Introduction to Econometrics Professor Sabuj Kumar Mandal Department of Humanities and Social Sciences Indian Institute of Technology Madras Lecture 22 Multiple linear regression model and application of F Statistics Part - 1

(Refer Slide Time: 00:14)



PTEL	risery		Model	257362.373	2	1/8681.18	Prob	> F	0.0000	 Variables 		
	 Filter commands here 	0	Nestonal	100313.02/	01	1/42.8/3	— Adj P	R-squared :	0.6981	K Filter variables	here	
	# Command	31	Total	363678	63	5772.6666	7 Root	MSE :	41.748	Name	Label	
	2 reg cm fir									fr	ER	
			cm	Coef.	Std. Err.	t	P> t	[95% Conf.	Interval]	pgnp	PGNP	
			flr	-2.231586	.2099472	-10.63	0.000	-2.651401	-1.81177	tfr	TFR	
			_cons	0056466 263.6416	.0020033 11.59318	-2.82 22.74	0.000	0096524 240.4596	0016408 286.8236			
		. гед	g cm flr									
			Source	SS	df	MS	F(1,	er of obs = 62)	64 125.65			
			Model Residual	243515.049 120162.951	1	243515.04	9 Prob	> F = =	0.0000			
							Adj F	R-squared :	0.6643	Properties		
			Total	363678	63	5772.6666	7 Root	RSE	44.024	# Variables		
		-	(8	Conf	Std Een		P> +	Inter Conf	Tetervall	Name		
			ciii	coer.	scu. err.	· ·	estel.	[95% CONT.	Incerval	Type		
			flr	-2.390496	.2132625	-11.21	0.000	-2.816802	-1.96419	Format		
				24510455	ALC LL TOP	11.50	01000			Notes		
										* Data	default	
	and the second se									* Filename		
	2 3	Comm	ound							1 Notes		
	2.5	CONTRACT OF CONTRACT OF CONTRACT	10110	9						Variables	4	
1										Observations	64 240	
	RatyS 162											- 6
	StaySE 160 He Edt Data Graphics St	anisics Uber Window	Help									- 6
Annual I	E StatuSE 160 Re Edit Data Graphic St P E E E E E E E E E E E E History	atistics User Window	Help				— F(1,	62)	125.65	∧ Variables		- 8 T
"Innull	E Steavier Stad The fast pran Conprise St The fast pran Conprise St The State State Mark Constraints free	atitic ber Window	Help Nodel Recidual	243515.049	1	243515.04	- F(1, 19 Prob	62) =	125.65 0.0000 0.6595	* Variables	here	- 8 T
A A A A A A A A A A A A A A A A A A A	Status E 163 Re Edit Data Condition Si Re Edit Data Condition Si Reconstructures Reconstructures Per constructures Per co		Help Nodel Residual	243515. 0 49 120162.951	1 62	243515.0/ 1938.1121	— F(1, 19 Prob 11 R-squ Adj F	62) = > F = uared = R-squared =	 125.65 0.0000 0.6596 0.6643 	Variables Variables Variables Name	here Label	- 8 T
	StatyS 100 The fat this cristics is The fat this cristic is The fa	unitic Urr Wirdw	Nodel Residual Total	243515.049 120162.951 363678	1 62 63	243515.04 1938.1121 5772.6666	— F(1, 19 Prob 11 R-squ — Adj F 17 Root	62)	 125.65 0.0000 0.6596 0.6643 44,624 	 * Variables Variables Intervandees Name on	here Label OM	- 8 T
	Stanyt Hol The Edit Date Conject IS The Edit Date Conject IS The Contract IS The Contr		Hep Model Residual Total	243515.049 120162.951 363678	1 62 63	243515.04 1938.1121 5772.6660	— F(1, 19 Prob 11 R-squ Adj F 7 Root	62) = > F = ared = R-squared = NSE =	 125.65 0.0000 0.6696 0.6643 44.024 	 Variables Toto resultion Name cm fit open 	here Label CM FRR pecup	- 8 T
	SeqUESS The fait this depice is the fait the depice is the depice is		Hep Nodel Residual Total cm	243515.049 120162.951 363678 Coef.	1 62 63 Std. Err.	243515.04 1938.1121 5772.6666 t	- F(1, 19 Prob 11 R-squ - Adj F 77 Root P> t	62)	 125.65 0.0000 0.6596 0.6643 44.024 Interval] 	➤ Variables Name Om fr pgsp tfr	here Label CM FLR PGNP TLR	- 8 T
	Beq.0 56	Little: Urr Window	Model Residual Total cm cons	243515.049 120162.951 363678 Coef. -2.390496 263.8635	1 62 63 Std. Err. .2132625 12.22499	243515.04 1938.1121 5772.6660 t -11.21 21.58	- F(1, 19 Prob 11 R-squ - Adj F 57 Root P> t 0.000 0.000	62) = = = = = = = = = = = = = = = = = = =	 125.65 0.0000 0.6696 0.6641 1.16419 288.3009 	 Variables The variables Inne ars fit popp the 	here Label CM FLR PGNP TFR	- 8 T
	Bayl 13 Bayl 13 Bayl 14 B	atific Lier Window	Model Residual Total cm cons g cm pgap	243515.049 120162.951 363678 Coef. -2.390496 263.8635	1 62 63 Std. Err. .2132625 12.22499	243515.04 1938.1121 5772.6666 t -11.21 21.58	- F(1, 19 Prob 11 R-squ - Adj F 57 Root P> t 0.000 0.000	62) = = = = = = = = = = = = = = = = = = =	 125.65 0.0000 0.6543 44.024 Interval] -1.96419 288.3009 	× Valdes Vane Name on 1999 fr	here Label CM FR PGSP TFR	- 5 T
	succession s		Nodel Residual Total cm flr _cons sg cm pgrp Source	243515.049 120162.951 363678 Coef. -2.330496 263.8635 SS	1 62 63 Std. Err. .2132625 12.22499 df	243515.04 1938.1121 5772.6666 t -11.21 21.58 7/5	F(1, 19 Prob 11 R-sq: Adj F 77 Root P> t 0.000 0.000 Numbo	62) = = = = = = = = = = = = = = = = = = =	 125.65 0.0000 0.6593 0.6543 44.024 Interval] -1.96419 288.3009 644 	 Valables Pre-strates Name Rr prop- fr fr off 	here Label CM FLR PGNP TFR	- 6 T
	E Sections The fact has device as the Comment 1 region (page 2 region (page) 3 region (page)		Hep Model Residual Total cm flr _cons g cm pgrp Source	243515.049 120162.951 363678 Coef. -2.390496 263.8635 SS	1 62 63 Std. Err. .2132625 12.22499 df	243515.0 1938.1127 5772.6666 t -11.21 21.58	- F(1, 19 Prob 11 R-squ 77 Root P> t 0.000 0.000 F(1,1)	62)	 125.65 0.0000 0.6643 44.624 Interval] -1.96419 288.3009 64 12.36 0.0001 	* Woldes Networks Network Meme on 500 de	latel CM FIR PGNP TFR	- 8 Tł
	E 8ag(2 kg) Re 64 Das Gapta Si E 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2		Nodel Nodel Total Cem flr _cons g ca pgap Source Nodel Residual	243515.049 120162.951 363678 Coef. -2.390496 263.8635 55 60449.4605 303228.539	1 62 63 5td. Err. .2132625 12.22499 df 1 1 62	243515.00 1938.1127 5772.6666 t t -11.21 21.58 66449.464	- F(1, 19 Prob 11 R-squ P> t 0.000 0.000 F(1, 15 Prob	62) > F = Jarred = RSE = [95% Conf 2.816802 239.4261 = 62) = 62) = F = Jarred = 62 = F = Jarred = F =	 125.65 0.0000 0.6696 0.6643 1.96419 288.3009 12.36 0.0008 0.1662 	× Valdes Normalities norm popp th boundary	hore Label CM FIR FGSP TFR	- 8 *1
	E sect ha Te te the Costen S C sect the section of the section o		Model Residual Com flr _cons source Model Residual Total	243515.049 220162.951 363678 Coef. -2.390496 55 55 55 66449.4605 53 363678	1 1 62 63 5td. Err. 2132625 12.22499 df 1 1 62 63	243515.04 1938.1121 5772.6660 t -11.21 21.58 905 66449.466 4599.722 5772.6660	 F(1, 19 Prob Prob Ps)[t] Adj fi Prob Adj fi Adj fi Adj fi Root 	62) = = = = = = = = = = = = = = = = = = =	 125.65 0.0000 0.6693 0.6643 44.024 Interval] -1.96419 288.3009 64 12.36 0.0063 0.1662 0.1562 0.1562 0.934 	 A Value Name Name Mare Mare SPD dr 	Intere Label CM FIR PGNP TFR	- 8 T
	Sancti H3 Tre fait no Gapto II Tr		Model Residual Total Com fir _cons g cm pgrp Source Nodel Residual Total	241515.6.49 228152.951 343678 Coef. -2.396496 62849.6635 55 55 55 66849.46635 363678 363678 Coef.	1 62 63 5td. Err. ,2131625 12,22499 df 1 62 63 5td. Err.	243515.00 1938.1127 5772.6664 t -11.21 21.58 //55 5772.6664 4599.7222 5772.6664	- F(1, 19 Prob 11 R-squ 7 Root P> t 0.000 0.000 10 R-squ 10 R-squ	62)	 125.65 0.0000 0.6643 0.6643 44.024 1.56419 283.3009 0.16528 0.9008 0.1528 0.9334 1.1528 1.1528<td> ★ Wolds Non-solids Name BOP dr BOP dr SOP dr A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid</td><td>here Label CM FLR PGNP TFR TFR</td><td>- 6 T</td>	 ★ Wolds Non-solids Name BOP dr BOP dr SOP dr A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid A solid	here Label CM FLR PGNP TFR TFR	- 6 T
	E Sand' Hol Te tak Das Gapta Si E tak Das Gapta Si E tak Das Gapta Si E tak Das Gapta Si E tak Das Gapta E tak Das Gapt		Nodel Residual Total Can fun ga m pgop Source Nodel Residual Total Can Residual Total	243515.049 228163.951 363678 00ef. -2.390496 66049.4605 55 66049.4605 363578 3635678 00ef. -0113645	1 62 63 5td. Err. .2131625 12.22499 df 1 62 63 5td. Err.	243515.00 1938.1127 5772.6660 t -11.21 21.58 7572.6660 4590.7222 5772.6660 t t	 F(1, 1) Prob R-squ Adj F P> t 0.000 0.000 F(1, 1) R-squ Adj F F(3, 1) R-squ Adj F R-squ Adj F R-squ P> t 0.000 	62) → F → wared → F → NSE → [553 Conf- 62) → F → 239.4061 → F → 239.4061 → F → RSE → F → RSE → F → (553 Conf- RSE → (553 Conf- (553 Conf- (125.65 9.4000 9.6643 44.624 1.15419 283.3009 64 9.1528 69.534 1.15419 1.1541	 ≤ Velable Non-watches Non-watches non- non-	hore Label CM FLR PGAP TFR TFR	- 6 T
	Sanget Hol In the first non-dependence In the fir		Nobel Residual Total Cas fur cas g cn pgro Source Nobel Statal Total Cas Cas Cas Cas Cas Cas Cas Cas Cas Cas	243515.649 28162.951 363678 Coef, -2.39456 263.8635 55 566449.4665 380228.539 380228.539 383678 Coef, 411365	1 62 63 5td. Err. .2137625 12.22499 df 1 62 63 5td. Err.	243515.04 1938.1121 5772.6664 t -11.21 21.58 9/5 66449.466 4599.7822 5772.6664 t t -3.52 15.99	- F(1, 19 Prob 1 R-sqs - Adj f - Adj f - Adj f - 0.000 0.000 - F(1, 15 Prob 9 R-sqs - Adj f - Add f	62) → F +	 125.66 9.000 9.600 9.601 9.611 9.641 9.641 9.643 9.644 <li< td=""><td>> Valaba Non contain Name em fr Page fr fr</td><td>here Itabel CM FR PGP TFR TFR</td><td>- 8 T</td></li<>	> Valaba Non contain Name em fr Page fr fr	here Itabel CM FR PGP TFR TFR	- 8 T
	E sact Hd Te ta to o Gate d E ta to constant E constant Constant 2 regon pop		Nodell Residual Total Come g ca papo Residual Total Total Source Nodel Total Come Source Nodel Source Nodel Come Source S	243515. 649 228152. 951 363678 Coef. 55 55 66049. 4605 55 66049. 4605 55 Coef. . e8113645 157. 4244	1 62 63 5td. Err. ,2132625 11.22499 df 1 63 5td. Err. ,0032325 5td. Err.	243515.00 1938.1127 t -111.21 21.58 60449.66 4590.7628 5772.6666 4590.7628 t t -3.52 5.59	 F(1, 19) Prob 11 R-sq; R-sq; Root P> t 0.000 0.000 0.000 R-sq; R-sq;	62) → F → arred → F R5E + Capared → R5E + Ca	 115.65 0.0000 0.6695 0.6695 4.6.024 1.05413 2.83.3009 2.83.3009 2.83.3009 2.83.534 1.528 2.9.554 1.528 2.5.554 1.528 1.528	 ▲ Wolds ▲ The worlds Nume BOP 6 Popelies â < >> b < >> b < >> b < >> c < > c < >> <pc <="">> c < >> <pc <="">> c < >> <pc <="">> c < >> c <</pc></pc></pc>	Nors Lubel CM FR 700P TR	- 8 TI
	E sand to3 Te ta no data si Te ta no data si		Nopel Residual Total Cene Cene Cone Cone Cone Cone Residual Total Total Cone Cone Cone Cone Cone Cone Cone Cone	243515.649 228162.951 363678 Coef. -2.390496 55 55 566449.4605 55 363678 363678 Coef. -0113645 197.4244	1 62 63 5td. Err. .213629 df 1 2 63 5td. Err. .4092225 5td. Err. .4092225 9,845583	243515.04 1938.1127 5772.6664 t -11.21 21.58 775 66444.04 4598.7822 t t -3.52 15.99	- F(1, 19 Proto 17 Rotot P> t 8.000 - Kont P> t 8.000 - Kont P> t 8.000 - Kont - Kont	63) γ γ =	 125.65 8.000 8.663 44.624 44.624 1.55419 1.55439 8.1625 6.9.534 1.1647-0.11 2.1.52 2.1.52<	 ▲ Voidels ▲ Normalise ■ Normalise ■ Normalise ■ Poperion ● O ■ Normalise ■ O ■ Normalise ■ O ■ O<td>ter Label OM FR FR FR FR FR FR FR FR FR FR</td><td>- 8 </td>	ter Label OM FR FR FR FR FR FR FR FR FR FR	- 8
	 RacK 10 RacK 10		Nobel 1 Residual 1 Total 1 Centre 1 Source 1 Nobel 1 Total 1 Total 1 Total 1 Total 1 Total 1 Centre 1 Source 1	241515.649 228162.951 363678 Coef. 2283.8635 55 55 560449.4685 30222.539 363678 Coef. e8113645 157.4244	1 62 63 5td. Err. .2132625 12.22499 df 1 62 63 5td. Err. .0022225 5td. Err.	243515.04 1938.1127 5772.6666 t -11.21 21.58 905 68449.466 4599.7823 5772.6666 t t -3.52 15.99	 F(1, 19 Prof. F-sqi. Adj F Rost P> t 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 	62) + F + squared + - squared + (5% Conf, - (5% Conf,	 125.65 8.000 8.663 8.6643 44.624 44.624	A Vatales None order None None Poper fr fr Poper fr fr Poper fr fr	Not Laboratory of the second s	- 8
	Sack Hd Re dd ha o Garle d Gomad Gomad gan pop gan		Hep Model 1 Total 0 Car 2 Car 2 Car	243515.649 228162.951 363678 Coef. -2.39496 68494.4605 55 55 68494.4605 263.8635 363678578 36	1 62 63 5td. Err. .2132625 12.22499 df 1 62 63 5td. Err. .213625 83 5td. Err. 	243515.00 1938.1121 5772.6666 t -11.21 21.58 905 5772.6666 4590.7422 5772.6666 t t -3.52 5772.6666	- F(1, 9 Prob 1 R-sq. - Adj J 77 Root - F(1, 15 Prob 9 R-sq. - F(1, 15 Prob 9 R-sq. - Root - P) t 0.000 - P) t 0.000	62) 1 → F - t-squared 1 (555 corf of the squared 1 (553 corf of the squared 1 562 - 574 control of the squared 1 575 corf	 125.65 0.000 0.6643 0.6643 1.6643 1.6643 1.6643 2.83.3009 64 1.2643 2.83.3009 1.6643 1.6643	 ▲ Wolds. ▲ The solution ■ Non-solution ■ Non-solution	box Label CM TR TR TR TR TR	- 8 • 1
	E sack al Te ta ta no deale si E de la constant I c		Nopel Residual Total Total filr filr filr filr filr filr filr fil	243515.049 363673 363678 50 55 55 55 56049.4605 53 363678 303222.539 363678 55 56049.4605 55 56049.4605 51 52 53 53 53 53 53 53 53 53 53 53 53 53 53	1 62 63 5td. Err. 2135635 12.22499 6f 1 1 62 63 5td. Err. @823235 5td. Err.	245515.04 1938.1127 5772.6666 t -11.21 21.58 105 66449.466 4599.7827 t t -3.52 15.99	- F(1, 19 Frob 11 R-45(77 Root - F(1, 9.000 9 R-30 9 R-30 9 R-30 9 R-30 9 R-30 9 R-30 9 P> t 0.000 P> t 0.000	62) > F - squared - squared - squared - squared - 7.85892 2 39.4281 - 7.85892 2 39.4281 - 7.85892 - 7.8592 - 7.952 - 7.9522 - 7.9522 - 7.9522 - 7.9522 - 7.	 125.65 0.000 0.6693 0.6643 44.627 1.96419 1.96419 1.643 0.1528 65.934 0.1528 65.934 Interval] Interval 	 ▲ Walks ▲ Walks ▲ Walks ■ New ■ New ■ SPD fr SPD fr fr fr	Iver Ladd CM CM RA POP RR RA A A A A A A A	- 8 T.

Welcome once again. So we were discussing about multiple linear regression model in our previous class and the example what we were talking about is child mortality rate (that was our dependent variable) equals to beta 0 plus beta 1, female literacy rate, FLR and then beta 2 per capita GNP plus ui. And this is a state level data.

So, in our last class we learnt how to estimate this multiple linear regression model and how to interpret this beta 1 and beta 2. Once again I will just repeat. Once you estimate this beta 1 hat, then the interpretation of beta 1 hat is basically, for unit change in FLR, child mortality rate changes by beta 1 hat amount, keeping the impact of PGNP constant.

And we have also learnt how to keep the impact of PGNP constant in different steps and then we saw that actually when we keep the impact of PGNP constant following several steps and when we estimate the model at one go including all the variables, the magnitude of the beta hat turns out to be the same. That is how we justified the multiple linear regression model interpretation.

So, today what we will do? We have also learnt previously the justification of FLR from the point of view of omitted variable bias. So, that means even if our objective is to estimate the impact of FLR only on child mortality rate, do I need to include PGNP also in the model? And we said that yes, even if our objective is to estimate the impact of FLR on child mortality rate, we should also include PGNP in the model so as to avoid omitted variable bias.

Now, with the same example and same data set, we will just see the severity of omitted variable bias in an empirical set up. So, let us see the omitted variable bias. So, we will take the same example. So, first what we will do? We will run the complete model which is given by CM and then FLR and PGNP. This is our complete model. And what is the model of FLR on PGNP? The coefficient of FLR is minus 2.231.

So, that means from the complete model, beta 1 hat equals to minus 2.231. Now, the next model, what we are going to estimate is CM equals to beta0 plus beta1 FLR plus let us say u1. And from this model, the estimated coefficient we will say beta1 tilde, as we have mentioned theoretically also, we see what is the value of beta1 tilde and then we will examine whether beta1 tilde is actually equals to beta1 hat or not. If they are equal, then we will say that there is no bias involved, beta 1 tilde now we are going to estimate.

So, we will include only CM (child mortality rate) on FLR. So, that means we are not including PGNP in the model. Now, we will see what is the coefficient? Coefficient is minus 2.39. So, we can easily understand from this that beta1 hat, that means this implies beta1 hat is actually not equals to beta1 tilde. It is 2.231, this is minus 2.39, that is increase in the value of the beta1 tilde. So, that means there is a bias when we include, when we do not include the model, other relevant variable in the model.

Similarly, what we can do? We will look at the estimate of PGNP also from the complete model, what is the beta2 hat? So, beta2 hat from the complete model, beta2 hat equals to, that means from PGNP is minus 0.005. And then, we will include, we will try to estimate beta2 tilde also

from the incomplete model and that shows, so we will include PGNP as explanatory variable and what is the value? Minus 0.01. So, that means we can say that beta2 hat is not equal to beta2 tilde, rather than beta2 tilde becomes almost double.

So, with this example, we can understand the severity of this omitted variable bias when we do not include an important variable, explanatory variable in the model. So, that means that gives you the proper justification of multiple linear regression model. Even if our objective is to estimate the impact of a particular variable in the model, we must include other relevant or important variables as control in the model. Otherwise our model will suffer from omitted variable bias and this example clearly demonstrate the severity of the omitted variable bias.

(Refer Slide Time: 08:41)

C.M. = (70.774.) @





Now, one more thing that also happens when we do not include, what happens to the goodness of fit? So, from the complete model, beta0 plus beta1 FLR plus beta2 PGNP plus ui. So, from the complete model we will try to get the R square and we will call it as R square 1. So, what is the R square value from the complete model, the goodness of fit? If we look at the complete model, then we see the R square value is 0.7077.

And then, when we run the other 2 models, CM i equals to let us say beta0 plus beta1 FLR plus u1, what is the R square value? Let us say this is called R square2, and then we will run this model also, CM equals to beta0 plus beta1 PGNP plus u2 and we will call this R square as let us say R square3. So, we will now report the R square2 and R square3 and we will see.

So, when we include only FLR in the model and then we see the R square value is 0.6696, this is 0.6696. And then when we include only PGNP in the model, then we get the R square value is 0.1662. Now, in a complete model, R square value is 0.7077, that means we can say that FLR and PGNP, these two variables together, they can explain 70.77 percent variation in child mortality rate. So, from this, what we can say that, FLR and PGNP together can explain 70.77 percent variation in child mortality rate.

Now, a natural question comes to our mind; out of these 70.77 percent explanation, how much of this is explained by FLR and how much of this is explained by PGNP? So, one obviously would tend to say that out of this 70.77, probably 66.96 percent is explained by FLR and remaining would be PGNP. But if you look at, 70.77 minus 66.69 is not actually 16.62 percent.

So, that means we cannot actually allocate the R square among these two explanatory variables. What I say? I will just repeat here. PGNP and FLR together can explain 70.77 percent variation in child mortality rate because that is the R square from the complete model. Now, a natural question comes to our mind. Out of this 70.77 percent, how much percentage of this R square is due to FLR and how much of this R square is due to PGNP? Can we allocate this 70.77 percent among these two variables?

So, if we would like to allocate these 70.77 among these two variables, so when we include FLR in the model, then we see that 60.96 percent variation is explained by FLR itself. So, that means 70.77 percent minus this percentage should be explained by PGNP, but actually if we look at, PGNP is explaining much more than that. So, that means we cannot say, we cannot actually, we cannot, can we allocate R square which is 70.77 percent among the two explanatory variables? The answer is actually no, we cannot.

Why this is no? Because you see, if you allocate 60.96 percent on FLR, then the remaining 70.77 minus this much should have been explained by PGNP, but PGNP is explaining much more than that. So, that means it is actually not possible to allocate this R square. Why? Because of the intercorrelation among the two explanatory variable, if you calculate this intercorrelation, among these two you can see what is the intercorrelation between PGNP and FLR, you can easily calculate that.

What is the command for that? corr PGNP and FLR if you give, then you see that 0.2685 is their intercorrelation. Because of this intercorrelation, we cannot say how much percentage of this R square is due to FLR and how much percentage of this R square is due to PGNP. That is one thing we have to keep in mind.

But one thing what we observe that compared to the partial models, when we include one more variable in the model, R square is actually increasing. So, whenever, let us say that this is our initial model CM is the dependent variable and FLR is the independent one. And then R square is 0.6696. If we include PGNP also in the model, then the complete R square, complete model gives R square equals to 70.77.

So, adding one extra variable is giving additional R square in the model. Similarly, when we have only PGNP in the model, R square is 0.1662, however when we add FLR in the model, that

gives R square equals to 0.7077. Now, the question here that comes to our mind; can we compare these two R square, that means can we compare R1 square and R2 square or R1 square and R3 square and say that R1 square is better, let us say this is our model 1, this is model 2, this is model 3.

So, can we compare R1 square and R2 square or R1 square with R3 square and say that model 1 is actually greater than both model 2 and 3? This is a question; this is an interesting question. So, compared to these 2 partial models, our complete model gives better R square. So, since complete model gives better R square, when we compare 2 and 1, we see that there is some change in the R square value, that means explanatory power of the model increases when we add PGNP also in the model, because R square increases from 0.6696 to 0.7077.

Similarly, compared to 3 and 1, when we add FLR also in the model, R square increases from 0.1662 to 0.7077. Because of this comparison, can we say that model 1 is actually better than model 2 and 3? What we are posing is can we compare R1 square with R2 square and R3 square? And say and decide which model is better than the other one? The answer is again no. We cannot decide the suitability of a particular model based on only R square. What is the reason? Let us try to explain that.

(Refer Slide Time: 20:37)

So that means can we involve ourselves in a game, I will call it game, in a game of maximizing R square? Apparently, it may look like we will keep on adding variables in the model as long as

our R square increases. So, we are basically trying to compare these two models, CMi equals to beta0 plus beta1 FLR plus beta2 PGNP, let us say this is ui and then, the second model CM equals to beta0 plus only beta1 FLR plus let us say u1i. These are the 2 models we are comparing.

So, if we add additional variable in the model, R square will definitely increase. Why this is so? What is the measurement of R square? R square is basically ESS by TSS and this we can again manipulate as TSS minus RSS and then divide it by TSS which is equals to 1 minus RSS by TSS. And what is RSS? RSS as we know, summation of ui hat square divided by, what is TSS? yi minus y bar whole square. This is basically our R square.

Now, when we add one extra variable in the model, what will happen to our ui hat square or RSS? The residual sum of square will definitely decrease or at least, it will not increase. So, that means when an additional, when an extra variable is added in the model, RSS which is actually equals to summation ui hat square, RSS will decrease or at least it will not increase.

And what will happen to TSS? But, TSS remains unchanged. So, if the denominator remains unchanged and the numerator decreases due to an additional variable in the model, what will happen to R square? So, R square therefore, R square will increase when additional variable in the model. So, we can say that R square is basically a non-decreasing function of any additional or extra explanatory variable.

So, this says R square is a non-decreasing function of extra explanatory variable. So, this is the thing that happens. When we add additional variable, explanatory power of the model increases, R square will increase. But, does that mean that we will keep on adding variable in the model and we will say that this model is better than this one because this gives higher R square value. In this model R square equals to 0.70 and here it is R square equals to 0.66. We cannot say that this model is better.

Why? Because this model involves, these two models, these two R square are not directly comparable. Why this is so? Because in the first model, we have two explanatory variables while in the second model, we have only one explanatory variable even though the dependent variable of both the models are same, that is child mortality rate.

So, when two models involves different number of explanatory variable, how can you say that this model is better than this? Yes, we are getting additional R square but in the process we are also losing something. In this world nothing comes free of cost. So, that means a natural question comes to our mind. This additional explanatory power comes with the cost of what? That is the question.

So, the question that we would like to answer here, the incremental value in R square, what is the cost of the increment in R square? This is an important question that I am asking. Is there any cost involved? When we add one extra variable to experience little higher level of R square? The cost is there and what is the cost? If we recall, every additional variable we include in our model, we use extra degrees of freedom.

If we recall that if you include too many variables in the model, then the degrees of freedom, based on which your beta1 hat, beta2 hat, these parameters are estimated, would be lesser and lesser. And if you have too few observations freely moving to estimate your parameters, then the reliability of the parameters comes down drastically.

So, that means the cost is in terms of, what is the cost? Cost is lower degrees of freedom based on which, parameters are estimated. This is the cost. The more variable we add in the model, lesser and lesser would be degrees of freedom in that model, based on which our parameters beta1 hat, beta2 hat and so on, they are estimated. Lower the degrees of freedom, lower would be reliability of the parameters.

So, that is one thing we keep, we have to keep in our mind. So, that means we need to give some penalty to the R square measure while adding the additional variable in the model. And what is the penalty that we give? The penalty that you have to give is the degrees of freedom. That means we have to adjust this R square, that means the numerator and denominator with their corresponding degrees of freedom. What is the degrees of freedom for this RSS as you know? This is n minus k.

What is the degrees of freedom for the denominator? This is n minus 1. That means now from this we can understand that since the numerator involves n minus k, as you include more variables in the model, k will also increase. And as a result of which this numerator will also come down, R square will also come down. That means R square will increase but not like the

previous cases. Earlier, it was increasing more but now I have adjusted the R square, that means the denominator and the numerator both, by their corresponding degrees of freedom.

Since the denominator is not depending on the parameters, see here it is n minus k, so that means it remains again unchanged. But the numerator, it is the degrees of freedom n minus k, more the explanatory variable you add in the model, more would be the value of k, so that means n minus k will decrease and that means that this summation ui hat square will come down. And that gives some penalty on the addition of explanatory variable.