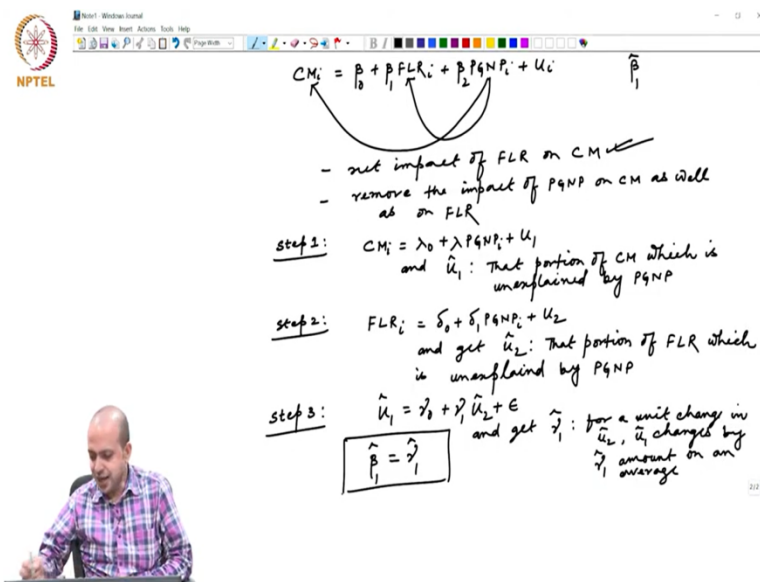


Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras

Application of STATA for hypothesis testing and introduction to multiple linear regression model Part - 4

(Refer Slide Time: 00:14)



NPTEL

$$CM_i = \beta_0 + \beta_1 FLR_i + \beta_2 PGNP_i + u_i \quad \hat{\beta}_1$$

- net impact of FLR on CM
- remove the impact of PGNP on CM as well as on FLR

step 1: $CM_i = \lambda_0 + \lambda_1 PGNP_i + u_1$
 and \hat{u}_1 : That portion of CM which is unexplained by PGNP

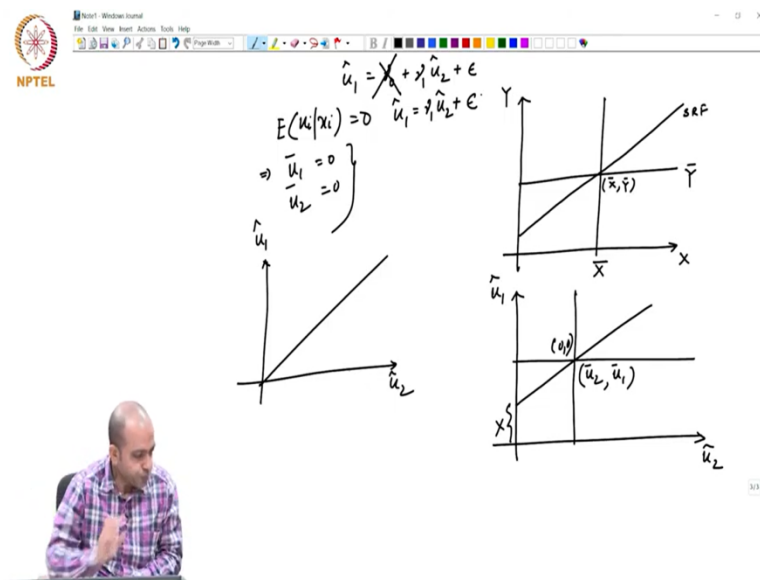
step 2: $FLR_i = \delta_0 + \delta_1 PGNP_i + u_2$
 and get \hat{u}_2 : That portion of FLR which is unexplained by PGNP

step 3: $\hat{u}_1 = \gamma_0 + \gamma_1 \hat{u}_2 + E$
 and get $\hat{\gamma}_1$: for a unit change in \hat{u}_2 , \hat{u}_1 changes by $\hat{\gamma}_1$ amount on an average

$\hat{\beta}_1 = \hat{\gamma}_1$

So, I will make you understand what is the mistake I have committed in step 3.

(Refer Slide Time: 00:18)



NPTEL

$$\hat{u}_1 = \lambda_0 + \lambda_1 \hat{u}_2 + E$$

$$E(u_i | x_i) = 0 \Rightarrow \begin{cases} \bar{u}_1 = 0 \\ \bar{u}_2 = 0 \end{cases}$$

$$\hat{u}_1 = \gamma_0 + \gamma_1 \hat{u}_2 + E$$

The slide contains three graphs:

- Top Graph:** A scatter plot of Y vs X showing a regression line (SRF) and a horizontal line at \bar{Y} . The intersection is at (\bar{X}, \bar{Y}) .
- Bottom-Left Graph:** A scatter plot of \hat{u}_1 vs \hat{u}_2 showing a regression line passing through the origin (0,0).
- Bottom-Right Graph:** A scatter plot of \hat{u}_1 vs \hat{u}_2 showing a regression line passing through the origin (0,0) and a point (\hat{u}_2, \hat{u}_1) .

In step 3, basically what we did, \hat{u}_1 equals to γ_0 plus $\gamma_1 \hat{u}_2$ plus ϵ . This is the equation we have specified. Now, if you recall, one of the properties of our sample regression function is that it must pass through the origin.

That means, this is basically \bar{x} and \bar{y} . This is the property that sample regression function must pass through the sample mean or average. That means here when I am running a regression of y on x , it should pass from \bar{x} and \bar{y} . That is one of the important property that we have discussed. This is the SRF. Now, in this regression, what we are doing? If you replace x by \hat{u}_2 and your dependent variable here is \hat{u}_1 , then that should also pass through the origin.

So, that means this should be \bar{u}_2 and \bar{u}_1 . So, that should also pass through the sample average. Now, my question is what is the value of \bar{u}_2 and \bar{u}_1 ? If you recall, one of our assumptions what we mentioned is that expectation of u_i given x_i is equal to 0. If that is the case, following this assumption we can say that \bar{u}_1 equals to 0 and \bar{u}_2 is also equal to 0. And that means, this is the co-ordinate called 0, 0. So, that means this intercept should not be there in this equation.

So, that means when you are running an equation of \hat{u}_2 on \hat{u}_1 , that regression must pass through the origin. This point is 0, 0. So, that means this intercept should not be there. If the intercept is there, that means that cannot be the co-ordinate of 0,0. So, that means in this equation, when I am specifying this term should not be 0, should not be there. So, our equation would be \hat{u}_1 equals to $\gamma_1 \hat{u}_2$ plus ϵ . No intercept should be there and that is coming from the property- one of the important property of the sample regression function- that sample regression function must pass through the origin.

And since then u_1 and u_2 , they both have 0 mean, the sample average of u_1 and u_2 is 0, so the line must pass through the origin, there should not be any intercept. So, this is basically the theoretical portion. That means this is the theory of how to keep the impact of other factor constant. Now, these steps, what we have learnt now, we will try to perform using the data set.

(Refer Slide Time: 04:39)

The screenshot shows the Stata Data Editor interface. The main window displays a dataset with 25 rows and 4 columns: cm, flr, pgnp, and tfr. The data values are as follows:

cm	flr	pgnp	tfr
123	37	1870	6.66
204	22	130	6.15
202	36	310	7
197	45	570	6.25
96	76	2050	3.81
209	26	200	6.44
170	45	670	6.19
240	29	300	5.89
241	11	120	5.89
55	55	270	2.16
75	87	1100	3.93
129	55	900	5.89
14	93	1710	3.5
105	31	1150	7.41
94	77	1160	4.21
96	80	1270	5
148	30	500	5.27
98	69	660	5.27
111	43	420	6.5
118	47	1000	6.12
209	17	200	6.19
189	270	5.85	
126	560	6.16	
17	4260	1.8	
112	240	4.75	
		4.1	

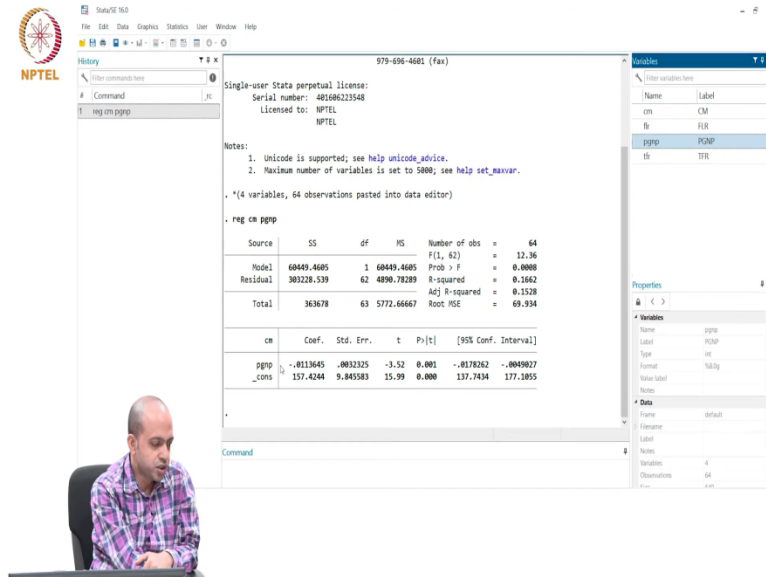
The right-hand side of the interface shows the 'Variables' list with the following details:

Name	Label	Type	Format	Value label
cm	CM	int	%d0g	
flr	FLR	byte	%d0g	
pgnp	PGNP	int	%d0g	
tfr	TFR	float	%d0g	

So, this is the data what we are getting. Look at this. So, first I will show you the data. This is the data, this is child mortality rate, this is female literacy rate, this is per capita GNP and this is total fertility rate, some other variable and we have state level data. So, what we will do now? We will try to follow that step.

(Refer Slide Time: 05:10)

The screenshot shows the Stata command window. The command entered is `reg cm pgnp`. The output shows the Stata logo and version information (16.0), copyright (1985-2019 StataCorp LLC), and contact information for StataCorp. It also displays the user's perpetual license details (Serial number: 401606223548, Licensed to: NPTEL) and notes about Unicode support and the maximum number of variables (5000). The command window shows the command `reg cm pgnp` and the output `There are no items to show.`



So, we have child mortality rate, FLR and PGNP and we want to get the net impact of FLR on child mortality rate. So, step 1, what we need to do? We need to run a regression where the dependent variable is CM and independent variable would be PGNP because we need to remove the impact of PGNP from both CM and FLR. Since we are interested to get net impact of FLR, we are eliminating the impact of PGNP. So, first step, we are removing the impact of PGNP of CM. How to do that?

We will run a regression, **reg CM PGNP**. This is how we will run the regression. You can easily interpret this coefficient also. See there is a negative relationship between per capita GNP and CM. As income of a particular state increases, child mortality decreases by 0.01 unit on an average, which is expected also. As income increases, people become more aware, people take food, good rest and they give proper care in terms of medicine, so on and so forth. And as a result of which, child mortality rate goes down.

So, once we estimate this model, then what we need to do? From this regression, we need to collect the predicted value of the error term. Now, to collect the predicted value of the error term u_1 hat from this regression, what we need to do? We need to put a specific command. We need to give a specific command and the command is very simple. I am writing the command also.

(Refer Slide Time: 07:13)

The image shows a video lecture with a man at a computer. A whiteboard in the background contains handwritten notes:

- predict u_1 , residual — step 1
- predict u_2 , residual — step 2
- reg u_1 , u_2 , u_3 — step 3

The Stata software window displays the following information:

Single-user Stata perpetual license:
Serial number: 481606213548
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.

*(4 variables, 64 observations pasted into data editor)

```
. reg cm pgrp
```

Source	SS	df	MS	Number of obs =	F(1, 62)	Prob > F
Model	60449.4685	1	60449.4685			0.0000
Residual	303228.539	62	4890.78939			0.1662
Total	363678	63	5772.66667			

Adj R-squared = 0.1528
Root MSE = 69.934

	cm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pgrp		-.0113645	.0032325	-3.52	0.001	-.0178262	-.0049027
_cons		157.4244	9.84583	15.99	0.000	137.7434	177.1055

```
. predict u1,residual
```

Command

Variables:

Name	Label
cm	CM
flr	FLR
pgrp	PGRP
trr	TRR

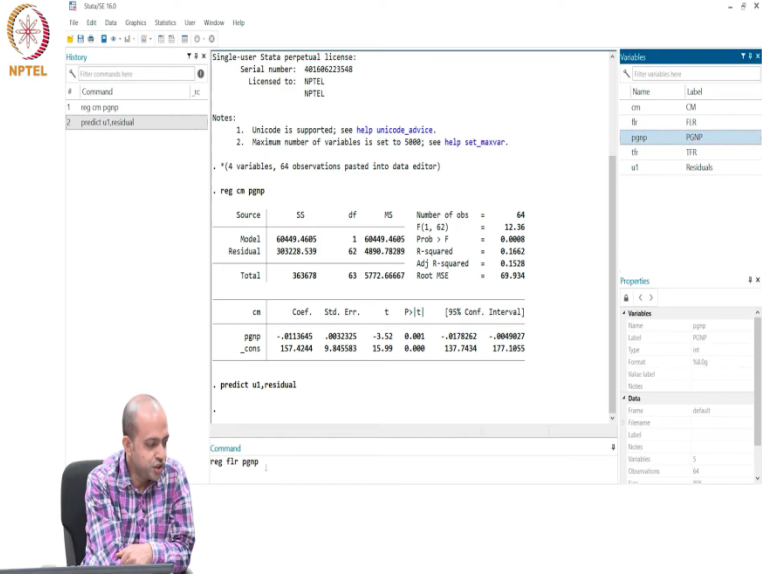
Properties:

Variables:

Name	Label	Type	Format	Value Label	Missing
pgrp	PGRP	int	%d		

Data:

Frame	Display	Variables	Observations
default		4	64



Stata/SE 16.0
File Edit Data Graphics Statistics User Window Help

History
Filter commands here
Command |_jt
1 reg cm pppp
2 predict u1,residual

Single-user Stata perpetual license:
Serial number: 401606223548
Licensed to: NPTEL
NPTEL

Notes:
1. Unicode is supported; see help unicode_notice.
2. Maximum number of variables is set to 5000; see help set_maxvar.

*(4 variables, 64 observations pasted into data editor)

```
. reg cm pppp
```

Source	SS	df	MS	Number of obs	=	64
Model	60649.4605	1	60649.4605	F(1, 62)	=	12.36
Residual	303228.539	62	4890.78989	Prob > F	=	0.0008
Total	363878	63	5772.66667	R-squared	=	0.1662
				Adj R-squared	=	0.1528
				Root MSE	=	69.934

cm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pppp	-.0113645	.0032325	-3.52	0.001	-.0178262	-.0049027
_cons	157.4244	9.045583	15.99	0.000	137.7434	177.1055

```
. predict u1,residual
```

```
reg flr pppp
```

Variables

Name	Label
cm	CM
flr	FLR
pppp	PCNP
u1	TRR
u1	Residuals

Properties

Variables

Name	Label	Type	Format	Value-label	Notes
pppp		int	%10g		
cm		int	%10g		
flr		int	%10g		
u1		float	%10g		

Data

Name	Label	Format	Value-label	Notes
pppp		%10g		
cm		%10g		
flr		%10g		
u1		%10g		

Variables

Name	Label	Type	Format	Value-label	Notes
pppp		int	%10g		
cm		int	%10g		
flr		int	%10g		
u1		float	%10g		

Data

Name	Label	Format	Value-label	Notes
pppp		%10g		
cm		%10g		
flr		%10g		
u1		%10g		

Variables

Name	Label	Type	Format	Value-label	Notes
pppp		int	%10g		
cm		int	%10g		
flr		int	%10g		
u1		float	%10g		

Data

Name	Label	Format	Value-label	Notes
pppp		%10g		
cm		%10g		
flr		%10g		
u1		%10g		

Variables

Name	Label	Type	Format	Value-label	Notes
pppp		int	%10g		
cm		int	%10g		
flr		int	%10g		
u1		float	%10g		

Data

Name	Label	Format	Value-label	Notes
pppp		%10g		
cm		%10g		
flr		%10g		
u1		%10g		

Stata/SE 16.0

History

```

1. reg cm pgrp
2. predict u1, residual
3. reg flr pgrp

```

Residual	303228.539	62	4890.78289	R-squared = 0.1642
Total	363670	63	5772.66667	Adj R-squared = 0.1528
				Root MSE = 69.934

cm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pgrp	-.0113645	.0032325	-3.52	0.001	-.0178262 -.0049027
_cons	157.4244	9.845583	15.99	0.000	137.7434 177.1055

```

. predict u1, residual
. reg flr pgrp

```

Source	SS	df	MS	Number of obs = 64
Model	3072.80504	1	3072.80504	F(1, 62) = 4.82
Residual	39540.945	62	637.757277	Prob > F = 0.0329
Total	42613.75	63	676.40873	R-squared = 0.0721
				Adj R-squared = 0.0571
				Root MSE = 25.254

flr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pgrp	-.0025622	.0011673	2.20	0.032	-.0042289 -.0008956
_cons	47.59716	3.55533	13.39	0.000	40.49016 54.70416

Command

```

.

```

Variables

Name	Label
cm	CM
flr	FLR
pgrp	PGRP
u1	TTR
u1	Residuals

Properties

Variables

Name	Label	Type	Format	Value label	Notes
pgrp		int			
flr		float			
u1		float			
u2		float			

Data

Name	Filename	Label	Notes	Variables	Observations	From
				5	64	me

Stata/SE 16.0

History

```

1. reg cm pgrp
2. predict u1, residual
3. reg flr pgrp
4. predict u2, residual

```

Command

```

.

```

Variables

Name	Label
cm	CM
flr	FLR
pgrp	PGRP
u1	TTR
u1	Residuals
u2	Residuals

Properties

Variables

Name	Label	Type	Format	Value label	Notes
pgrp		int			
flr		float			
u1		float			
u2		float			

Data

Name	Filename	Label	Notes	Variables	Observations	From
				5	64	me

The command is **predict**, and then you have to give some name. Predict command is always used for predicting something. That means when I am writing u_1 hat, u_2 hat, y hat, the hat means the predicted value, that is why predict command. What you are predicting? Some name you are giving. Predict u_1 . What is u_1 basically? Residual. This is the command. Similarly, for u_2 , **predict u_2 , residual**. This is the command. So, now we will put this command here. We will say, this is predict u_1 , residual.

Now, the moment I put the command, Stata will immediately predict the u_1 hat and look at here, u_1 is already included as a variable now. Whether it is predicted or not, how will you check? You

just look at in the variable list, if the variable is appearing, that means Stata has already predicted the variable and stored it also.

Now, what is the next step? In the next step, we need to remove the impact of PGNP on FLR. So, again regress FLR on PGNP, this is how. And the impact is, look at here, positive that means as per capita income GNP increases, FLR also increases. And the variable is significant at what percentage level can you think of from the p value? P value is 0.032.

So, if you multiply this value with 100, what will come? 3.2 which is greater than 1 but less than 5. That means this PGNP is significant at 5 percent level. This is how you have to interpret this coefficient. So, again from this, we need to predict and same command, predict u2 and then you put residual. So, once you put this command, then Stata will immediately predict the residual also. And how will you check whether it is predicted or not? Look at here, u2 is also included.

Now, what we have to do? We have to run a regression that means in step 3, u1 hat is regressed on u2 hat with no constant term. And what would be the command for that in step 3? Reg u1 on u2 and you need to specify Stata that I do not want any constant term. And what is the command for that? We need to specify '**no cons**', this is the command in step 3. This is step 2, this is step 3, step 1. In step 1, we will regress and then we will collect this.

(Refer Slide Time: 11:47)

The screenshot shows the Stata software interface with the following components:

- Command History:**

```

1 reg cm pgnp
2 predict u1,residual
3 reg flr pgnp
4 predict u2,residual

```
- Regression Results:**

	cm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pgnp		-.8113645	.0802325	-3.52	0.001	-.9178262 -.6949027
_cons		157.4244	9.845583	15.99	0.000	137.7434 177.1055

	flr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pgnp		.0025622	.0011673	2.20	0.032	.0002289 .0048956
_cons		47.59716	3.55533	13.39	0.000	40.49016 54.70416
- ANOVA Table:**

Source	SS	df	MS	Number of obs = 64
Model	3072.80504	1	3072.80504	F(1, 62) = 4.82
Residual	39540.945	62	637.75717	Prob > F = 0.0319
Total	42623.75	63	676.40873	R-squared = 0.0721
				Adj R-squared = 0.0571
				Root MSE = 25.254
- Command:**

```
reg u1 u2,nocons
```
- Variables List:**

Name	Label
cm	CM
flr	FLR
pgnp	PGNP
u1	Residual
u2	Residual

The screenshot shows the Stata/SE 16.0 interface. The main window displays the results of a regression model. The command history shows the following steps:

```

1. reg cm pgnp
2. predict u2,residual
3. reg flr pgnp
4. predict u2,residual
5. reg u1 u2,nocons

```

The regression results for the final model (reg u1 u2,nocons) are as follows:

Source	SS	df	MS	Number of obs	F(1, 63)	Prob > F
Model	196912.912	1	196912.912	64	116.49	0.0000
Residual	106315.622	63	1687.54955			0.6494
Total	303228.534	64	4737.94504			41.08

u1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
u2	-2.231586	.2065878	-10.80	0.000	-2.644419 -1.818753

The inset video shows a man in a plaid shirt speaking, likely providing a verbal explanation of the statistical results.

In step 3, what we need to do, we will regress u1 on u2 and then we will put no cons. So, what is the coefficient here? Coefficient of u2 is minus 2.231. So, that means for a unit change in u2, u1 changes by minus 2.231 unit on an average. But then, what is its relation with the original equation? Since \hat{u}_2 is basically the purified value of FLR and u1 is the purified value of child mortality rate, that means, the interpretation of this would be when FLR increases by 1 unit, then child mortality rate decreases by 2.231 unit.

So, u2 this 2.231 is basically the net impact of FLR on child mortality rate. Since this is negative, we can say that female literacy rate and child mortality rate, they are negatively related and which is very true also. As we said in the beginning that as female literacy rate increases, then mothers become more aware of what type of proper care they need to take during pregnancy and after delivery, child mortality rate will decrease. That is why the negative sign.

So, this is the step by step procedure what we should follow to keep the impact of other factor constant. But that does not mean that whenever we want to get the net impact of a particular variable in a multiple regression model, we will follow all these steps. Rather, at a single step itself, we can get the net impact. That is the beauty of this statistical software. How to do that? What was our original model? Our original model was $CM = \beta_0 + \beta_1 FLR + \beta_2 PGNP$.

(Refer Slide Time: 14:36)

Stata/SE 16.0

History

1. reg cm pgnp
2. predict u2,residual
3. reg flr pgnp
4. predict u2,residual
5. reg u1 u2,noncons

```

. reg cm flr pgnp
      Source      SS       df       MS       Number of obs   =      64
      F(2, 63)    = 116.60
      Prob > F    = 0.0000
      R-squared   = 0.6494
      Adj R-squared = 0.6438
      Root MSE   = 41.408

      Total      301328.534      64      4737.94584

      u1      Coef.   Std. Err.   t    P>|t|   [95% Conf. Interval]
      -----+-----
      flr     -2.231586   .2065878   -10.80  0.000   -2.644419   -1.818753
      pgnp     .8055466   .0920833    8.74  0.000   .6214201   .9896731
      _cons   283.4416   11.59318   24.46  0.000   259.7480   307.1352

      . predict u2,residual
      . reg u1 u2,noncons
      Source      SS       df       MS       Number of obs   =      64
      F(1, 63)    = 116.60
      Prob > F    = 0.0000
      R-squared   = 0.6494
      Adj R-squared = 0.6438
      Root MSE   = 41.408

      Total      301328.534      64      4737.94584

      u1      Coef.   Std. Err.   t    P>|t|   [95% Conf. Interval]
      -----+-----
      u1      1.0000000   .0000000    1.00  0.000   .9999999   1.0000001
      u2     -2.231586   .2065878   -10.80  0.000   -2.644419   -1.818753
  
```

Command

```
reg cm flr pgnp
```

Variables

Name	Label
cm	CM
flr	FLR
pgnp	PGNP
u1	Residual
u2	Residual

Properties

Variables

Name	Label	Type	Format	Value-label	Missing
cm	CM	int	%12.0g		
flr	FLR	int	%12.0g		
pgnp	PGNP	int	%12.0g		
u1	Residual	float	%12.0g		
u2	Residual	float	%12.0g		

Stata/SE 16.0

History

1. reg cm pgnp
2. predict u2,residual
3. reg flr pgnp
4. predict u2,residual
5. reg u1 u2,noncons
6. reg cm flr pgnp

```

. reg cm flr pgnp
      Source      SS       df       MS       Number of obs   =      64
      F(2, 63)    = 73.83
      Prob > F    = 0.0000
      R-squared   = 0.7077
      Adj R-squared = 0.6981
      Root MSE   = 41.748

      Total      363678       63      5772.66667

      cm      Coef.   Std. Err.   t    P>|t|   [95% Conf. Interval]
      -----+-----
      flr     -2.231586   .2099472   -10.63  0.000   -2.651401   -1.81177
      pgnp     .8055466   .0920833    8.74  0.000   .6214201   .9896731
      _cons   283.4416   11.59318   24.46  0.000   259.7480   307.1352

      . reg cm flr pgnp
      Source      SS       df       MS       Number of obs   =      64
      F(2, 63)    = 116.60
      Prob > F    = 0.0000
      R-squared   = 0.6494
      Adj R-squared = 0.6438
      Root MSE   = 41.408

      Total      301328.534      64      4737.94584

      u1      Coef.   Std. Err.   t    P>|t|   [95% Conf. Interval]
      -----+-----
      u1      1.0000000   .0000000    1.00  0.000   .9999999   1.0000001
      u2     -2.231586   .2065878   -10.80  0.000   -2.644419   -1.818753
  
```

Command

```
reg cm flr pgnp
```

Variables

Name	Label
cm	CM
flr	FLR
pgnp	PGNP
u1	Residual
u2	Residual

Properties

Variables

Name	Label	Type	Format	Value-label	Missing
cm	CM	int	%12.0g		
flr	FLR	int	%12.0g		
pgnp	PGNP	int	%12.0g		
u1	Residual	float	%12.0g		
u2	Residual	float	%12.0g		

So, if you specify that regression, how will you specify? Our dependent variable was CM, then independent variables were FLR and PGNP. This is our complete model that we wanted to run and if we run this, now look at the beauty. What is the coefficient of FLR? minus 2.231. What was the coefficient what we got following the steps?

If you follow, that is also minus 2.231586. So, this coefficient is exactly following with the coefficient what we got earlier. So, that means what we can say? That the theoretically what we have derived and empirically what we have estimated, both are matching. So, in multiple linear regression model, the coefficient that we get is basically the net impact.

So, the statistical software Stata is actually doing everything for us. So, that means all those steps, step 1, step 2, step 3, what we discussed earlier to keep the other factor constant, Stata is basically doing that and Stata is supplying us directly the net impact of each and every variable. That is the beauty of this multiple linear regression model. And that is the beauty of this software. But we must know all the steps, what we actually should do to understand the interpretation of multiple linear regression model in a better way.

If we do not know the steps, then we may not be able to appreciate what we are actually getting from this. This is how a multiple linear regression model is estimated and then, what we will do? As we have learnt the hypothesis testing part and goodness of fit measure previously, we will try to understand all those result from here. First of all, what is the R square here? The R square here is 0.7077. What does it mean? Can you remember? Do you recall the meaning of R square? So, R square is 0.7077.

(Refer Slide Time: 17:40)

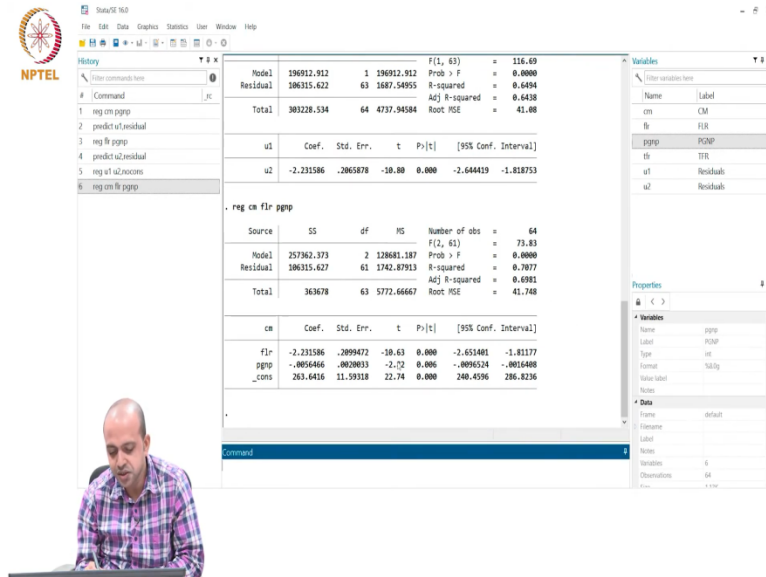
NPTEL

$H_0: \beta_1 = 0$
 $H_0: \beta_2 = 0$

predict u_1 , residual — step 1
 predict u_2 , residual — step 2
 reg $u_1, u_2, nocons$ — step 3

$R^2 = 0.7077$
 $R^2 = \frac{ESS}{TSS}$
 \rightarrow 70.77% of total variation in CM is explained by FLR and P4NP

Significance of the explanatory variables
 FLR $\rightarrow t = |-10.63| = 10.63 \sim t_{n-2}, \alpha = 5\%$
 P4NP $\rightarrow t = |-2.82| = 2.82 \sim t_{n-2}, \alpha = 5\%$
 FLR $\rightarrow p = 0.000$
 P4NP $\rightarrow p = 0.006$
 Both FLR & P4NP are sig at 1% level



First of all, R square is 0.7077. What is the meaning of this? The meaning of this, as R square equals to all, we defined earlier, R square is ESS by TSS. So, that means we can say, the meaning is 70.77 percent of total variation in CM is explained by FLR and PGNP. These are the 2 explanatory variables we have included in our model. So, that means we can say that out of total variation, 70.77 percent of total variation in CM is explained by the 2 explanatory variables FLR and PGNP. And that is what the goodness of fit is- 70.77 percent.

So, our model can explain 70.77 percent variation in child mortality rate. That is the interpretation of R square. This is the first thing we need. Then secondly, the significance of the explanatory variable. How will you check? So, first of all for FLR, what is the t value? Minus 10.63. So, corresponding to FLR, the t value equals to minus 10.63. So, that means what we need to do? We need to take modulus of this. It would become 10.63.

And this is called calculated t value and this we should compare with the tabulated one and what should be the degrees of freedom? Total number of observations is 64 and as we said, the t will always follow a t distribution with n-2 degrees of freedom. So that means 62 degrees of freedom and 5 percent, 1 percent and 10 percent, this we have to compare from the table.

So, alpha equals to 5 percent, we need to get the value. And since this is 10.63, even without comparing also, any value beyond 3 or 4, by rule of thumb we can say that this calculated value would be greater than the tabulated one. And since this is greater than the tabulated one, we have to reject our null hypothesis. What was our null hypothesis?

Here we can say that we have two null hypotheses; first one is β_1 equals to 0 and second null hypothesis is β_2 equals to 0. These are the two null hypotheses. So, that means neither FLR nor PGNP has any impact on child mortality rate. If this is greater than this, then we have to reject the null. We have to say that, yes FLR has significant impact on child mortality rate. This is by level of significance approach.

Similarly, for the PGNP, what is the t value? Minus 2.82. So, that means for PGNP, what is the value? Minus 2.82. This also, you take modulus, then it would become 2.82 and will follow then $n-2$ degrees of freedom, α equals to 5 percent, we need to compare. Since it is 2.82, we need to compare whether it is actually greater than the tabulated or not. So, that means going by this t values, we need to compare with the tabulated value.

But as I said earlier, Stata is also giving us the p value which is called exact level of significance. And look at the p value, p value is 0.000 and 0.006 for PGNP. So, if you multiply by 100, both the values are less than 1. What I said? So, that means p equals to first of all 0.000 for PGNP and p equals to 0.006 for, this is for PGNP and this is for FLR.

So, we have to multiply p with 100 and for both of these cases, this is less than 1. This is 0.0, this is 0.06, both are less than 1. So, that means using the p value, we can easily understand both FLR and PGNP are significant at 1 percent level. So, following the p approach, we do not need to compare this tabulated with the calculated one. So, this is called level of significance approach. Now, what we will do?

(Refer Slide Time: 24:28)

Handwritten Notes:

$TSS = \sum (y_i - \bar{y})^2$
 $ESS = \sum (y_i - \hat{y}_i)^2$
 $RSS = \sum (\hat{y}_i - \bar{y})^2$
 $RSS = R^2 \cdot TSS$

Confidence Interval Approach:

FLR: $(-2.65) - (-1.81) \rightarrow$ 95% confidence interval
 $\beta_1 = 0$
 - estimated interval does not capture 0
 \Rightarrow Reject $H_0: \beta_1 = 0$

PGMP: $(-0.009) - (-0.001) \rightarrow$ 95% confidence interval
 - estimated interval does not capture 0
 \Rightarrow Reject $H_0: \beta_2 = 0$

Statistical Software Output:

Model	SS	df	MS	F(1, 63)	Prob > F
Residual	106315.622	63	1687.54955		0.0000
Total	303228.534	64	4737.94584		0.6494

Variable	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
u1					
u2	-2.23586	.2065878	-10.80	0.000	-2.64419 -1.818753

Source	SS	df	MS	Number of obs
Model	257362.373	2	128681.187	64
Residual	106315.627	61	1742.87913	
Total	363678	63	5772.66667	

Variable	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
cm					
flr	-2.23586	.2099472	-10.63	0.000	-2.651401 -1.81177
pgmp	-.0054566	.0020813	-2.62	0.000	-.0095251 -.0014080
_cons	263.4416	11.59318	22.74	0.000	240.4596 286.8236

We will see the confidence interval approach. There, corresponding to FLR, what is the confidence interval that we got? The confidence interval is minus 2.65 to minus 1.81. That means we can easily understand this particular interval does not capture the hypothesized value of the population parameter. What was the hypothesized value? We have hypothesized beta1 is 0.

So, this interval does not capture 0. So, that means estimated interval does not capture 0. Then what is the implication? Reject the null hypothesis since the interval fails to capture the hypothesized value of the population parameter. So reject H_0 which is beta1 equals to 0. That means, when we reject this, we will say that beta 1 is actually significantly different from 0. So, that means FLR has significant impact on child mortality rate.

Then what is the interval for PGNP? PGNP's interval is minus 0.009 to minus 0.001. This interval also, since both are negative, it does not capture 0. So, that means estimated interval does not capture 0. So, reject H_0 which is β_2 equals to 0. And when it is rejected, we will say that even PGNP also has significant impact on child mortality rate.

But following this interval estimation approach, we can only say that whether the variable is significant at 5 percent level or not because this is 95 percent confidence interval. So, Stata will always give only 95 percent confidence interval. This is also 95 percent confidence interval. If we want to know whether it is significant at 1 percent, obviously if it is 5 percent, then automatically it would become significant at 10 percent, it is absolutely no problem as we said earlier. But whether it is significant at 1 percent level or not, to understand that, we need to construct the interval at 99 percent confidence interval.

But following the p value, we can easily say that it is significant even at 1 percent level. So, that is the advantage of using p value over t and this confidence interval. This is how we can interpret the coefficients actually. Now, what about this ANOVA table? From this ANOVA table again, as we said this particular table is called ANOVA- Analysis of Variance. So Stata is again supplying model sum of square, residual sum of square and total sum of square. That means when the total sum of square is basically this, 363678, that is called total sum of square or TSS.

TSS equals to 363678, that is summation y_i minus \bar{y} whole square. Out of this, how much our model is able to explain? 257362.373, that is called ESS or in Stata's language it is called model sum of square. And what is the remaining? Remaining portion is called residual sum of square or RSS that is, 106315.

And then, again corresponding to degrees of freedom, what would be the degrees of freedom for TSS? Total number of observation is 64 and once again I would like to remind you that TSS is basically summation y_i minus \bar{y} whole square. Since there is only one restriction imposed, this is called degrees of freedom for TSS and should be $n-1$. And what is the degrees of freedom for RSS?

The RSS equals to y_i minus $\hat{\alpha}$ minus $\hat{\beta} x_i$ square. Our model was y_i equals to α plus βx_i plus u_i . So, I would like to remind you, before we estimate RSS, we must estimate $\hat{\alpha}$ and $\hat{\beta}$. So, we are putting 2 restrictions here in terms of $\hat{\alpha}$ and $\hat{\beta}$,

that is why degrees of freedom for RSS would be $n-2$. Or in general, we can say that this is $n-k$ where k is total number of parameters to be estimated from the model.

And then, you can easily say that degrees of freedom for ESS would be $(n-1) - (n-k)$ which is equal to k . k is the degrees of freedom where k is the total number of parameters. So, in this model, we have 3 parameters to be estimated, alpha, beta and this constant term. That is why you see the degrees of freedom for model is $n-k$, k equals to 3, 1 2 3, that is why it is 61. $(64-3)$

And if you take this 63 minus 61 equals to 2. Is this clear? k equals to 3 here, because three parameters we are going to estimate; two regression coefficient and one constant term. That is why RSS degrees of freedom is n minus k . 64 minus 3 equals to 61 . And then 63 minus 61 equals to 2 . So, this is how we have to understand the ANOVA table which clearly shows the decomposition of total variation in our y_i .

But one sum of certain other measures which are also supplied by Stata which is another F statistic, look at the F statistic value which is 73.83 and then, another R square which is called adjusted R square. These things we have not yet discussed. We will understand the importance of F statistic and adjusted R square in our next class. So, in our next class, we will discuss how to use these measures; F statistics and R square, adjusted R square which is also supplied by Stata. That way, will discuss in our next class. Thank you.