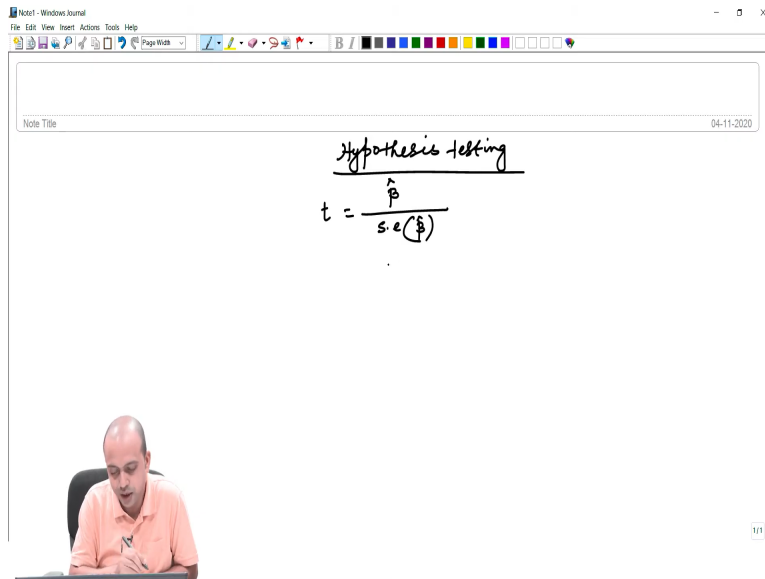
$\alpha =$ size of critical region
 $=$ probability of committing type I error
 $=$ level of sig.
 $= 5\%$



Now, if you, if we go back to our data set and do the estimation, this is our result. Look at the result. So, we have estimated a consumption function and we have only one explanatory variable income and the coefficient of income is 0.5090, that means for a unit change in income on an average, consumption changes by 0.5090 unit.

This is the interpretation that we have learnt but now what we will do? We will see whether this particular income coefficient, it is statistically significant or not. And by checking the statistical significance, basically we are ensuring that the value that we got here is not by a chance factor, rather it is due to a statistical significance.

(Refer Slide Time: 02:14)



Now, how will you do that? If you go back to our discussion yesterday, then we said that what is the decision rule of hypothesis testing? The decision rule says that we have to first construct the t

statistic and that is $\frac{\hat{\beta}}{s.e\hat{\beta}}$. And in this particular case, if you look at, what is our $\hat{\beta}$? $\hat{\beta}$ is 0.5090 and what is the standard error? Standard error is 0.035. 0.5090 divided by 0.0357 - if you do that, then what you will get?

You will get the calculated value of the t which is supplied by Stata - 14.24. So, this is the calculated value and this calculated value we have to compare with the tabulated value, that is what we learnt. That means basically, in terms of the diagram, what is our calculated value? Calculated value is 14.24, now we want to know what is the tabulated value.

And if you look at your t table, and specify your α , that means this area is what we said, that 2.5 percent, this is also 2.5 percent, that means we said that α equals to 5 percent or it is equal to 0.05. Let us specify the area, the size of the critical region is 5 percent, so that means I am saying that α is equals to size of the critical region or we can also say that this is probability of committing type one error or level of significance.

These are the three interpretations of α , and that is equal to 5 percent. So, that means we are saying that we will commit 5 type one errors in 100 cases. That is the pre-specified level of alpha

we have kept. Now, if you look at your table, now here you see that total number of observations is 10 and as I said yesterday, that t will follow $n-2$ degrees of freedom, that means 8 degrees of freedom.

So, what you have to see? At 8 degrees of freedom and 5 percent level of significance, what is the value of t ? That means this is equal to, I will say, this is equals to 2.306 at α equals to 5 percent and degrees of freedom equals to 8. At 9 degrees of freedom, this is 2.306. Now, when your calculated value is 14.24 and the tabulated value is only 2.306, you can understand that means you are actually easily entering into the critical region.

So, that means what I said, that calculated t , so t calculated which is 14.24 is actually greater than t tabulated which is 2.306. And even at the 1 percent level also, you will see that your calculated value is greater than the tabulated one, that you can verify from the t table. So, that means this implies that reject your null hypothesis, reject H_0 .

And what is our H_0 , we mentioned? H_0 was β equals to 0. That means income does not have any impact on consumption. Now, from this diagram, if you reject your H_0 , so that means we can establish our claim that yes income has significant impact on consumption. From the diagram, one thing is very clear that if a particular variable is significant at 5 percent, will it be significant at 10 percent? The answer is yes.

Because if you increase the size of this critical region, then obviously, it would be easy, it would be easier for you to enter into the critical region. The more you increase, more you have about the size of the critical region, it would become easier for the calculated value to enter into the critical region. If I reverse the question. If the variable is significant at 5 percent, will it be significant at 1 percent also?

That, we may not say anything about. That we have to check. Because if you reduce the size of the critical region, then calculated value may or may not enter into the critical region. It is very simple right? You increase the size, it become easier for calculated value to enter, if you reduce the size, then it would be difficult for the calculated value to enter into the critical region.

So, once the variable is significant at a specified level, that means it would be automatically significant at higher level of significant also. A variable which is significant at 5 percent, it is

definitely significant at 10 percent. But if the variable is significant at 5 percent, it may or may not be significant at 1 percent. That we have to again check.

So, this is the area of non-rejection and this is the area of rejection which is called the critical region. Critical region, level of significance and probability of committing type one error are all related concepts. When I am specifying alpha equals to 5, that means probability of committing type one error is actually 0.05. That means out of 100 cases, you are actually committing your type one error 5 times.

Now, this is what we can learn, so that means to, if you want to take the decision based on the t statistic, what you have to do? You have to take the calculated value, you have to compare with the tabulated value and based on whether your calculated value is greater than the tabulated one or not, you have to decide whether to reject your null or not.

But this Stata has made our life simple. By supplying additional information. What is that additional information? That additional information is here. Look at, they have also given something called p. What is this p? Now, you have to keep in mind that this p, what Stata has given as p, is actually called exact level of significance. Or you can say that this is the exact number of type one errors.

Now, why this p is required? As I said, this t is telling that this is significant at 5 percent but I do not know whether it is significant at 1 percent level also. I need to again check the value of t at 1 percent level and then decide. But if you take p, since this is called exact level of significance, that means I exactly know the size of the critical region.

And if you go by the p value, then you do not have to compare that tabulated with the calculated one, rather what you know, if the value of p, let us say, in this case, what is the value of p? 0.000. That means you are actually not committing type one error at all. But then as I said earlier, when your decision making involves some kind of probability distribution, because it is t divided by standard error of β , beta hat divided by standard error, that determine your t, so that means probability distribution is involved there. How is it possible that you are not committing any type one error?

Actually, the meaning of this is, the p value is very small, 0.000000000 some value will come. It is such a small value that Stata is reporting it as 0. So, that means in 100 cases you may not commit any mistake, maybe 1000 cases, something like that. So, probability of committing type one error is very small when your p value is small. So, based on the p value, statistician or econometrician, they have made a simple rule. What is the rule?

You have to multiply the p value with 100. And if after multiplication, if this becomes less than 1, then we will say that the variable is significant at 1 percent. Then, you multiply this by 100. And if this becomes greater than 1 but less than 5, then we will say that the variable is significant at 5 percent. And if the variable p value after multiplication, if it is greater than 5, but less than 10, then we will say that this is, the variable is significant at 10 percent.

And if after multiplication, if it is greater than 10, then we will say that the variable is actually insignificant. For example, let us say the p value is 0.12. So, if you multiply that by 100, that means p multiplied by 100 would be equal to 12. So, that means out of 100 cases, you are committing 12 type one errors and as I said earlier, we cannot afford of making 12 errors, more than 10 errors, we cannot afford to make. Then we will say that the variable is actually insignificant.

So, this simple rule if you remember, just take the p value, lower the p value higher would be the level of significance. Just multiply that by 100 and follow this simple rule. Because we do not report the result, while interpreting we do not say the variable is significant at 2.3 percent or 4.5 percent or 1.2 percent, nothing like that.

We will only say whether it is significant at 1 percent, 5 percent or 10 percent. That means all this exact level of significance will be converted into 1 percent, 5 percent and 10 percent after multiplying this by 100. If it is less than 1, then it is significant at 1 percent, if it is less than 5 but greater than 1, then significant at 5 percent, greater than 5 but less than 10, it is significant at 10 percent only.

And after multiplication, if it is greater than 10, then we will say that this is insignificant actually. This is the simple rule that we apply to see whether the variable is significant, following the, comparing t with the tabulated one and applying this p. So, this is called level of significance

approach. But there is an alternative way of checking the significance of the variable. That is called the interval approach, confidence interval approach.

(Refer Slide Time: 17:15)

Confidence interval approach of checking the significance:

$$y_i = \alpha + \beta x_i + u_i$$

$$H_0: \beta = 0$$

(1- α)

whether this interval captured the hypothesised value of the true population parameter or not

$$\hat{\beta} \pm t_{\alpha/2} \cdot s.e.(\hat{\beta})$$

979-696-4681 (fax)

Single-user Stata perpetual license:
Serial number: 48166623548
Licensed to: NPTEL
NPTEL

Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5808; see help set_maxvar.

*(2 variables, 10 observations pasted into data editor)

```
. reg consumption income
```

Source	SS	df	MS	Number of obs	F(1, 8)	Prob > F
Model	8552.72727	1	8552.72727	10	282.87	0.0000
Residual	337.272727	8	42.1590909		R-squared	0.9621
Total	8890	9	987.777778		Adj R-squared	0.9573
					Root MSE	6.493

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
consumption					
income	-.5098909	.8357428	14.24	0.000	-4266678 .591514
._cons	24.45455	6.413817	3.81	0.005	9.664256 39.24483

Confidence interval approach of checking the significance:

$y_i = \alpha + \beta x_i + u_i$
 $H_0: \beta = 0$

region of non-rejection

Decision rule:
 (1) If the estimated confidence interval captures the hypothesized value of the true population parameter, do not reject H_0 .
 (2) If it is not captured by the interval, reject H_0 .

whether this interval captures the hypothesized value of the true population parameter or not-

$\hat{\beta} \pm t_{\alpha/2} S.E.(\hat{\beta})$

LL = $0.5090 - 2.306 * 0.0357$
 = 0.4266
 UL = $0.5090 + 2.306 * 0.0357$
 = 0.5915

So, confidence interval approach of checking the significance. Now, what I said? When you are estimating this type of model $Y = \alpha + \beta x_i + u_i$ and your null hypothesis is β equals to 0, that means hypothesized value of the true population parameter is 0, then in interval approach, what you have to do?

You have to just check, let us say in this interval approach this 1 minus alpha, this region, whether this region called region of confidence or region of non-rejection, whether that hypothesized value of the true population parameter is captured by this interval or not.

So, what you have to check? Whether this interval captures H_0 , captures the hypothesized value of the true population parameter or not. And how this interval is determined? The interval is determined as $\hat{\beta} \pm t_{\alpha/2} S.E.(\hat{\beta})$. So, that means from this particular example, what is $\hat{\beta}$? $\hat{\beta}$ is 0.5090 and standard error is 0.0357 and the calculated value of t is 2.36.

If you put all these values, that means the upper limit, this is called lower limit of the interval and this is called upper limit. So, that means lower limit of the interval equals to 0.5090 plus, what is t alpha by 2? 2.306. And what is standard error of beta hat? That is basically 0.0357. If you do that, then you will get the lower limit of the interval and that is actually given by Stata.

What is the value? Look at this. Stata has given you 95 percent confidence interval. That means Stata have specified alpha equals to 5. What is the value? Value is 0.4266. This would become 0.4266. That is the lower limit of the interval. What would be the upper limit of the interval?

And upper limit would be beta hat plus this, 0.5090 plus 2.306 multiplied by 0.0357. And that would become your upper limit which is 0.5915. This is how the confidence interval is constructed using this particular formula. One thing you have to keep in mind. Stata will always give you the confidence interval specifying alpha equals to 5 percent, that means this is 95 percent confidence interval.

If you would like to get confidence interval specifying alpha equals to 1 percent or alpha equals to 10 percent, that means 90 percent and 99 percent confidence interval, you can always calculate manually. Stata will not give you. Stata will give you only one single interval that is called 95 percent confidence interval.

And calculating 99 percent or 90 percent confidence interval is very easy. Instead of 2.306, you have to specify t value at 1 percent and 10 percent, that is all. So, this is how the interval is constructed and as I said earlier, the standard error determines the size of the region. That means the width, range of this interval is determined by the beta because beta hat is constant, fixed, t alpha by t is also fixed for a given number of observation.

Size or region of non-rejection or this is called region of non-rejection is actually determined by your standard error of beta hat. Higher the standard error, higher would be the size and if this size is higher, so obviously it would be highly probable that each and every hypothesized value of the true population would be captured by this interval.

If your size of the interval is much wider, a much wider interval will automatically capture the true hypothesized value, that is not a great deal. So, that means your decision making would be precise if your standard error is precise. That is why while estimating the model, one of our desirable property we said that that should be the minimum variance property should be there. Because if the variance is minimum, standard error is minimum, then your interval would also become very precise.

Higher the size of the interval, larger the size of the interval, less would be the preciseness in your decision making. That we have to keep in mind. This is how you construct your interval. Then what is the decision rule? If the estimated confidence interval captures the hypothesized value of the true population parameter, then the decision-making rule says you do not reject your null.

What I am saying, you do not reject your null. So, if your hypothesized value of the true population parameter is captured by this interval, then the decision-making rule says you should not reject your null. And if you do not reject your null, then what would be your decision? That means what would be the inference?

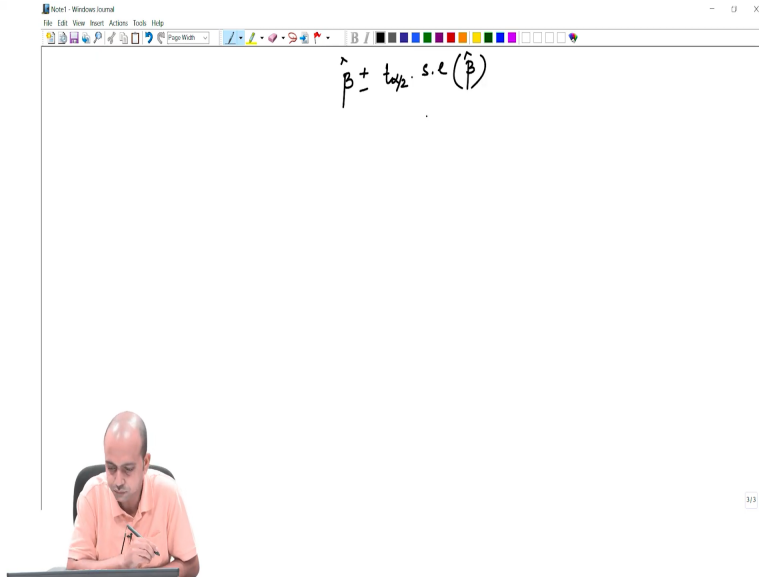
That means variable is not significant. That means a variable is not significant. This is the decision-making rule. Now, in this particular case, look at what is happening. What is the interval? 0.4266 and 0.5090. Now, in this interval, do you think the true hypothesized value captured? Is 0 captured by this interval? Since it is in the positive axis, you can understand that 0 is not captured.

And if 0 is not captured by your estimated interval, then you have to reject your null. And if you reject your null, that means variable is significant. So, that means we can say that whether you follow level of significance approach or interval estimation approach, both these approaches are leading to same result.

It should not so happen that a variable is significant when you follow the confidence interval approach, but it is otherwise, when you follow the level of significance approach, that means you are comparing calculated t with the tabulated t, it should not so happen. So, what we are saying then? So, that means population parameter, if the estimated confidence captures the hypothesized value of the true population parameter, then the decision is - do not reject your null, H_0 .

And if it is not captured by the interval, then reject H_0 . And that is what is happening in this particular case. So, that means 0 does not lie between 0.42 and 0.59. So, that means here it is 0.42 and here it is 0.59, so 0 is not there in this interval. 0 lies outside the interval. So, we are rejecting the null and claiming that income does have significant impact on consumption.

(Refer Slide Time: 30:45)



Stata/SE 16.0

File Edit Data Graphics Statistics User Window Help

History

```

1 . reg consumption income

```

979-696-4681 (fax)

Single-user Stata perpetual license:
 Serial number: 48168622548
 Licensed to: NPTEL

Notes:

- Unicode is supported; see help unicode.advice.
- Maximum number of variables is set to 5000; see help set_maxvar.

*(2 variables, 10 observations pasted into data editor)

```

. reg consumption income

```

Source	SS	df	MS	Number of obs	=	10
Model	852.72727	1	852.72727	F(1, 8)	=	282.87
Residual	337.272727	8	42.1590909	Prob > F	=	0.0000
				R-squared	=	0.9621
				Adj R-squared	=	0.9573
Total	8890	9	987.777778	Root MSE	=	6.493

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
income	.59090909	.0357428	14.24	0.000	-.4266678 .591514
_cons	24.45455	6.413817	3.81	0.005	9.664256 39.24483

Command

Variables

Name	Label
consumption	Consumption
income	

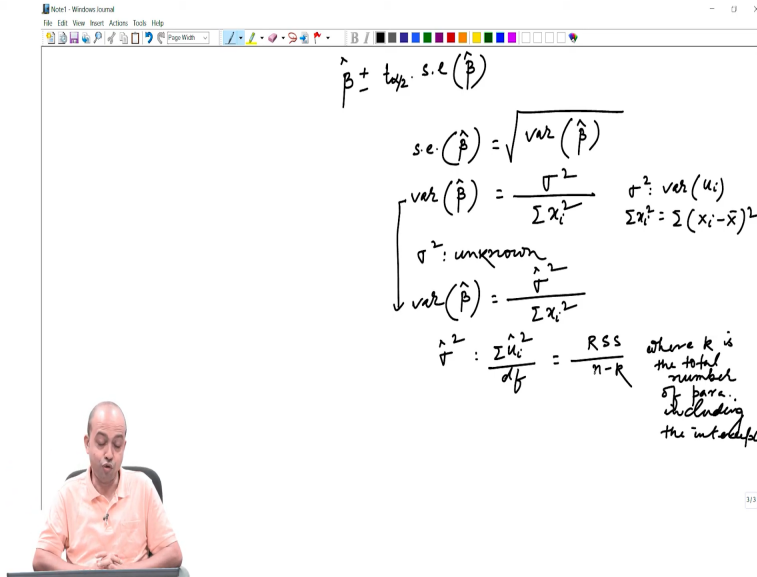
Properties

Variables

Name	Label	Type	Format	Value label	Notes
income		int	%d0g		

Data

Filename	Label	Notes	Variables	Observations	View
default			2	10	ok



Now, one more thing what we have to remember, we said that the interval is estimated using this formula $\hat{\beta} \pm t_{\alpha/2} S.E.(\hat{\beta})$. Now, sometimes, by mistake, what we do? We estimate the interval and we see this $\hat{\beta}$ is actually lying in this interval or not.

For example, here after estimating the interval you will see that your beta hat 0.50 is lying in the interval and as I said, decision making rule says when the interval captures, the value then you do not reject your null. And if you do not reject your null, then your variable would be insignificant. But that is actually not the case by level of significance approach, your t says variable is significant. That is a mistake.

We should never examine whether this beta hat is actually lying in the interval or not. Rather what I said, $\hat{\beta}$ will only play a role in the construction of the interval. Once the interval is constructed, we will only see whether the hypothesized value of the population parameter is lying here or not. And what is the hypothesized value? β equals to 0. So, we will always see whether 0 is lying here or not, not this beta hat. That is one thing.

And secondly, since standard error of $\hat{\beta}$ is playing a role there, we must also know how is the standard error calculated. See, standard error of beta hat is basically root of variance of beta hat,

and how is this variance of beta hat calculated? Variance of beta hat is calculated using this formula. Sigma square divided by standard xi square.

And what is sigma square? Sigma square is basically, variance of u_i . And summation xi square is basically, sum of x_i minus \bar{x} whole square. But the sigma square is actually unknown to you, sigma square is an unknown population parameter. So, if sigma square is unknown, then how do you calculate variance of beta hat and standard error of beta hat?

In that context, we have to keep in mind that when something is unknown in statistics or econometrics, statistician say that you replace the unknown population parameter with its sample counterpart. Since sigma square is unknown, then what do you do? You calculate your variance of beta hat using its sample counterpart which is basically sigma hat square divided by summation xi square. And what is sigma hat square?

Sigma hat square is basically summation u_i hat square divided by its degrees of freedom. That means sigma hat square is nothing but your residual sum of square. And what is the degrees of freedom? It is $n - k$ where k is the total number of parameters to be estimated including the intercept, where k is the total number of parameter including the intercept. That is how you calculate your sigma hat square and that is how you calculate your variance of beta hat.

And why the degrees of freedom for RSS is $n - k$? That we have discussed in our previous class when we are talking about the ANOVA table. We discussed if you go back to our previous lecture, to that particular section we have clearly explained what is the degrees of freedom for TSS, what is the degrees of freedom for RSS and what is the degrees of freedom for ESS.

So, that means one thing we have learnt here, when we have a problem with unknown population parameter while computing some statistical measure, our rule says we have to replace that with the true, with the sample counterpart. Sigma square is unknown, so we are replacing by sigma hat square which is nothing but summation u_i hat square or RSS by its degrees of freedom. So, this is how you can actually work with your hypothesis testing.

So, so far we discussed the estimation and hypothesis testing in the context of a two variable linear regression model, that means we have only one explanatory variable in the right hand side. Now, what will happen when there are multiple explanatory variables in the right-hand side?

That we will discuss in our next class. So, in our next class we will talk about multiple linear regression model. Thank you.