Handling Large Scale Unit Level Data Using STATA Professor Pratap C. Mohanty Department of Humanities and Social Sciences Indian Institute of Technology, Roorkee Lecture 33 Binary Response Models - II

So, welcome once again to the NPTEL MOOC module on Handling Large Scale Unit Level Data using STATA. This is our lecture in continuation to the previous lecture on binary response models. Last lecture specifically we addressed the problems of the very foundation model, very basic model called linear probability model what are the probabilistic structure of the dependent variable is attached with 1 and 0.

But there are a number of limitations while we apply the OLS method. And we have already clarified that it has problem with R square. It also exceeds the limit of the R square and then it is problem with non-normality. It is problem with standard errors and where the BLUE properties are not fulfilled. So, in that continuation we are going to test all those things with the help of data.

(Refer Slide Time: 01:35)



The data we are going to consider is through our practice data. We have filtered from our 73rd round of National Sample Survey 2015-16. We are also interested in getting those factors that explain the choices of women being necessity entrepreneurs. So, as we have already given you the background during our data and introduction to National Level Data where we defined what

do we mean by necessity entrepreneur, based on the number of workers within the entrepreneurs, and necessity entrepreneurs whether women is operating within that structure or not. And accordingly we are going to emphasize. So, when we have necessity entrepreneur or non-necessity entrepreneur or opportunity entrepreneurs we coded into two, in a dichotomous format that is simply called a dummy variable. A mixture of qualitative and quantitative variables are used in the explanatory variables.

(Refer Slide Time: 02:36)



The purpose here is to show how to use various data analysis commands. It does not cover the whole research process such as verification of assumptions, models diagnostics and various follow-up analysis. We just wanted to mention how we can best use the commands.

(Refer Slide Time: 02:56)



Given our model we have simplified through the existing variables. As highlighted in the dependent variable and mentioned already that our purpose is to understand the factors, explanation behind the enterprise to be necessity type with 1 if it is necessity based and 0 if it is non-necessity or opportunity-based.

We have broadly 3 types of variables included, continuous variable, categorical variable and dummy variables. Continuous variable we have in the model, we are going to show it in STATA also. Enterprise age, square of that age, square of the age can be taken as I already mentioned because it might be nonlinear with the variable but not with the coefficient. So, squaring of a variable is not going to break the relationship of the model between dependent and independent.

So, dummy variables like sectors, rural, urban; mixed activities whether they are having or not that is we simply coded as 0 and 1; any problem faced and not assistance received or not; whether they are traditionally operating or non-traditional format. Similarly categorical variable like social category; we have activity group etc.

(Refer Slide Time: 04:32)



Coming to the understanding through the data we are going to open the data right now before you.

(Refer Slide Time: 05:10)



I am going to switch up to the STATA window and in the STATA window I am going to open the data only. So, that data is here. So, it is here. In the computer I have that data and we will provide the data to you also. I am going to open this binary choice model data before you. The variable as we just said is going to be discussed right now. Like we have the variables like enterprise type

that we discussed, other binary groups like sector dummy, their assistance received or not, similarly there are ordinal, some categorical variable also, categorical variable like social categories, social groups are there.

Now, what we wanted to understand is the following. After opening the data we wanted to simply describe it. How to describe it then? Simply type des then. Okay if you type des it will show you. So, simply type des, describe is going to give the nature of the data. It talks about whether your data storage is in byte form or in string form.

(Refer Slide Time: 06:10)

And a control of con	T 4	X . des				 Variables 		
Contrast Image: Status	commands here					+ Effer variable		
(Autom) (Autom) <t< td=""><td>and the second</td><td>contains data from G: V</td><td>Dinary choi</td><td>ce model/BC</td><td>n practice. dta</td><td>Nome</td><td>Label</td></t<>	and the second	contains data from G: V	Dinary choi	ce model/BC	n practice. dta	Nome	Label	
at (Jakaparatoria) iiii: 13,24 iiii: 13,24 iiii: 13,24 iii: 13,24 iii: 13,24 iiii: 13,24 iiii: 13,24 iii: 13,24 iii: 13,24 iii: 13,24 iii: 13,24 iii: 13,24 iii: 13,24 iii: 13,24 iii: 13,24 ii:: 13,24 ii:: 13,24 ii:: 13,24 ii:: 13,24 i:: 13,24 <	timano juc	vars: 15			28 San 2020 15:08	name	Laber	
<pre>maintent etcase display while writable and type frame label writable label writable and type frame label writable label writable and type frame label writable and type label writable and type label writable frame label writable frame label writable frame label writable and type label writable frame label writable and type label writable frame label writable and type label writable and type label writable frame label writable and type label writable frame label writable and type label</pre>	'G/binary choice mode	nize: 175 284				sector	Sector	
microsoft dipple wind microsoft diple micro						mix_activity	Whether pursui	
<pre>unriable name type format intel wurnahe lace1 wetches hype tills in interd_string's provide name activity (yes=1, so=2) mathematic strenger byte tills in the definition of provide name activity (yes=1, so=2) mathematic strenger byte tills in the definition of provide name activity (yes=1, so=2) mathematic strenger byte tills in the definition of provide name activity (yes=1, so=2) mathematic strenger byte tills in the definition of provide name activity (yes=1, so=2) mathematic strenger byte tills in the definition of provide name activity (yes=1, so=2) mathematic strenger byte tills in the definition of specific strenge strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger name activity (yes=1, so=2) mathematic strenger byte tills in the strenger strenger strenger strenger name mathematic strenger byte tills in the strenger streng</pre>		storage	display	value		social_catego.	if item 204 is '1'	
<pre>setser by the visit of the starging maintain and an extra grant and a starging maintain and and a starging maintain and and and and and a starging maintain and and and and and and and and and an</pre>		variable name type	format	label	variable label	bank_account	Whether the en	
<pre>meters by f 10 by meters.</pre>						nature_opera.	. Nature of opera	
<pre>image: image: imag</pre>		sector byte	\$10,09	sectornum		prbim facd	Did the enterpt	
ist_strike bits modul_strike modul_strik modul_strike modul_stri			4		Sector	AND TAKEN IN	Did the enterne	
Monther particular, stated activity/(yes-1, no-2)) Inter 26 is 1' is 0 + 1' is 0', scall grapp of the omar/major particular Inter 26 is 1' is 0 + 1' is 0', scall grapp of the omar/major particular main scount byte 11:0 10:1 10:1 10:1 10:0 10:0 main scount byte 11:0 0:1 10:1 10:0		mix_actvity byte	\$10.0g	mixed_act	ivity	and the states	Crater of the en	
<pre>becligitatepure by: 10.0g billing torop if its 20 is 11 to 5 '', setting turns of the owner/hapte partner hash_accent by: 10.0g billing torop instance, red by: 10.0g billing instance, red by: 10.0g billing torop insta</pre>					Whether pursuing mixed activity?(yes-1,no-2)	groworcscaus	Status of the en	
bal account bit is 10 is 11 is 10 is 11 is 10 is		social_category byte	\$10.0g	social_gr	oup	activity_group	manufacturing	
<pre>bad_execut byt 116.0 bad_execut. bata_execut byt 116.0 bad_execut. bata_execut byt 116.0 bad_execut. bata_execut byt 116.0 bad_execut. bata_execut.byt 116.0 bad_execut.byt 116.0 bad_execut. bata_execut.byt 116.0 bad_execut.byt 116.</pre>					if item 204 is '1' to '5', social group of the owner/major partner	trad_sec	traditional and	
Market the attrighter solution way hold account/post office saving hold account/ market operation Market the attrighter solution way hold account/post office saving hold account/ market operation Market the attrighter solution way hold account/post office saving hold account market operation Market the attrighter solution way hold account/post office saving hold account for market operation Market operation <th c<="" td=""><td></td><td>bank_account byte</td><td>\$14.0g</td><td>bank_acco</td><td>rant</td><td>enterprise_age</td><td>el enterprise age-</td></th>	<td></td> <td>bank_account byte</td> <td>\$14.0g</td> <td>bank_acco</td> <td>rant</td> <td>enterprise_age</td> <td>el enterprise age-</td>		bank_account byte	\$14.0g	bank_acco	rant	enterprise_age	el enterprise age-
water great by to 10.0 minute opertime Monor of speartime Monor of speartime Monor of speartime printing and the first of the strapping near the many problem in its spearatime during last 316 days (yet) Monor of spearatime Monor of the strapping near the many problem in its spearatime during last 316 days (yet) printing and the first of the strapping near the strapping near the strapping near the strapping optime is the strapping near the st					Whether the enterprise maintain any bank account/post office saving bank account	southstates	RECODE of stat	
pala de mor si a de partin de partin de partin de partin de la de de de la de de la de de de la de de de la de de de la de de de de la de		nature_operat~n byte	\$10.0g	nature_op	ert	location_entr.	. RECODE of loc	
pixla, feed byte 10.0 bit the entrypice area of possibles in the spectrum during last 186 days (yet- grand, this byte 10.0 bit the entrypice area of possibles from the spectrum during last 186 days (yet- grand, this byte 10.0 bit the entrypice area of possibles from the spectrum during last 186 days (yet- entrypice area of the start) Interpice area of the start)					Nature of operation			
minimized pres dys 100 g into the softwardprine microsite and principles to the softward principles and the softwardprine microsite and and the softwardprine micros		prblm_facd byte	\$10.0g		Did the enterptise face any problem in its operation during last 365 days? (yes-	Properties		
piezed jitusi grif tilling intrillinginger intrillinginger </td <td></td> <td>assistance_rovd byte</td> <td>\$10.0g</td> <td></td> <td>Did the enterprise receive any assistance from the government during last three</td> <td>8</td> <td></td>		assistance_rovd byte	\$10.0g		Did the enterprise receive any assistance from the government during last three	8		
activity group activity group mainfacturing task and services mainfacturing task and services the task_set filest 11.6 m/s task set sets filest activity group filest 11.6 m/s task sets sets activity group filest 11.6 m/s task sets filest activity group filest 11.6 m/s filest filest activity group filest filest filest fil		growth status byte	\$10.0g		status of the enterprise over the last 3 years	4 Variables		
tark per files 11.0 millioni per per settingting per setting ting per setting per <td></td> <td>accivity_group float</td> <td>\$9.0g</td> <td>activity_</td> <td>group</td> <td>II Name</td> <td></td>		accivity_group float	\$9.0g	activity_	group	II Name		
integring and the state is a state in the state of the state is a state of the stat			114.00		manufacturing trade and services	Label		
Second state System State		trad sec float	\$14.0g	trad_sec	creditional and non-teditional sector	Los .		
ication_stree_synt 0.53 battion MCCOM of Linearius strengths (Decetion of the strengths) to the strength of th		enterprise age ficat	\$9.0g		enterprise age zoie-initial year of operation	()pe		
sep preze front 10.0g extension age entryste type byte 10.0g MCCCE of extension age bered by: Command 4		location entrue bute	\$9.04	location	RECORE of location enterprise (location of the enterprise)	Petinal		
extrapring type byte 18.5p MECODE of extrapring type (Extrapring type during the last 345 days)		are smare float	19.00	1000001000	antermrise and ' entermrise ane	Value label		
Sorted by:		enterprise type hyte	19.00		RECORE of enterprise type (Enterprise type during the last 365 days)	Notes		
Sorted Syn Command Command						4 Data		
Command Command		Sorted by:				Filename	BCM_practic	
- Rom Unavie Commund I Unavier						Label		
Command Command						 Notes 		
Command B Council of C						Variables		
Command F Size 11						Observations		
		Command				Size	171.18K	
Memory Marine						Memory	6414	

We have several times mentioned that if it is in string form mathematical estimation or numerical operation is not possible. We need to destring it and then you can operate it. Similarly it also gives number of observation how many observation it has and how many variables we are dealing with, fine. And along with that it also gives the variable label. The label can also be read from here, alright.

(Refer Slide Time: 06:55)

. su								
Variabl	e 0	os 1	Mean	Std. Dev.	мін	Мах	Summ	arize command
secto	- 64	2 1 51	5096	4998106	1	2	give	s information
mix actvit	y 6,4	2 1.97	0579	.168996	1	2	2	hout total
	y 6,4	2.95	0555	9366901	1	4	a	bouttotal
bank_accoun	t 6,4	2.60	8903	1.468721	1	4	obser	vation for each
nature_ope~	n 6,4	02 1.02	2643	.1913519	1	3	cater	gory (here we
prblm fac	d 6,4	2 1.6	7976	4666047	1	2	hav	e no missing
assistance~	d 6,4	1.9	9122	. 0932968	1	2	lav	e no missing
growth_sta~	s 6,4	1.76	1245	. 6294681	1	3	values) and also gives
activity_g~	p 6,4	2 1.	6565	8169363	1	3	idea at	out the type of
trad_se	c 6,4	1.26	6636	4422344	1	2	iuca at	loat the type of
enterpris~g	e 6,4	10.3	6445	7.077576	3	64	variab	ie: continuous,
southstate	s 6,4	1.65	1109	4766561	1	2	dumm	or categorical
location_e~	e 6,4	1.34	4578	4752671	1	2		
age_squar	e 6,4	157.	5062	265.6416	9	4096		
Valyanii (may choce model/bCM p Gaphics Samos 	ractices Z III To User Esc Tab	8) 0 , 1 0 q w	2 3 e r	⁸ 4 [%] 5 ∩ t y	6 [⊗] 7 [°] 8 u i ¢	9 0 9 10	+ _ Bkop []]] \ Def	Variables
Variation Contraction Contractions	attor Z III III III IIII IIII IIIIIIIIIIIII	a s a z	2 3 e r d x c	⁵ 4 [‰] 5 [°] t y f g h v b	6 ²⁰ 7 8 U i c n j k Alt 5		C 	• Variables • The standard two fully metter inter-
any choice model IAM p Graphics Substitut La Y + X = X Cre 0 Free model Substitut Subs	ador Z III III III IIII IIII IIIIIIIIIIIII	12.55	2 2 3 e r d x c	S 4 % 5 ^ t y f g h v b	6 [®] 7 8 U i 4 n j k Alt 8 enterprise app	9 0 P 1 1 ; ; >	- + = Bisp + = Nort 	Variables Variables Variables tere Name Liables Variables variage for
Varyann G Graphic Statistic the Statistic T 2 X Statistic terr 0 C terr	uvr Esc ab tivit Caps da s Sult tivit Caps da s Sult atio tab tivit Cons tab tivit Cons tab tivit Cons tab tivit Cons tab	1) 0 , 1 1 9 9 W a 5 2 13.09 13.09 13.09	2 3 e r d x c	S 4 % 5 ^ t y f g h v b	6 [©] 7 8 U i G N j k N M 8 esterprise apr is_type (Enterprise	y y o p t tri ; tri ; s typ during t		Variables
Visiyani 🦛	attor Caps	10 0 , 1 1 9 W a 5 2 AR 13.59 19.59	2 3 e r d x c	S 4 % 5 ° t y f g h v b	6 8 7 8 U I Q j k At 0 entreprine ap- tie_type (Interpri	P P Cu s type during t	Basp Basp Suit	Variables Variables Variables Ver Name Labor Name
Varyon Image: Comparison of Comp	and a solution of the solution	19:00 10:00 19:00 19:00 19:00 19:00 19:00 19:00	2 3 e r d x c	S 4 % 5 ° t y f g h v b	6 0 7 8 u i g j k n n An B antapris ge	P P Cu se type during t	Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million Image: 100 million <td>Vaides Vaides Vaides Vaides vaides see very date see, actors there perm, stad. Order autor, actors, there autor, actors, there</td>	Vaides Vaides Vaides Vaides vaides see very date see, actors there perm, stad. Order autor, actors, there
Any shoe moot (M s) Captus States - 4 - 3 - 3 - 3 Y 3 X (Pr - 2 - 4 - 4 - 3 - 3 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1	anter a general a construction of the construc	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 3 e r d X C	A S A S A F A S A F A S A S A S A S A S	6 8 7 8 8 0 1 1	C g P 0 P U U F 2 Cu P Cu P to type during to	An Last 365 days)	Variables Variables Variables Vere causes were Nemer Landers were nature green. Nemer
any choice model IAM get Graphics Stands	and a second sec	1 0 1 0 9 W 0 x 1 0 9 W 0 x 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 3 e r d x c	3 4 % 5 1 1 Y 9 1 V B 1 1 V B 1 1 V B 1 1 V B 1 1 V B 3 1	6 0 7 8 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k	Ctri a	 Bop Bop Def Shit Shit Fin Shit Fin 	Variables
way choice model. (M. j. Graynes Stannes	anna Canada anna Canad	20 0 1 1 0 9 W 0 5 2 48 19:00 10:00 19:00	2 3 e r d x c	5 4 % 5 1 1 7 9 1 1 7 9 1 1 7 9 1 1 7 9 1 1 0 6 1 24 1 1	6 [®] 7 [®] 0 j k n n j Alt n bitstyrtte spe tist tist 2 4	P P O	An Last 355 days)	Variables Variables Ariter understens here Name Under Name Name Under
control of the second sec	And a set of the set o	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 3 e r d x c x c x c x c x x x x x x x x x x x x	3 4 % 5 1 1 7 9 1 1 7 9 1 1 7 9 1 1 0 1 1 10000 0 0 1 10000 1 1 1 101 1 1 1	6 0 7 8 0 1 0 0 1 4 0 1 4 0 1 4 0 1 4 0 1 4 0 1 4 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1	P P C C C C C C C C C C C C C C C C C C	Ar Last 355 days)	Variables
Capital Section 2014	Annu and a second a secon	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 3 e r d x c x x x x x x x x x x x x x x x x x x	3 4 % 5 4 1 2 7 9 14 1 7 9 14 10 10 1 1 10 10 10 10 10 1 1 10<	6 0 7 8 8 10 10 10 10 10 10 10 10 10 10 10 10 10	C 9 P 0 P I C P I C P I P I P I P I P I P I P I P I P I P I	Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage Image: 100 minipage <td< td=""><td>Vaidele Vaidele Vaidele Vaidele Vaidele vereinen vere</td></td<>	Vaidele Vaidele Vaidele Vaidele Vaidele vereinen vere
xy choice in solid (24)	Construction C	005 Maa 197000 197000 197000 197000 19700 19700 19700 19700 197	2 3 e r d x c x x x x x x x x x x x x x	4 % 5 1 1 7 9 1 1 7 9 1 1 7 1 1 1 1 1 1	6 0 7 8 0 j k 0 a a 0 a a 0 a a 0 a b 1 a b 1 a b 1 a b 1 a b 1 a b 1 a b 1 a b 2 a <t< td=""><td>P P CH CH CH CH P CH P CH P CH P CH P C</td><td>A Last 165 days)</td><td>Variables Variables Variables</td></t<>	P P CH CH CH CH P CH P CH P CH P CH P C	A Last 165 days)	Variables
	Construction C	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 3 3 9 r 4 c 7 x c 8 x c	I N S I Y I Y I Y I Y I I Y I I Y I I Y I I I Y I	6 0 7 8 0 j k 0 Alt 0 0 0 Alt 0 0 Charger 14 type (Exterput 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2	P C		Variables
	anna an anna an anna an anna an anna an an	0 1 0 1 0 4 1 0 4 3 5 2 Att 31.00 5 422 1.5150 402 402 2.402 402 402 2.402 402 402 1.402 1.402 402 1.402 1.402 402 1.402 1.402	2 3 e r d x c x x x x x x x x x x x x x	3 4 % 5 6 I Y I Y I I I Y I	6 0 7 8 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 10 j k 2 k j 2 k j 2 j j 2 j j 2 j j 2 j j 2 j j 3 j j 4 j j	P Ctri *	2 J J J J J J J J J J J J J J J J J J J	Variable Variable Ariter understener Name Under Ariter understener Name Under Ariter understener Ander under Ander under Ariter Ander Ander Ariter Ander Ander Ariter Ander Ariter
y there moved K44 g	Anton 2 I I I I I I I I I I I I I I I I I I	O I	2 3 4 7 8 9 9 9 9 9 9 9 9 9 9 9 9 9	S A N S A I Y J J J J V J J J J J J V J	6 0 7 8 0 j k n n j k n n antropyrine system system 10 traps (finiseprine system) 2 2 4 3 2 2 4 3 2 2 4 4 2 2 4 4 2 3 2 2 4 4 2 3 2 2 4 4	P V P V Cut P To type during t	An Last 365 days)	Variables
	Variable	0 0 q W q W q W q W q W q W q W q W q W q W q W q W q W q W q W q W q W q 1.5105 q 2.5535 q 2.5535 q 2.5535 q 2.5535 q 2.5535 q 2.5535 q 1.9224 q 1.9224 q 1.9422 q 1.9425 q 1.5425	2 3 4 7 3 6 3 6 3 7 3 10 3 10	4 % 5 1 t y t y t f g t y t t y t g t t y t g t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t	6 7 8 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 10 j k 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j	P 0 P (C 1) C 1 P (C 1) C 1 P (C	20 Last 165 days)	Vaiable Vaiable Vaiable Vaiable New Example vaiable
	Construction C	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 3 2 9 7 4 9 7 9 7 9 7 9 7 9 7 9 7 9 7 9 7	I N S I I I I I I I I I I I I I I I I I I I I	6 0 7 8 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 0 j k 1 k j 2 j k 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 2 j j 3 j j 4 j j 2 j j 4 j j 3 j j 4 j	P C	An Last 365 days)	 Variable Variable Variable Variable Meri Variable Variable

After the describe we are also interested to understand some minimum background related to the summary. Summary, how it is there? So, simply summary of the entire variable. It gives in different format. It gives; their averages are also present, number of observation present. The minimum value, maximum value is there.

From the minimum and maximum value you can infer that these variables are dummy type because 1 and 2 is there. And some other categories are there. 1 to 4 there are categorical type.

Then from 3 to 64, 9 to 4000, this very clearly seem that there are continuous variables. So, we have clarity on this aspect as well.

(Refer Slide Time: 07:52)

Now we will us	se LP	M to fi	ind th	e pr	obabili	ity of	wome	n being a
mentioned exp	lanat	ory var	ased iables	on	the III	near	compil	hation o
regress enter	prise_	type	enter	prise	e_age	age	_squar	e secto
mix_activity pr	rblm_ rise s	facd d ocial c	issista ateao	nce_ rv_h	_rcvd ank a	trad_ ccour	sec so	uthstates th status
activity aroup	lise s	ociui_c	utego	IY D	unk_u	LCOUR	t yrow	III_SIUIUS
activity_group								
2.0								
swayanı 👲								
swayan 🧕								
swayaan 🧕								_
Swayani G	ess enterpr	ise_type ente se social_cat	rprise_age a egory bank_a	ige_squar	re sector mix. growth_status	_actvity pr activity_(blm_facd assis]
Syvayanı 6	ess enterpri ion_entrpri Source	ise_type ente se social_cato SS	rprise_age a egory bank_a df	nge_squar nccount g MS	re sector mix growth_status Number of ob	_actvity pr activity_c os =	blm_facd assistrop roup 6,492]
Swayan 6	ess enterpr ion_entrpri Source Model 3	ise_type ente se social_cato SS 03.270181	rprise_age a egory bank_a df 13 23.3	nge_squar nccount g MS 3284754	re sector mix growth_status Number of ob F(13, 6478) Prob > F	_actvity pr activity_c cs = = 2 = (blm_facd assis roup 6,492 11.11 .0000	
Swayan 6	ess enterpri ion_entrpri Source Model 3 sidual 7	ise_type ente se social_cat SS 03.270181 15.848119	rprise_age a egory bank_ df 13 23.3 6,478 .110	uge_squar account g MS 3284754 3504495	re sector mix_ prowth_status Number of ot F(13, 6478) Prob > F R-squared Add R-squared	_actvity pn activity_c = 2 = 2 = 2 = 2 = 2 = 2 = 2 = 2 = 2 = 2	blm_facd assistrop roup 6,492 11.11 .0000 .2976 .2962	
Syveyant G	ess enterprison_entrpr	ise_type ente se social_aat SS 03.270181 15.84019 1019.1183	rprise_age a df 13 23.3 6,478 .110 6,491 .157	NS 12504495 1004822	re sector mix prowth_status Number of ot F(13, 6470) Prob > F R-squared Adj R-square Root MSE	_actvity pa activity_c = 2 = (= (ed = (blm_facd assist roup 6,492 11.11 0000 2976 2962 2362 33242	
Swayan C	ess enterpri Source Source Source Source Total Status	ise_type ente social_cat SS 03.270101 15.84019 1019.1103 e Coef	rprise_age a egory bank_s df 13 23.3 6,478 .110 6,478 .110 6,491 .157 . Std. Err	uge_squar lccount g MS 10504495 7004822 L L L L L L L L L L L L L L L L L L	re sector mix proth_status Number of ol F(13, 6478) Prob > F R-squared Aost MSE ; P> t	_activity pr activity_c = 2 = 2 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0	blm_facd assis roup 6,492 11.11 .0000 2.2976 2.2952 33242 . Interval)	
Swayan C	ess enterpri ion_entrpri Source Model 3 sidual 7 Total Prprise_typ terprise_ag	ise_type ente se social_cat SS 03.270101 15.84013 1019.1103 e Coef e .000392	rprise_age a egory bank_s df 13 23.3 6,476 .110 6,491 .151 . Std. Err 1 .0016422	Ige_squar Inccount g MS 284754 9504495 7004822 0 0 0 0 0 0 2 0 2 0 2 0 2 0 2 0 2 0	re sector mix proth_status Number of of F(13, 6470) Prob > F R-squared Adj R-squared Adj R-squared Adj R-squared Sector MSE P> t 4 0.811	_actvity pr activity_c = 2 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0	blm_facd assister roup 6,492 11.11 0000 2.2976 2.9752 33242 . Intervall . OD36128	
Swayan C	ess enterprison_entrpris Source Model 3 Sidual 7 Total Prrise_typ terprise_aguar secto	ise_type ente se sozial_est SS 03.270181 15.848119 1019.1183 e Coef e .000392 e000372 03398	rprise_age a df 13 23.3 6.478 .110 6.491 .157 . Std. Erri 1 .0016423 8 .0000433	HS HS HS HS HS HS HS HS HS HS	re sector mix provth_status Number of ol Fil3, (470) Prob > F R-square Adj R-square Adj R-square Adj R-square Status H 0.811 22 0.187 9 0.000	_actvity pr activity_c = 2 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0	blm_facd assis roup 6,492 11.11 0000 1.2976 1.2962 33742 	
Swegen C	ess enterpri Source Model 3 sidual 7 Total erprise_typ terprise_ag age_squar secto mix_actvit	ise_type ente es social_cat SS 03.270101 15.84811 1019 1103 e Coef e .000392 e000392 r .003950 y .01064	rprise_age a egory bank_s df 13 23.3 6,478 .11(6,491 .157 . Std. Err 1 .0016422 8 .00000433 8 .00245654	Age_squar MS 2284754 3284754 004622 ↓ 	re sector mix_ provth_status Number of ol F(13, 6470) Prob > F R-square Adg R-square Adg R-square Adg R-square Sector SE D> t R4 0.811 12 0.187 19 0.000	_actvity pr activity_c = 2 = 0 = 0 = 0 = 0 = 0 = 0 [55% Conf 0028286 0028286 0001436 0002436 000436	blm_facd assis roup 6,492 11.11 .0000 1.2976 2.2962 3.3242 	
Swegen C	ess enterpri Source Model 3 sidual 7 Total erprise_typ terprise_ag age_squar secto mix_actvit prblm_fac	ise_type ente es social_cat SS 03.270181 15.848119 1019.1183 e Ccef e .000352 e .000057 c .00058 e .000057 c .00058 d .027862 d .027862	rprise_age a df 13 23.3 6.478 .11 6.491 .157 . Std. Err 1 .0016422 8 .008503 8 .008503 5 .002455 5 .0022705 5 .04475	Age_square MS 2284754 3284754 004622 0 3 - 1.3 3 - 1.3 3 - 1.3 1 - 4.4 1 - 3.0 4 - 2 1 - 4.4 2 - 3.5 1 - 4.4 2 - 3.5 2 - 4.5 2	re sector mix provth_status Number of ot F(13, 6478) Prob > F R-square Root MSE ; P> t 24 0.811 12 0.187 19 0.000 13 0.664 11 0.003	actvity pr activity_c = 2 = 0 = 0 ed = 0 [55% Conf 0028286 0001436 0034214 .0036888 014176	blm_facd assi roup 6,492 11.11 0000 1.2976 2.2976 2.2976 2.33242 0.0036128 0.000281 017289 0.002081 0.000281 0.00000000000000000000000000000000000	
Sworth E	sss enterpri ion_entrpri Source Model 3 sidual 7 Total erprise_typ terprise_typ terprise_typ secto mix_actvit prbim_fac stance_rcv trad_se	ise_type ente se social_cat SS 03.270181 15.848119 019.1183 e Ccef e000392 e000392 e000392 e003986 d027862 d010164 c38778	rprise_age a egory bank_c df 13 23.1 6.470 .110 6.491 .157 . Std. Err 1.0016423 8.0005037 8.0095037 5.009507 5.0095707 5.0095707	Hge_equate HS 2284754 3284754754 3284754 3284754 3284754 3284754 3284754 3	re sector mix provth_status Number of ol F(13, 6478) Prob > F R-squared Adv Root MSE 2 2 2 2 2 2 2 3 3 3 3 4 0 8 1 1 0 0 3 3 0 6 6 1 0 0 0 3 3 0 6 8 1 0 0 0 3 1 0 0 0 3 1 0 0 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	_actvity pa activity_c = 2 = 2 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0	blm_facd assi roup 6,492 11.11 .0000 .2376 .2352 .33242 .000281 .000281 .000281 .0002841 .0002841 .000841 .1861522 .15518152	,
Swayan C	Ass enterpri Source Model 3 sidual 7 Total Frprise_typ terprise_squar sector mix_actvit prblm_fac istance_rev trad_se	ise_type ente se social_cat SS 03.270181 15.848119 1019.1183 e Coef e .000352 e .000357 r -003358 y .010684 d .027862 d .101644 s -0068497	rprise_age a egory bank_ df 13 23.3 6,478 .110 6,491 .157 1 .0016423 8 .002503 8 .0024565 5 .0092707 5 .0092707 7 .008856	Age_Pquares MS MS A224754 A224	re sector mix provth_status Number of ol F(13, 6470) Prob > F Prob > F R-squared Adj R-squared Adj R-squared Adj R-squared Do 0.11 20 0.187 10 0.003 10	_actuity pn activity_cr = 2 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0	blm_facd assii roup 6,492 11.11 0.0000 2.2762 33242 . Interval] 0000 0.000201 0172809 0.08841 0.0460361 1891522 1.591522 1.591522	
Swegan C	ss enterpri Source Model 3 Total Total erprise_up prblm_destance_rov tradies outhstate	Ise_type ente se social_cat SS 03.270181 15.848119 1019 1183 e Ccef e .000392 e -000397 y .000392 d .001684 d .027662 d .010684 d .02762 d .010684 d .02762 d .010164 d .02762 d .01164 d .02762 d .01164 d .02762 d .00564 d .005644 d .005644 d .005644 d .005644 d .005644 d .005644 d .0056444 d .0056444 d .005644444444444444444444444444444444444	rprise_age a egory bank df 13 23.3 6,478 .110 6,491 .157 1 .0016423 8 .008503 8 .024565 5 .009377 9 .008957 7 .008856 5 .0093024	HS HS 284754 5004495 0004822 ↓ 000000 ↓ 00000 ↓ 00000 ↓ 00000 ↓ 00000 ↓ 000000 ↓ 000000 ↓ 000000 ↓ 0000000 ↓ 00000000	re sector mix provth_status Number of ol 7(13, 6470) Prob > F R-squared Adg R-square Adg R-square Adg R-square Adg R-square Adg R-square Adg R-square 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,	_activity pr activitys = 2 2 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0	blm_facd assi roup 6,492 11.11 0000 2.276 2.33242 . Interval] . 0036128 0.0036128 0.0036128 0.003641 1.0036426 	
Swegen C	sss enterpri Source	ise_type ente es social_cat SS 03.270181 15.848113 1019 1183 e Coef e .000352 e -00352 e -00352 y .010684 d .01144 c .138778 s -008497 e .356378 y - 023045 c .023045	rprise_age a df 13 23.3 6,478 .110 6,491 .157 . Std. Erri 1 .0016423 8 .0005037 9 .024555 5 .009977 7 .008856 5 .009977 9 .009897 7 .008856 5 .009977	HS	re sector mix_ provth_status provth_status Prob > F R-squared Adj R-square Adj R-square Adj R-square Sector Status Prob > F R-squared Adj R-square Adj R-square Adj R-square Distance Prob > F R-squared Adj R-square Adj R-square Adj R-square Adj R-square Adj R-square Adj R-square Distance Prob > F R-squared Adj R-square Adj R-square Adj R-square Distance Prob > F R-squared Adj R-square Adj R-square Distance Prob > F R-squared Adj R-square Distance Prob > F R-squared Adj R-square Adj R-square Distance Dis	_activity pr activity_s = 2 2 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0	blm_facd assist roup 6,492 11.11 0000 1.2976 1.2962 3.3742 0.0036128 0.0036128 0.003614 0.012869 0.05841 0.486152 1.561811 0.008646 - 3.371618 - 0.015418	
Sworth C	ess enterpri Nodel 3 Nodel 3 rprise_type terprise_type prblm_factors problm_factors problm_factors protection factors production factors production factors production factors production factors production factors production factors production factors production factors	ise_type ente es social_cat SS 03.270181 15.848119 015.9183 e Ccef e .000392 e .000057 c .000392 e .000057 c .000392 e .000392	rprise_age a egory bank df 13 23.1 6.470 .11(6.491 .157 . Std. Err 1.0016423 8.0005037 6.004505 5.0098207 5.0098026 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 9.008856 5.009826 5.00986 5.009826 5.009826 5.009826 5.009826 5.00986 5.009826 5.009826 5.00986 5.00986 5.00986 5.00986 5.00986 5.009856556 5.009856556 5.0098565556 5.0098565556565656555556555555555555555555	Age_Pquak MS 2284754 2284754 2284754 20004822 2 2 2 3 -1.5 -5.0.4 -5.0.4 -5.5.5	re sector mix provth_status Ful3. (478) Prob > F R-squared Action Prob > F R-squared Action Prob > F R-squared Action Prob > F R-squared Root MSE Research Root MSE Root MSE Research Root MSE Root MSE Research Root MSE Research Root MSE Research Root MSE Research Root MSE Root MSE	_activity pr activity_composition = : : = : = : = : = : = : = : = : = : =	blm_facd assi roup 6,492 11.11 1.0000 2.2376 2.2372 1.000281 0.000281 0.000281 0.000281 0.000281 0.0008446 0.3371618 0.034561 0.034561 0.034561 0.034561 0.034561 0.034561 0.034561	;

So, let us make a move to the next understanding. That is I think perfectly fine. Let us move to how we can apply LPM specially to understand the probability of women, probability of women being a necessity entrepreneur based on those variables we discussed or the linear combinations of the explanatory variables. So, simply linear model we all know that we need to simply regress

dependent variable and the independent variable. Irrespective of dependent variable to be having only discrete choices is not the matter of concern so far as LPM technique is concerned.

So, now we are going to apply. We are going to copy it and then we will go to run it. So, simply all the variables are there. We are simply going to run it with the data. So, once I enter, will get the result. Look at how clearly we establish the result with the reg command, regression command. What is interesting to note in this, it is similar to the ordinary least square result.

It gives R square, adjusted R square which we used to discuss earlier. It also gives F statistics. And F statistics is significant here. So, that means there is no problem with the overall model of the LPM. But there are some integreties, some internal problem that we need to explore. That how it creates problems we are going to define in our other slides. So, this is what the result we have shown.

(Refer Slide Time: 09:36)



	age_square 6,43	157.5062 265.64	6 9	4096	 Variables 	Ŧ
Filter commands here	enterpris-pe 6,49	8049908 .396238	93 0	1	+ Effer varial	
Command rc	. regress enterprise type	enterprise age age squ	are sector mix actv	ity prblm facd assistance rowd trad sec southstates loca	tion . Name	Label
use 'G\binary choice mode	> ntrprise social_categos	y bank account growth :	tatus activity_grou	P	sector	Sector
des			Marker 14 161		mix_actvity	Whether pursuing
sum	Source 55	dr MS	Number of obs	= 6,492 = 211 11	social_categ	p., if item 204 is 'T t
regress enterprise_type ent	Model 303.27018	13 23.3284754	Prob > F	- 0.0000	bank_accour	nt. Whether the ente
predict (pmpred, xb	Residual 715.84811	9 6,478 .11050449	R-squared	- 0.2976	nature_oper	a Nature of operat
			Adj R-squared	= 0.2962	prbim_facd	Did the enterptis
	Total 1019.11	6,491 .15700482	ROOT MSE	33242	assistance_n	c Did the enterpris
					growth_stati	us Status of the ent
	enterprise type	Coef. Std. Err.	t P> t (95	A Conf. Interval)	activity_grou	up manufacturing to
					trad_sec	traditional and r
	enterprise_age .	003921 .0016429 0	24 0.81100	28286 .0036128	enterprise_a	ige-enterprise-age+
	age_square .	0000578 .0000438 -1		01436 .0000281	southstates	RECODE of state
	nix actvity	106848 .0245654 0	43 0 664 - 03	74714 058841	location_ent	tr RECODE of local
	prbln_facd	278625 .0092707	0.01 0.003 .00	96888 .0460361	Properties	
	assistance_rcvd .1	011645 .044374 2	28 0.023 .01	41768 .1881522		
	trad_sec .	1387789 .0098974 14	1.02 0.000 .11	93767 .1581811	4 Variables	
	location entroprise	0084977 .0088568 Lgc 8563785 .0098028 -30		55557	Name	
	social category - 4	230459 .0045359 -!	.08 0.00003	19378014154	Label	
	sector correlation in		8.70 0.000 .01	94183 .0307064	Type	
	bank_account .4	250623 .0028/91 8				
	bank_account .4 growth_status .4	411171 .0068707 5	.98 0.000 .02	76482 .0545861	Format	
	bank_account .4 growth_status .4 activity_group4	0250623 .0028/91 8 0411171 .0068707 5 0838133 .0057389 -14	0.98 0.000 .02 0.60 0.00009	76482 .0545861 506350725632 96533 1.18733	E Value label	
	bank_account .4 growth_status .4 activity_group4 	00250623 .0028791 8 0411171 .0068707 5 0838133 .0057389 -14 978637 .1066657 5	5.98 0.000 .02 1.60 0.00009 9.17 0.000 .76	76482 .0545861 506350725632 95371 1.187737	E Value label Notes	
	bank_account	2250623 .0028791 8 2411171 .0068707 9 2838133 .0057389 -14 .978637 .1066657 9	5.98 0.000 .02 5.60 0.00009 9.17 0.000 .76	76482 .0545861 506350725632 95371 1.187737	Value label Notes	
	bank account .0 growth_status .0 activity_group0 . predict lpmpred, xb	2200633 .0028791 8 M11171 .0068707 9 838133 .0057389 -14 978637 .1066657 9	5.98 0.000 .02 1.60 0.00009 9.17 0.000 .76	7642 .6545961 5655 - 675582 95371 1.187737	Value label Notes	BCM_practice
	ban_account	22502253.0028731 M41171.0068707 M238133.0057389 -14 978637.1066657	8.98 0.000 .02 8.60 0.00009 9.17 0.000 .76	7442 0545961 2655 - 072562 85371 1.187737	Value label Notes Value label Notes Filename Label	BCM_practice
	bank_sccount	220022930 00229331 0 2411271 0068707 9 28381333 0057389 -14 978637 1066657 9	6.98 0.000 .02 6.60 0.00009 9.17 0.000 .76	7442 0343441 9655 - 07242 98371 1.387737	Value label Notes Data Filename Label Notes	BCM_practice
	bank_sccount growth_status 	2200223 .0028731 1 98381371 .0068707 1 9838333 .0057389 -14 978637 .1066657 1	5.98 0.000 .02 .60 0.00009 9.17 0.000 .76	7442 545841 5655 - 675522 59371 1.147737	Value label Notes Value label Notes Label Notes Conservations	BCM_practice 16 5 6.492
	bank_account	220623 0022971 1 141111 0.064970 1 9838133 0.057389 -14 978637 1066657 1	998 0.000 .02 8.60 0.00009 9.17 0.000 .76	9442 - 34444 1953 - 601942 9971 - 1,40797	Value label Science label	8CM_practice 16 5 6.492 202.88K
	bank account d growth status d construction of the status d construction o	230633 0022971 1 141111 0.0064977 1 9398133 0.0557899 -1 978637 1.066657 1	5.98 0.000 .02 .80 0.00099 9.17 0.000 .76	7442 343443 9555 - 421942 9571 1.187117	Value label Value lab	BCM.practice 16 5 6.492 202.88K 64M

How we can interpret the result? This is the result we have derived just now. How we are going to interpret it? Basically as we know that our dependent variable is not just Y, it is P. It is probability of enterprise type 1 that is necessity entrepreneurs is equal to the constant term. This is the constant term, the intercept term is here. It is given 0.98637, is constant.

So, that is plus each coefficient time says explanatory variables. So, each coefficient like first variable, enterprise age, so age coefficient is this times this enterprise age plus, then second variable its coefficient and the second variable, third variable and its coefficient. So, if you just add everything, so that is all about the model. That is the expected value of your model.

So, once we add all those things and we can estimate the predicted values or the probability value of enterprise having 1 or to be necessity entrepreneur. In that case since there are so many variables and each one unit change and its impact on the dependent variable is difficult without software. So, manually we are not calculating. Basically we are predicting probability of success, predicting the model using all those coefficients. So, in that case we will use the predictability. And we define a variable with this option that will give a linear predictions.

So, here is our prediction option. So, we will simply run this command in STATA. So, just a minute. So, it is here. Then you can go to the previous slide. Probably box is selected. Box can be avoided. l p m p r e d, we can do that alright, p r e d x b. So, this is what the predicted variable we have defined. And now it is visible. The predicted variable is visible.

(Refer Slide Time: 12:41)



Ipmpred[1] 25443387	
lpmpred	 Variables
. 2544339	E Filter variables here
.2424133	R Nore Intel
.2742338	Element linear celetion
.2742330	E springeta College prediction
.2838348	
.2045169	C recipion of the Million Control of the Mill
2848471	back account, Whather the ends
1 .28025	abue opera. Nature of operati
.2903354	C while fact. Did the extension
.3920074	assistance in Did the enterprise
.2933027	Consult status Status of the end
.2937771	C artistic areas manufacturing to
.283000	trational and n
.2941359	
2941359	southstates BECODE of state
5 .2943792	Variables Snapshots
.2945192	Properties
2970565	4 Madebler
.29771	- variables
	name ipmpreu
	Tow first
.3132114	
3132114	Value label
4 . 31322114	
	4 Data
	Filename BCM practice.
	Label
	Notes
r - All Maria	Vanables 16
·	Observations 6.492
	5.ze 202.88K
	 Memory 64M

We can see and we can browse it. We can easily see how it looks like. But we can also sort it up and so simply sort and the variable defined. Now, we have sorted and now we can browse. So, through the browse we can see how it looks like? What is the probability? That is 25 percent, 25.44 percent the first enterprise has the probability to be necessity entrepreneur, of that particular enterprise could be a necessity entrepreneur. Similarly we can have other options also. But most importantly we wanted to check some important aspects of LPM, some problems of LPM. So, we need to define whether that is within the limit or not.

(Refer Slide Time: 13:32)

Stata/SE 15.1 - G\binary choice model\BC	M. practice da		_ 0 _ ×_
File Edit Data Graphics Statisti	cs. User Window Help		
📓 🖬 😂 🗄 🗟 - 🖬 - 🚿 - 📓 -	B = 0 · 0		
Review T # ×	> ntrprise social_category bank_account growth_status activity_group	 Variables 	₹₽×
★ Filter commands here	Source 55 df MS Number of obs = 6.452	+ Eiter variables	s here
# Commandrc	F(13, 6478) = 211.11	Name	Label
1 use 'G\binary choice mode	Model 303.270181 13 23.3284754 Prob > F = 0.0000	bank_account	Whether the enter
2 des	RESIDUAL 715.848119 6,478 .110504495 K-Squared = 0.2976	nature_opera	Nature of operation
3 sum	Total 1019.1183 6,491 .157004822 Root MSE = .33242	prbim_facd	Did the enterptise
4 regress enterprise_type ent		assistance_rc	Did the enterprise
5 predict (pmpred, xb)		growth_status	Status of the enter
6 sort lpmpred	enterprise_type Coef. Std. Err. t P>(t) [95% Conf. Interval]	activity_group	manufacturing tra
7 br (pmpred	enternrise and 0001921 0016429 0.24 0.811 - 0028286 0016128	trad_sec	traditional and no.
	age square - 0000578 0000438 -1.32 0.187 - 0001436 0000281	enterprise_age	enterprise age + 2
	sector0335588 .0085037 -3.99 0.00005062880172889	southstates	RECODE of staten
	mix_actvity .0106848 .0245654 0.43 0.6640374714 .058841	location_entr	RECODE of locati
	prbin facd .0278625 .0092707 3.01 0.003 .0096888 .0460561	age_square	enterprise age * e
	43513(date_icvd)01200500030 2.20 0.0230143061003222 Trad see 1387789 0000074 14.02 0.000 1103767 1563831	enterprise_ty.	RECODE of enterp
	southstates0084977 .0088568 -0.96 0.3370258599 .0088646	Ipmpred	Linear prediction +
	location_entrprise3563785 .0098028 -36.35 0.00037559513371618	Properties	ų×
	social_category0230459 .0045359 -5.08 0.0000319378014154	₽ • •	
	bank account .0250623 .0028791 8.70 0.000 .0194183 .0307064	 Variables 	
	growingstatus	Name	Ipropred
	Cons .978637 .1066657 9.17 0.000 .7695371 1.187737	Label	Linear prediction
		Type	float
		Format	%9.0g
	. predict lpmpred, xb	Walue label	
	sort lownrad	Notes	
		4 Data	
	. br 1pmpred	Filename	BCM_practice.dla
		Label	
		• Notes	
			16
	Command	Cuservations	0.922 302.88¥
	br 1pmpred if 1pmpred > 1	Memory	
G\Stata15			CAP NUM OVR
😗 🖉 🚺 🧇 🖉	🖳 🧿 🔚 🙀 🔟	- 🔋 📀 🖣 🕻	D 10/6/2020

(pmpre)(332) 1000(333)			
stables		Variables	
		filter variable	
1.000181		🗹 Name	Label
# 1.0001#J		🖉 Ipmpred	Linear prediction
44 1.000319		Sector	Sector
87 1.000319		mix_activity	Whether pursuing
40 1.000400		social_catego	if item 204 is '1'4
P 1.000475		☐ bank_account	Whether the ente
0 1.000589		inature_opera	Nature of operati
1.000734		prblm_facd	Did the enterptis
1.00183		assistance_rc	Did the enterprise
1.000882		growth_status	Status of the ente
4 1.000282		activity_group	manufacturing tr
1. 1. 000882		🗌 trad_sec	traditional and re
4 1.000852		🗌 enterprise_age	enterprise age » 2
1 000852		southstates	RECODE of states
4 1.009402		Variables Snapsh	1015
1.001001		Properties	
A 1 (A1228		4 Variables	
1 1 001237		Name	
3 1.001237	4		
3 1.001241		Type	float
4 1.001241		Format	%9.0g
5 1.001241		Value label	
4 1 00123		Notes	
7 1.001349		4 Data	
1.001478		: Filename	8CM_practice
1.001412		Label	
0 1.001412		Notes	
21 1 001412		Variables	16
1 001612		Observations	6,492
1 1.001435		Size	202.88K
		 Memory 	64M
	Vars: 1 of 16. Order: Dataset	Obs. 1,101 of 6,492 Filter: On Mod	le:Browse CAP

So, we can browse by greater than how many are exceeding 1. So, browse that variable. Then with if, that is exceeding greater than 1, alright. We need to check whether LPM has exceeded the probability limit of 0 and 1. Look at how many of them have exceeded the probability value, the predicted value. The predicted probabilities value is exceeded to 1. This is very clearly violating our very assumption of the LPM model.

(Refer Slide Time: 14:10)



□ Now we can browse the data and look for individual predicted probabilities of each enterprise.

browse

We can also sort the data by lpmpred:

sort lpmpred

Then browse to see which enterprise has lowest probability of being a necessity entrepreneur.

browse

							being necessity
prblm_face	assistance-d grow	th_sta-s activity_g-p	trad_sec en	terpris~ge southstat	es location_e~e	age_square ent	erpris-pe lpmpred
1 1	1	1 services	traditional	7	2 outsideHH	⊳ 49	0 .2544339
2	2	2 services	traditional	39	2 outsideHH	1521	0 .2624133
3	2	2 services	traditional	36	2 outsideHH	1296	0 .2742338
4	2	2 services	traditional	36	2 OUTSIGERM	1296	0 .2742338
Lt is al	so assu	² services 1 services med that	traditional traditional	cted pr	2 outsideHH 2 outsideHH	289 ties sl	0 .2742338 0 .2838348 hould lie



I wanted to move on and explain once again with these. We have already operated this browse, then sort, then also browse as per the interpretations, like we interpreted that 25.44 percent chances of being enterprise, that the first enterprise here in this list I have 25.4 percent chances of being necessity entrepreneur.

So, after browsing how many are exceeding and who are exceeding the 0 value or 1 value we have observed that there are many cases where the predicted value has exceeded 1. And we run this command browse. This variable we defined already. Then those who are greater, that has already been filtered. We can conclude that LPM is not fulfilling the condition of the limit of the expected values the predicted values.

(Refer Slide Time: 15:36)





We can look at a plot also used to understand their residual values against probabilities of all the observations in our data set. So, rvf plot is going to give us that result. The residual is going to be defined. It has come.

Look at the same thing can also be interpreted here. On the left hand, on the vertical axis residual these are fitted values. There are two points you can easily see. Even on the first line and nearby the another one there are some outlier, very clearly identified it is exceeding 1 value. One is here, that is exceeding that 1 value. So, this is what I wanted to discuss. rvf plot discusses the residuals and this is how it looks like.

So, we see here that for each Y hat the residuals are free to take on only two values, it can be concluded that the residuals are not normally distributed. So, residuals are not at all normally distributed and that too there are some outliers as well. So, we notice that from the rvf plot, that there are values of Y hat that fall even outside the range that is 0 and 1. So, this is because LPM places no constraint on the range of Y hat letting them range between negative infinity and infinity. So, we have probabilities greater than 1. This would indicate that there is a problem with the credibility of our model. Coming to another sophisticated model called Logit. Based on the earlier understanding that LPM is the simplest format where we simply apply the linear probabilistic structure but it attaches with number of shortcomings residuals as well as its predicted values are explaining that it is not good model.

We are discussing another binary choice model, binary response model called Logit. Logit model is applied where the distribution follows a logistic distribution or logistic structure. So, we are going to discuss that in a short while.

(Refer Slide Time: 17:56)

	INTRODUCTION
) LPM assumes that $P_i = E(Y=1 X)$ increases linearly with X that means the marginal or incremental effect of X remains constant throughout. But in reality one would expect that P_i is nonlinearly related to X_i .
	So, we need a probability model that characterize: □ As X _i increases, P _i = E(Y=1 X) but never fall beyond the range of 0 and 1.
	\Box The relationship between P_i and X_i is nonlinear.
	Device the sector of the secto
) - 5	Wayann 🔮
[Both theoretical and empirical consideration suggests that when the dependent variable is a binary variable, the shape of the response function will frequently curvilinear.
[Both theoretical and empirical consideration suggests that when the dependent variable is a binary variable, the shape of the response function will frequently curvilinear. The logit and probit function follows the sigmoid, or S-shaped curve. Showing nonlinear relationship between P_i and X_i that means, one approaches zero at slower and slower rates as X_i gets small and approaches one at slower and slower rates as X_i gets very large.
[Both theoretical and empirical consideration suggests that when the dependent variable is a binary variable, the shape of the response function will frequently curvilinear. The logit and probit function follows the sigmoid, or S-shaped curve. Showing nonlinear relationship between P_i and X_i that means, one approaches zero at slower and slower rates as X_i gets small and approaches one at slower and slower rates as X_i gets very large. The logit model follows cumulative logistic distribution function.

LPM in fact assumes that the probabilities value increases linearly with X. That means marginal or incremental effect of X remains constant throughout. But we have seen that the marginal impact should remain constant throughout but there are some problems in the model.

But in reality one would expect that Pi is nonlinearly related to Xi. So, it is not linear with Xi. As we need a probability model that characterizes the Xi's increases Pi that is with success as 1 but never fall beyond a range of 0 and 1. The relation between Pi and Xi is nonlinear. That is also required. Now onwards we will be emphasizing this aspect. Models with such characteristics are Logit and Probit models. Both theoretical and empirical considerations suggest that when the dependent variable is a binary variable the shape of the response function will frequently curvilinear. So, the Logit and Probit function follows the sigmoid curve or the S-shaped curve.

The sigmoid curve follows showing non linear relationship between Pi and Xi that is the probability of the Y with that of the explanatory variable should be nonlinear. That means one approaches 0 at slower and slower rate, approaches 1 at slower and gets small and approaches 1 at slower and slower rate at Xi gets very large. So, basically in the very extreme points the Pi value is approaching to a slower rate so far as sigmoid curve is concerned.

(Refer Slide Time: 20:06)



The Logit model follows a cumulative logistic distribution function. So, this looks like this. With the higher values, it approaches to a slower rate. Even at the lower value also approaches to a slower rate. So, that is all about the cumulative distribution function of a sigmoid curve or a Logit curve.

(Refer Slide Time: 20:26)



So, logistic regression follows number of important assumptions. Without that logistic regression is not wise to define. So, what are the important assumptions? The logistic regression does not require a linear relationship between the dependent and independent variables. That is one of the most important aspect.

So far in the earlier models we only discussed linear relationship. So, Logit and even Probit not necessarily require linear relationship. That can be converted to a linear format for estimation but original they do not require linear relationship. The dependant variable in logistic regression is

not measured on an interval or ratio scale that is on binary or in categorical format. The residuals do not need to be normally distributed.

We already define that residuals should not be forcefully kept or forcefully defined to be normally distributed because they have only binary choices at this moment. Not necessary even as per the assumption of the Logit. Homoscedasticity is also not required. So far we say there are problems of homoscedasticity. Even if it is not there, still we can define that Logit model is going to be one of the fittest model.

Logistic regression requires the observation to be independent of each other and it should come from repeated measurements or maths data. This requires little or no multicollinearity among the independent variables. So, independent variable should also not be correlated as per the assumption we normally take in all the models so far. This assumes that independent variables are linearly related to the log of odds.

So, once we define the odds ratio then log of those odds are linearly related to the independent variable. We are going to clarify this with the help of equation, log of odds. The Logit, logistic regression typically requires large sample size. Sample size if it is there better then this is one of the better models.

THE LOGIT MODEL	
 Also called logistic regression model. The Logit model is based on the following distribution function of the logistic distribution: 	cumulative
$P_i = E(Y_i = 1 X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}}$ Where, i is the ith observation. For simplification, let us denote, $Z_i = \beta_0 + \beta_1 X_i$ Equation (1) can be written as,	(1)
💿 _ syayan) 🧕	22

(Refer Slide Time: 22:45)

Coming to the application through the theory, coming to the clarification of the logistic regression model through the mathematical theory is as follows. Like this Logit model is also called a logistic regression model. Logit model is based on the following cumulative distribution function which we have already defined, that basically probability of success that is 1, follow a cumulative function or a sigmoid function which is generally expressed with the help of exponential diagram e to the power Z. e to the power Z is the distribution. Z here is the expected value, expected value of the model. 1 upon 1 plus e to the power minus Z. So, Z here is our standard expected value.

(Refer Slide Time: 23:56)



So this equation can be written like this. Since we said 1 plus e to the power minus Z, e to the power minus Zi, so this will certainly be equal then 1 upon e to the power Z. Then this will be equal to e to the power Z, 1 plus e to the power Z, then e to the power Z divided by 1 plus e to the power Z, alright.

So, this is what e to the power Zi then 1 plus e to the power Zi which is nothing but called the distribution function of the Logistic model. So, this represents this F function or the cumulative logistic function. The probability at the any point can be defined through this particular equation so far as the logistic distribution is concerned.

It is visible from the function that Zi ranges from minus infinity to infinity. Pi ranges from 0 to 1, alright. So, from the diagram if you just look at once again Zi varies from minus infinity to infinity but the probability limit varies from 0 to 1. So, the probability we have defined here, the Z values or the X values we defined. Z value is composed of X values is defined here, alright. So, I will just a minute, I will come back, estimation, we will come back, alright. So, let me explain another slide then we will take it forwarded to next class.

Coming to the discussion of the ranges of the Z function, the Z, the expected value of the probabilistic function and Pi that is 0 and 1. So, another important aspect is that Pi is nonlinearly related to Zi. So, Pi is, here Pi is nonlinearly related to Zi. Even in the diagram also it is nonlinear. So, Pi is nonlinear not only in X but also in betas. The beta values because beta values are where? Beta values are defined here.

So, the Pi is not just nonlinear to Xi, because it is e to the power something, so it is also nonlinear to beta values as well, alright. So, this means that OLS can never be used when it is completely nonlinear. So, this equation can be linearized to apply the OLS technique, can be linearized for further interpretation. (Refer Slide Time: 27:03)



So, accordingly we can define an odds ratio. So, this is the last slide of this lecture and then we will carry forward. So, that the probability of success is Pi and 1 minus Pi is the probability of failure, where probability of success we said it is e to the power Z divided by 1 plus e. So, it will be or 1 upon 1 plus e to power Zi then 1 minus this one will be equal to this.

Then the success divided by failure, success divided by failure, this is nothing but called odds ratio is Pi upon 1 minus Pi which is equal to, if you divide the success divided by the failure this is what we already defined, the success and failure, it boils down to e to the power Zi. So, Pi

upon 1 minus Pi is simply called odds ratio. This is what we said odds ratio. The log of odds ratio, if I take log of it, natural log of it, this is basically Zi, Zi. Log of Pi upon 1 minus Pi is nothing but only Zi, isn't it? So, Zi is basically beta not plus beta 1 Xi, beta i Xi, alright.

So, look at this is linearly related to the log of the odd ratio, which is the point I think we wanted to clarify somewhere in linear related to the log of odds. So, this is what we just clarified. And so odd ratio in favor of success that is basically the success to the failure ratio and we have clarified. So, if we take the natural log of that odds ratio that is nothing but beta not plus beta 1. If you have only one coefficient to be defined then this is only beta 1 Xi. Then the Li is the notation for log of odds ratio. It is also called logit. This is called logit and that can be estimated.

Hence the above model is named as Logit model. Li is not only linear in X but also linear in terms of parameters. So, this is also linear with X. Rest of the detail features of the Logit model and its application with STATA and why it gives way forward for Probit model we will continue from the next class. With this let me stop here. Thank you so much.