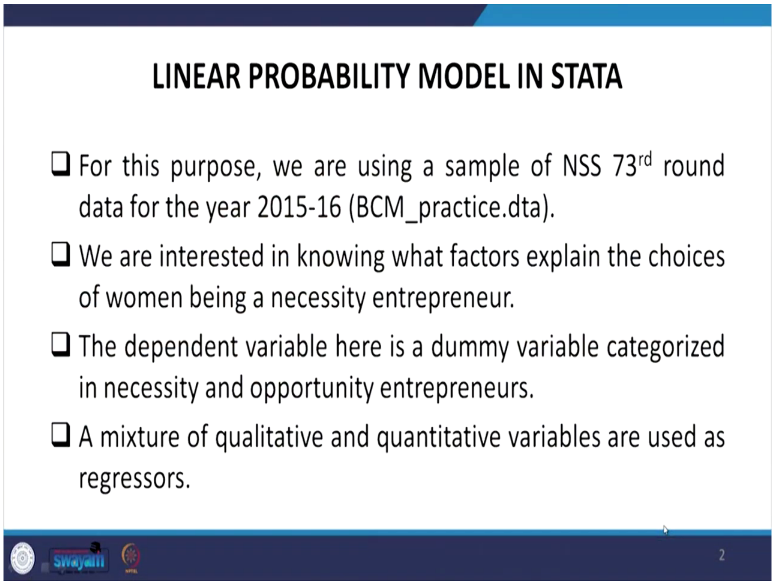
Handling Large Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 33
Binary Response Models - II

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)



LINEAR PROBABILITY MODEL IN STATA

❑ For this purpose, we are using a sample of NSS 73rd round data for the year 2015-16 (BCM_practice.dta).

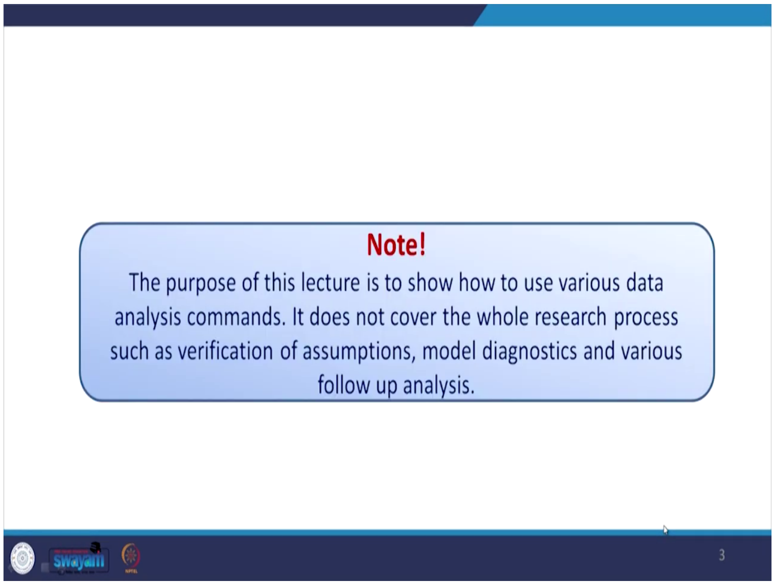
❑ We are interested in knowing what factors explain the choices of women being a necessity entrepreneur.

❑ The dependent variable here is a dummy variable categorized in necessity and opportunity entrepreneurs.

❑ A mixture of qualitative and quantitative variables are used as regressors.

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)

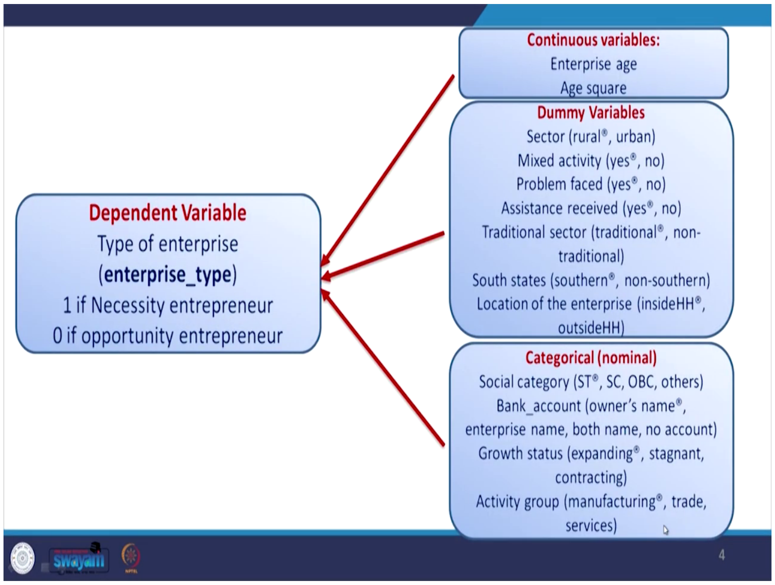


Note!

The purpose of this lecture is to show how to use various data analysis commands. It does not cover the whole research process such as verification of assumptions, model diagnostics and various follow up analysis.

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)



Continuous variables:
Enterprise age
Age square

Dummy Variables
Sector (rural®, urban)
Mixed activity (yes®, no)
Problem faced (yes®, no)
Assistance received (yes®, no)
Traditional sector (traditional®, non-traditional)
South states (southern®, non-southern)
Location of the enterprise (insideHH®, outsideHH)

Dependent Variable
Type of enterprise
(**enterprise_type**)
1 if Necessity entrepreneur
0 if opportunity entrepreneur

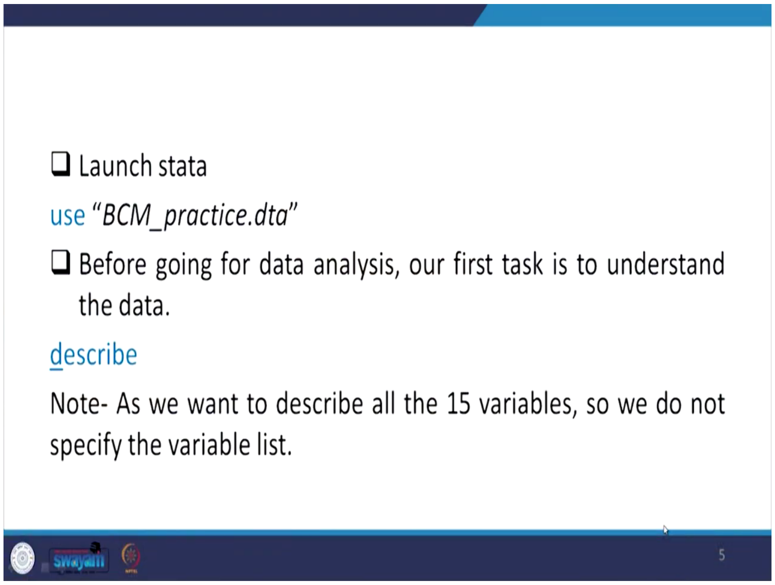
Categorical (nominal)
Social category (ST®, SC, OBC, others)
Bank_account (owner's name®, enterprise name, both name, no account)
Growth status (expanding®, stagnant, contracting)
Activity group (manufacturing®, trade, services)

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)



❑ Launch stata

use *"BCM_practice.dta"*

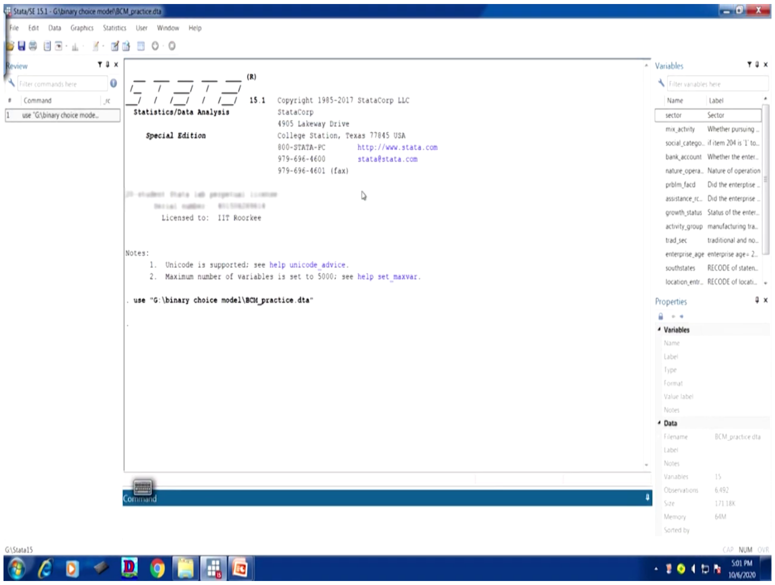
❑ Before going for data analysis, our first task is to understand the data.

describe

Note- As we want to describe all the 15 variables, so we do not specify the variable list.

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

(Refer Slide Time: 05:10)



#	Command
1	use "G:\binary choice mode...

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

[redacted]
       Licensed to:  IIT Roorkee

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

. use "G:\binary choice model\BCM_practice.dta"

.
```

Name	Label
sector	Sector
mix_activty	Whether pursuing ...
social_catego...	if item 204 is '1' to...
bank_account	Whether the enter...
nature_opera...	Nature of operation
prblm_facd	Did the enterprise ...
assistance_rc...	Did the enterprise ...
growth_status	Status of the enter...
activity_group	manufacturing tra...
trad_sec	traditional and no...
enterprise_age	enterprise age= 2...
southstates	RECODE of staten...
location_entr...	RECODE of locati...

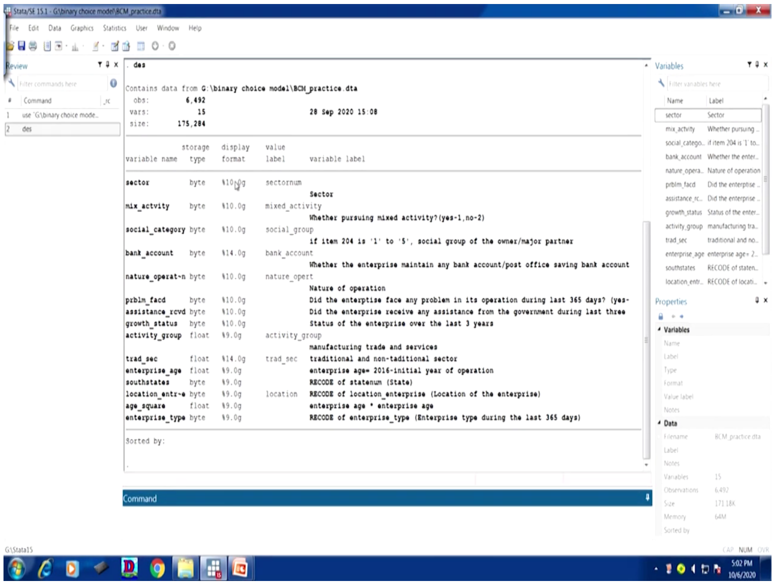
Variables
Name
Label
Type
Format
Value label
Notes

Data	
Filename	BCM_practice.dta
Label	
Notes	
Variables	15
Observations	6,492
Size	171.18K
Memory	64M
Sorted by	

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.

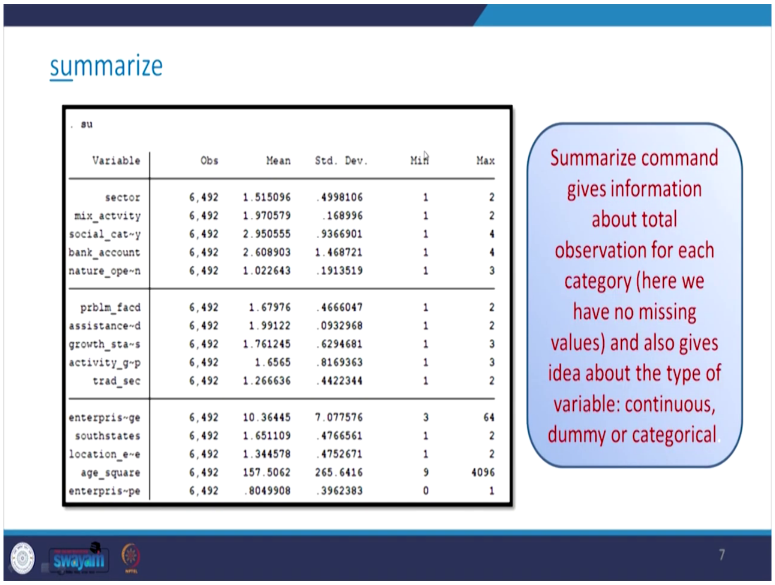
(Refer Slide Time: 06:10)



```
. des

Contains data from G:\binary choice model\BCM_practice.dta
  obs:         6,492
 vars:            15                          28 Sep 2020 15:08
 size:       175,284

              storage   display    value
variable name   type    format     label      variable label

sector          byte    %10.0g     sectornum
                                              Sector
mix_actvity     byte    %10.0g     mixed_activity
                                              Whether pursuing mixed activity?(yes-1,no-2)
social_category byte    %10.0g     social_group
                                              if item 204 is '1' to '5', social group of the owner/major partner
bank_account    byte    %14.0g     bank_account
                                              Whether the enterprise maintain any bank account/post office saving bank account
nature_operat~n byte    %10.0g     nature_opert
                                              Nature of operation
prblm_facd      byte    %10.0g                Did the enterptise face any problem in its operation during last 365 days? (yes-
assistance_rcvd byte    %10.0g                Did the enterprise receive any assistance from the government during last three
growth_status   byte    %10.0g                Status of the enterprise over the last 3 years
activity_group  float   %9.0g      activity_group
                                              manufacturing trade and services
trad_sec        float   %14.0g     trad_sec   traditional and non-taditional sector
enterprise_age  float   %9.0g                 enterprise age= 2016-initial year of operation
southstates     byte    %9.0g                 RECODE of statenum (State)
location_entr~e byte    %9.0g      location   RECODE of location_enterprise (Location of the enterprise)
age_square      float   %9.0g                 enterprise age * enterprise age
enterprise_type byte    %9.0g                 RECODE of enterprise_type (Enterprise type during the last 365 days)

Sorted by:
```

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.

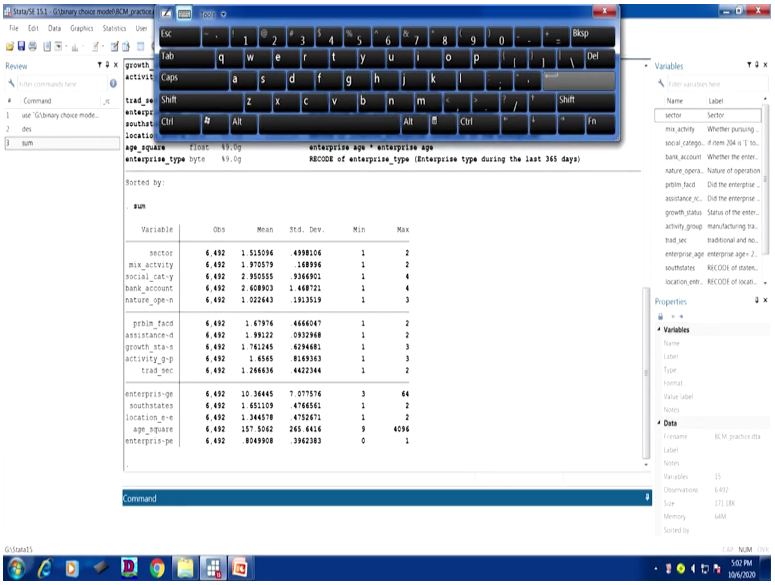
(Refer Slide Time: 06:55)



summarize

```
. su

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      sector |      6,492    1.515096    .4998106          1          2
 mix_actvity |      6,492    1.970579     .168996          1          2
social_cat~y |      6,492    2.950555    .9366901          1          4
bank_account |      6,492    2.608903    1.468721          1          4
nature_ope~n |      6,492    1.022643    .1913519          1          3
-------------+---------------------------------------------------------
  prblm_facd |      6,492     1.67976    .4666047          1          2
assistance~d |      6,492     1.99122    .0932968          1          2
growth_sta~s |      6,492    1.761245    .6294681          1          3
activity_g~p |      6,492      1.6565    .8169363          1          3
    trad_sec |      6,492    1.266636    .4422344          1          2
-------------+---------------------------------------------------------
enterpris~ge |      6,492    10.36445    7.077576          3         64
 southstates |      6,492    1.651109    .4766561          1          2
location_e~e |      6,492    1.344578    .4752671          1          2
  age_square |      6,492    157.5062    265.6416          9       4096
enterpris~pe |      6,492    .8049908    .3962383          0          1
```

Summarize command gives information about total observation for each category (here we have no missing values) and also gives idea about the type of variable: continuous, dummy or categorical



```
age_square      float   %9.0g                 enterprise age * enterprise age
enterprise_type byte    %9.0g                 RECODE of enterprise_type (Enterprise type during the last 365 days)

Sorted by:

. sum
```

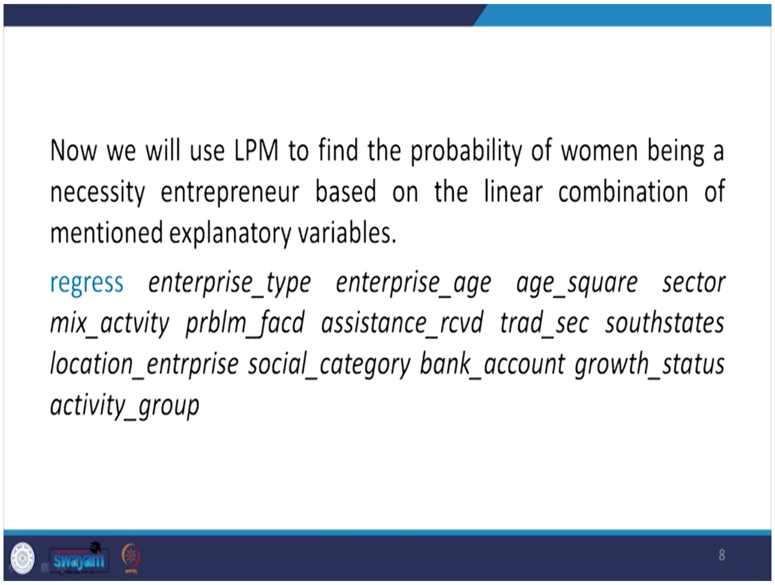
```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      sector |      6,492    1.515096    .4998106          1          2
 mix_actvity |      6,492    1.970579     .168996          1          2
social_cat~y |      6,492    2.950555    .9366901          1          4
bank_account |      6,492    2.608903    1.468721          1          4
nature_ope~n |      6,492    1.022643    .1913519          1          3
-------------+---------------------------------------------------------
  prblm_facd |      6,492     1.67976    .4666047          1          2
assistance~d |      6,492     1.99122    .0932968          1          2
growth_sta~s |      6,492    1.761245    .6294681          1          3
activity_g~p |      6,492      1.6565    .8169363          1          3
    trad_sec |      6,492    1.266636    .4422344          1          2
-------------+---------------------------------------------------------
enterpris~ge |      6,492    10.36445    7.077576          3         64
 southstates |      6,492    1.651109    .4766561          1          2
location_e~e |      6,492    1.344578    .4752671          1          2
  age_square |      6,492    157.5062    265.6416          9       4096
enterpris~pe |      6,492    .8049908    .3962383          0          1
```

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.

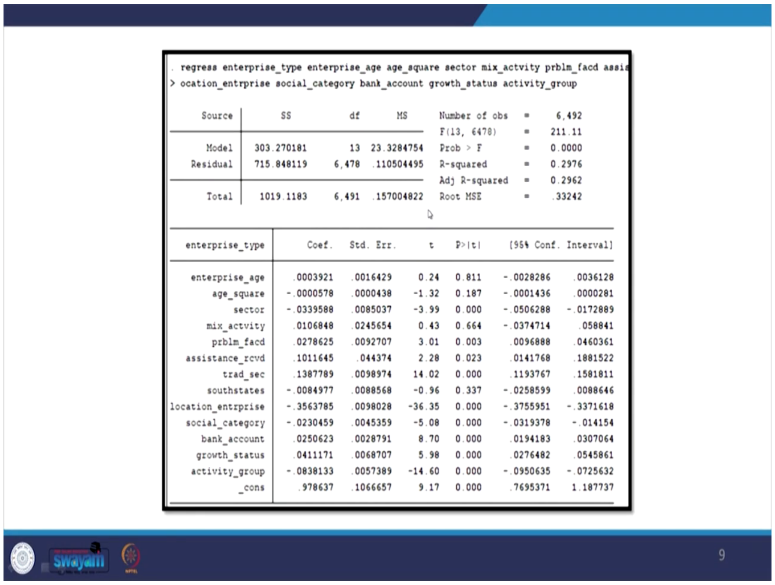
Then from 3 to 64, 9 to 4000, this very clearly seem that there are continuous variables. So, we have clarity on this aspect as well.

(Refer Slide Time: 07:52)



Now we will use LPM to find the probability of women being a necessity entrepreneur based on the linear combination of mentioned explanatory variables.

regress *enterprise_type enterprise_age age_square sector mix_actvity prblm_facd assistance_rcvd trad_sec southstates location_entrprise social_category bank_account growth_status activity_group*



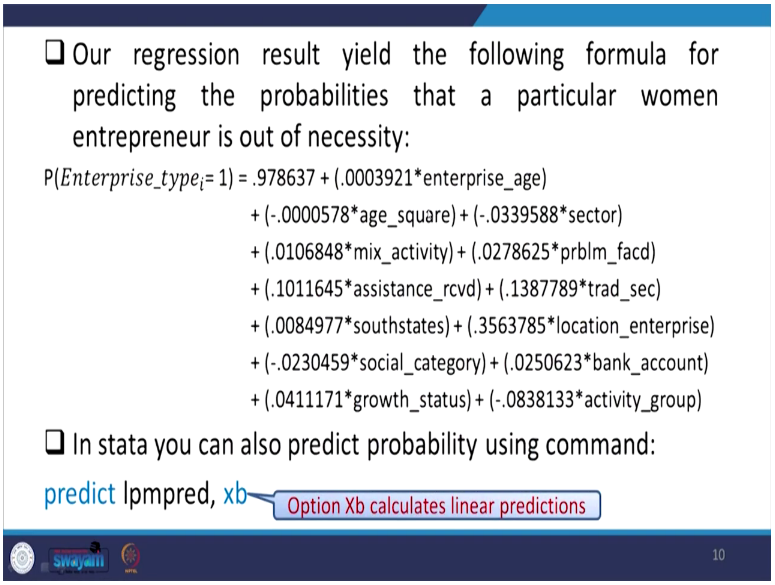
```
. regress enterprise_type enterprise_age age_square sector mix_actvity prblm_facd assis
> ocation_entrprise social_category bank_account growth_status activity_group

      Source |       SS           df       MS      Number of obs   =     6,492
-------------+----------------------------------   F(13, 6478)     =    211.11
       Model |  303.270181        13  23.3284754   Prob > F        =    0.0000
    Residual |  715.848119     6,478  .110504495   R-squared       =    0.2976
-------------+----------------------------------   Adj R-squared   =    0.2962
       Total |   1019.1183     6,491  .157004822   Root MSE        =    .33242

------------------------------------------------------------------------------------
   enterprise_type |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
    enterprise_age |   .0003921   .0016429     0.24   0.811    -.0028286    .0036128
        age_square |  -.0000578   .0000438    -1.32   0.187    -.0001436    .0000281
            sector |  -.0339588   .0085037    -3.99   0.000    -.0506288   -.0172889
       mix_actvity |   .0106848   .0245654     0.43   0.664    -.0374714     .058841
        prblm_facd |   .0278625   .0092707     3.01   0.003     .0096888    .0460361
   assistance_rcvd |   .1011645    .044374     2.28   0.023     .0141768    .1881522
          trad_sec |   .1387789   .0098974    14.02   0.000     .1193767    .1581811
       southstates |  -.0084977   .0088568    -0.96   0.337    -.0258599    .0088646
location_entrprise |  -.3563785   .0098028   -36.35   0.000    -.3755951   -.3371618
   social_category |  -.0230459   .0045359    -5.08   0.000    -.0319378    -.014154
      bank_account |   .0250623   .0028791     8.70   0.000     .0194183    .0307064
     growth_status |   .0411171   .0068707     5.98   0.000     .0276482    .0545861
    activity_group |  -.0838133   .0057389   -14.60   0.000    -.0950635   -.0725632
             _cons |    .978637   .1066657     9.17   0.000     .7695371    1.187737
------------------------------------------------------------------------------------
```

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)



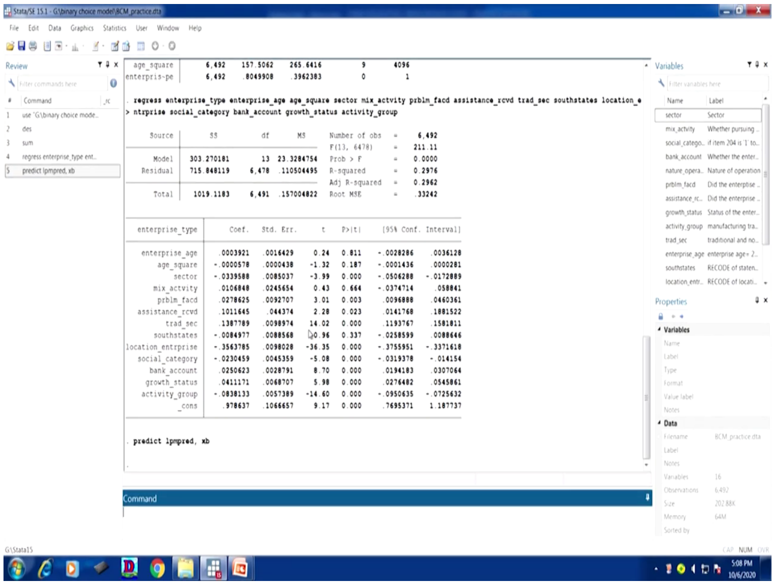
- Our regression result yield the following formula for predicting the probabilities that a particular women entrepreneur is out of necessity:

$$\begin{aligned} P(Enterprise\_type_i = 1) = {} & .978637 + (.0003921*\text{enterprise\_age}) \\ & + (-.0000578*\text{age\_square}) + (-.0339588*\text{sector}) \\ & + (.0106848*\text{mix\_activity}) + (.0278625*\text{prblm\_facd}) \\ & + (.1011645*\text{assistance\_rcvd}) + (.1387789*\text{trad\_sec}) \\ & + (.0084977*\text{southstates}) + (.3563785*\text{location\_enterprise}) \\ & + (-.0230459*\text{social\_category}) + (.0250623*\text{bank\_account}) \\ & + (.0411171*\text{growth\_status}) + (-.0838133*\text{activity\_group}) \end{aligned}$$

- In stata you can also predict probability using command:

predict lpmpred, xb

Option Xb calculates linear predictions



```
  age_square |      6,492    157.5062    265.6416          9       4096
enterpris~pe |      6,492    .8049908    .3962383          0          1

. regress enterprise_type enterprise_age age_square sector mix_actvity prblm_facd assistance_rcvd trad_sec southstates location_e
> ntrprise social_category bank_account growth_status activity_group

      Source |       SS           df       MS      Number of obs   =     6,492
-------------+----------------------------------   F(13, 6478)     =    211.11
       Model |  303.270181        13  23.3284754   Prob > F        =    0.0000
    Residual |  715.848119     6,478  .110504495   R-squared       =    0.2976
-------------+----------------------------------   Adj R-squared   =    0.2962
       Total |  1019.1183      6,491  .157004822   Root MSE        =    .33242

------------------------------------------------------------------------------------
   enterprise_type |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
    enterprise_age |   .0003921   .0016429     0.24   0.811    -.0028286    .0036128
        age_square |  -.0000578   .0000438    -1.32   0.187    -.0001436    .0000281
            sector |  -.0339588   .0085037    -3.99   0.000    -.0506288   -.0172889
       mix_actvity |   .0106848   .0245654     0.43   0.664    -.0374714     .058841
        prblm_facd |   .0278625   .0092707     3.01   0.003     .0096888    .0460361
   assistance_rcvd |   .1011645    .044374     2.28   0.023     .0141768    .1881522
          trad_sec |   .1387789   .0098974    14.02   0.000     .1193767    .1581811
       southstates |  -.0084977   .0088568    -0.96   0.337    -.0258599    .0088646
location_entrprise |  -.3563785   .0098028   -36.35   0.000    -.3755951   -.3371618
   social_category |  -.0230459   .0045359    -5.08   0.000    -.0319378    -.014154
      bank_account |   .0250623   .0028791     8.70   0.000     .0194183    .0307064
     growth_status |   .0411171   .0068707     5.98   0.000     .0276482    .0545861
    activity_group |  -.0838133   .0057389   -14.60   0.000    -.0950635   -.0725632
             _cons |    .978637   .1066657     9.17   0.000     .7695371    1.187737
------------------------------------------------------------------------------------

. predict lpmpred, xb
```

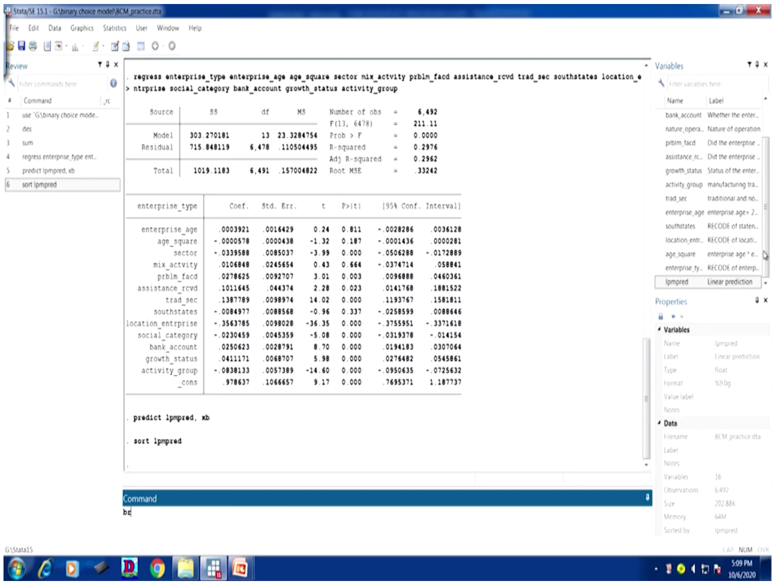
How we can interpret the result? This is the result we have derived just now. How we are going to interpret it? Basically as we know that our dependent variable is not just Y, it is P. It is probability of enterprise type 1 that is necessity entrepreneurs is equal to the constant term. This is the constant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

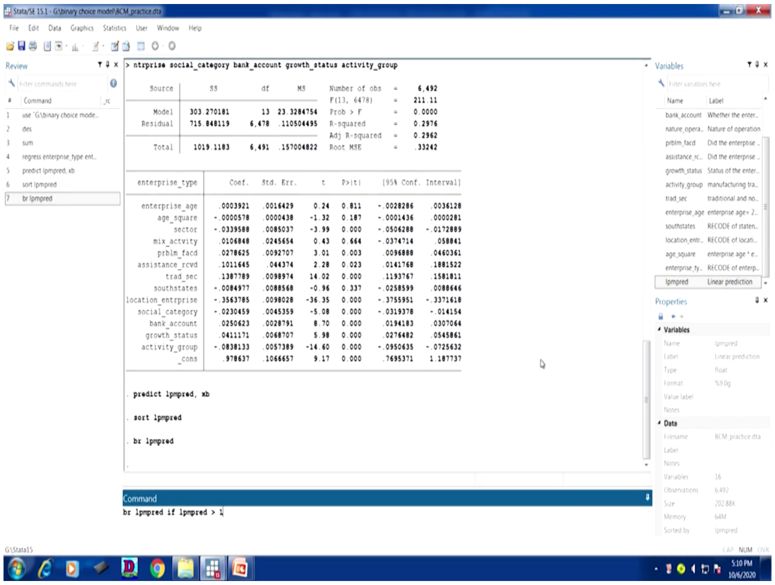
(Refer Slide Time: 12:41)


Stata/SE 15.1 - G:\binary choice model\BCM_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
1 use 'G:\binary choice mode...
2 des
3 sum
4 regress enterprise_type ent...
5 predict lpmpred, xb
6 sort lpmpred
Variables
Properties
Command
G:\Stata15

```
. regress enterprise_type enterprise_age age_square sector mix_actvity prblm_facd assistance_rcvd trad_sec southstates location_e
> ntrprise social_category bank_account growth_status activity_group

      Source |       SS           df       MS      Number of obs   =     6,492
-------------+----------------------------------   F(13, 6478)     =    211.11
       Model |  303.270181        13  23.3284754   Prob > F        =    0.0000
    Residual |  715.848119     6,478  .110504495   R-squared       =    0.2976
-------------+----------------------------------   Adj R-squared   =    0.2962
       Total |   1019.1183     6,491  .157004822   Root MSE        =    .33242

------------------------------------------------------------------------------
enterprise_type |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
 enterprise_age |   .0003921   .0016429     0.24   0.811    -.0028286    .0036128
     age_square |  -.0000578   .0000438    -1.32   0.187    -.0001436    .0000281
         sector |  -.0339588   .0085037    -3.99   0.000    -.0506288   -.0172889
    mix_actvity |   .0106848   .0245654     0.43   0.664    -.0374714     .058841
     prblm_facd |   .0278625   .0092707     3.01   0.003     .0096888    .0460361
assistance_rcvd |   .1011645    .044374     2.28   0.023     .0141768    .1881522
       trad_sec |   .1387789   .0098974    14.02   0.000     .1193767    .1581811
    southstates |  -.0084977   .0088568    -0.96   0.337    -.0258599    .0088646
location_entrprise | -.3563785 .0098028  -36.35   0.000    -.3755951   -.3371618
social_category |  -.0230459   .0045359    -5.08   0.000    -.0319378    -.014154
   bank_account |   .0250623   .0028791     8.70   0.000     .0194183    .0307064
  growth_status |   .0411171   .0068707     5.98   0.000     .0276482    .0545861
 activity_group |  -.0838133   .0057389   -14.60   0.000    -.0950635   -.0725632
          _cons |    .978637   .1066657     9.17   0.000     .7695371    1.187737
------------------------------------------------------------------------------

. predict lpmpred, xb

. sort lpmpred

.
```

Command

```
br
```

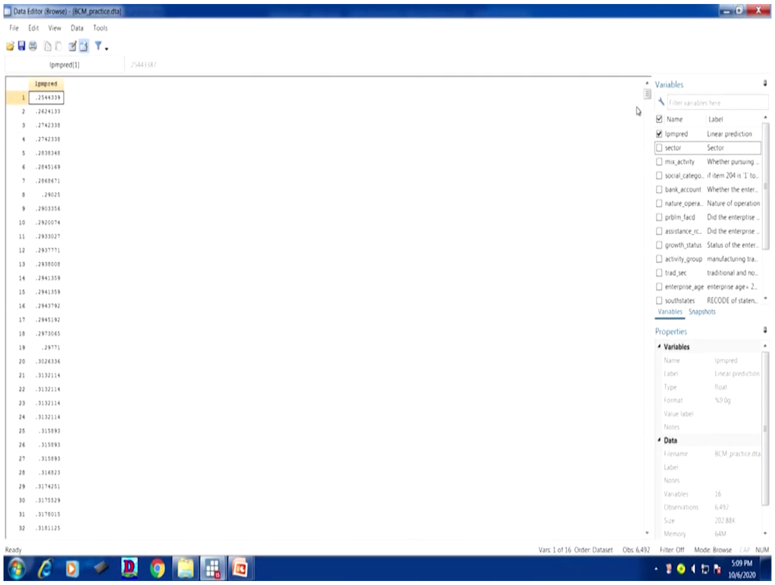

Stata/SE 15.1 - G:\binary choice model\BCM_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
1 use 'G:\binary choice mode...
2 des
3 sum
4 regress enterprise_type ent...
5 predict lpmpred, xb
6 sort lpmpred
7 br lpmpred
Variables
Properties
Command
G:\Stata15

```
> ntrprise social_category bank_account growth_status activity_group

      Source |       SS           df       MS      Number of obs   =     6,492
-------------+----------------------------------   F(13, 6478)     =    211.11
       Model |  303.270181        13  23.3284754   Prob > F        =    0.0000
    Residual |  715.848119     6,478  .110504495   R-squared       =    0.2976
-------------+----------------------------------   Adj R-squared   =    0.2962
       Total |   1019.1183     6,491  .157004822   Root MSE        =    .33242

------------------------------------------------------------------------------
enterprise_type |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
 enterprise_age |   .0003921   .0016429     0.24   0.811    -.0028286    .0036128
     age_square |  -.0000578   .0000438    -1.32   0.187    -.0001436    .0000281
         sector |  -.0339588   .0085037    -3.99   0.000    -.0506288   -.0172889
    mix_actvity |   .0106848   .0245654     0.43   0.664    -.0374714     .058841
     prblm_facd |   .0278625   .0092707     3.01   0.003     .0096888    .0460361
assistance_rcvd |   .1011645    .044374     2.28   0.023     .0141768    .1881522
       trad_sec |   .1387789   .0098974    14.02   0.000     .1193767    .1581811
    southstates |  -.0084977   .0088568    -0.96   0.337    -.0258599    .0088646
location_entrprise | -.3563785 .0098028  -36.35   0.000    -.3755951   -.3371618
social_category |  -.0230459   .0045359    -5.08   0.000    -.0319378    -.014154
   bank_account |   .0250623   .0028791     8.70   0.000     .0194183    .0307064
  growth_status |   .0411171   .0068707     5.98   0.000     .0276482    .0545861
 activity_group |  -.0838133   .0057389   -14.60   0.000    -.0950635   -.0725632
          _cons |    .978637   .1066657     9.17   0.000     .7695371    1.187737
------------------------------------------------------------------------------

. predict lpmpred, xb

. sort lpmpred
```

```
. br lpmpred

.
```

Command

```
br lpmpred if lpmpred > 1
```

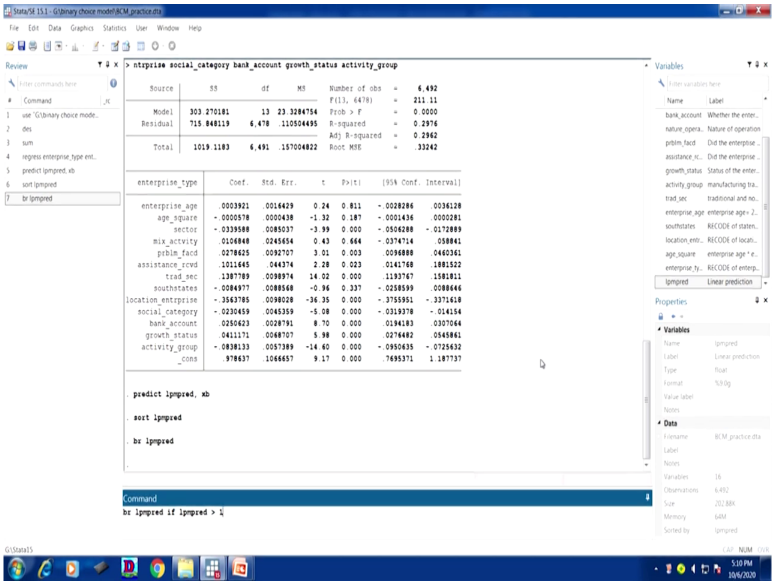


Data Editor (Browse) - [BCM_practice.dta]

```
lpmpred[1]      .25443387

     lpmpred
 1  .2544339
 2  .2624133
 3  .2742338
 4  .2742338
 5  .2838348
 6  .2845169
 7  .2868671
 8    .29025
 9  .2903356
10  .2920074
11  .2933027
12  .2937771
13  .2938008
14  .2941359
15  .2941359
16  .2943792
17  .2945192
18  .2973065
19    .29771
20  .3026336
21  .3132114
22  .3132114
23  .3132114
24  .3132114
25   .315893
26   .315893
27   .315893
28   .316823
29  .3174251
30  .3175529
31  .3178015
32  .3181125
```

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\BCM_practice.dta

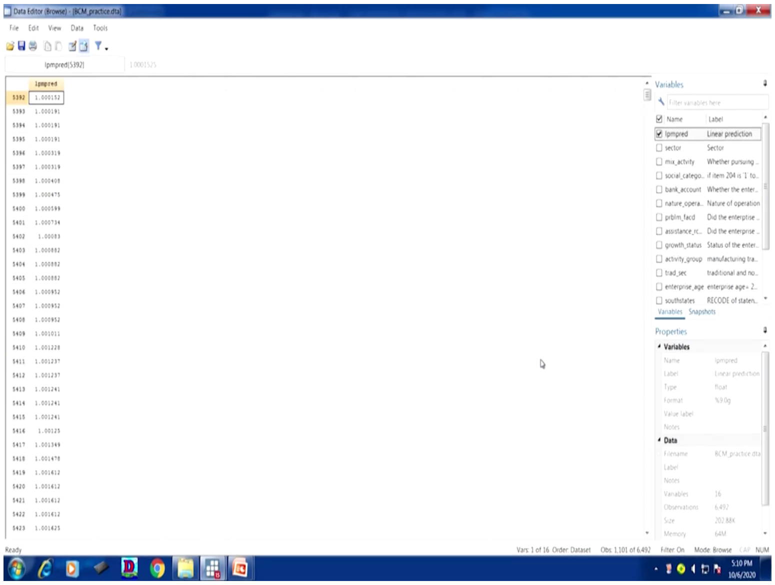
Review

#	Command
1	use "G:\binary choice mode...
2	des
3	sum
4	regress enterprise_type ent...
5	predict lpmpred, xb
6	sort lpmpred
7	br lpmpred

```
> ntrprise social_category bank_account growth_status activity_group

      Source |       SS           df       MS      Number of obs   =     6,492
-------------+----------------------------------   F(13, 6478)     =    211.11
       Model |  303.270181        13  23.3284754   Prob > F        =    0.0000
    Residual |  715.848119     6,478  .110504495   R-squared       =    0.2976
-------------+----------------------------------   Adj R-squared   =    0.2962
       Total |  1019.1183      6,491  .157004822   Root MSE        =    .33242

------------------------------------------------------------------------------
enterprise_type |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
 enterprise_age |   .0003921   .0016429     0.24   0.811    -.0028286    .0036128
     age_square |  -.0000578   .0000438    -1.32   0.187    -.0001436    .0000281
         sector |  -.0339588   .0085037    -3.99   0.000    -.0506288   -.0172889
    mix_actvity |   .0106848   .0245654     0.43   0.664    -.0374714     .058841
     prblm_facd |   .0278625   .0092707     3.01   0.003     .0096888    .0460361
assistance_rcvd |   .1011645    .044374     2.28   0.023     .0141768    .1881522
       trad_sec |   .1387789   .0098974    14.02   0.000     .1193767    .1581811
    southstates |  -.0084977   .0088568    -0.96   0.337    -.0258599    .0088646
location_entrprise | -.3563785 .0098028  -36.35   0.000    -.3755951   -.3371618
social_category |  -.0230459   .0045359    -5.08   0.000    -.0319378    -.014154
   bank_account |   .0250623   .0028791     8.70   0.000     .0194183    .0307064
  growth_status |   .0411171   .0068707     5.98   0.000     .0276482    .0545861
 activity_group |  -.0838133   .0057389   -14.60   0.000    -.0950635   -.0725632
          _cons |    .978637   .1066657     9.17   0.000     .7695371    1.187737
------------------------------------------------------------------------------

. predict lpmpred, xb

. sort lpmpred

. br lpmpred
```

Command

```
br lpmpred if lpmpred > 1
```



Data Editor (Browse) - [BCM_practice.dta]
File Edit View Data Tools

lpmpred[5392]   1.0001525

	lpmpred
5392	1.000152
5393	1.000191
5394	1.000191
5395	1.000191
5396	1.000319
5397	1.000319
5398	1.000408
5399	1.000475
5400	1.000599
5401	1.000734
5402	1.00083
5403	1.000882
5404	1.000882
5405	1.000882
5406	1.000952
5407	1.000952
5408	1.000952
5409	1.001011
5410	1.001228
5411	1.001237
5412	1.001237
5413	1.001241
5414	1.001241
5415	1.001241
5416	1.00125
5417	1.001349
5418	1.001478
5419	1.001612
5420	1.001612
5421	1.001612
5422	1.001612
5423	1.001625

Variables

☑ Name	Label
☑ lpmpred	Linear prediction
□ sector	Sector
□ mix_actvity	Whether pursuing ...
□ social_catego...	if item 204 is '1' to...
□ bank_account	Whether the enter...
□ nature_opera...	Nature of operation
□ prblm_facd	Did the enterptise ...
□ assistance_rc...	Did the enterprise ...
□ growth_status	Status of the enter...
□ activity_group	manufacturing tra...
□ trad_sec	traditional and no...
□ enterprise_age	enterprise age= 2...
□ southstates	RECODE of staten...

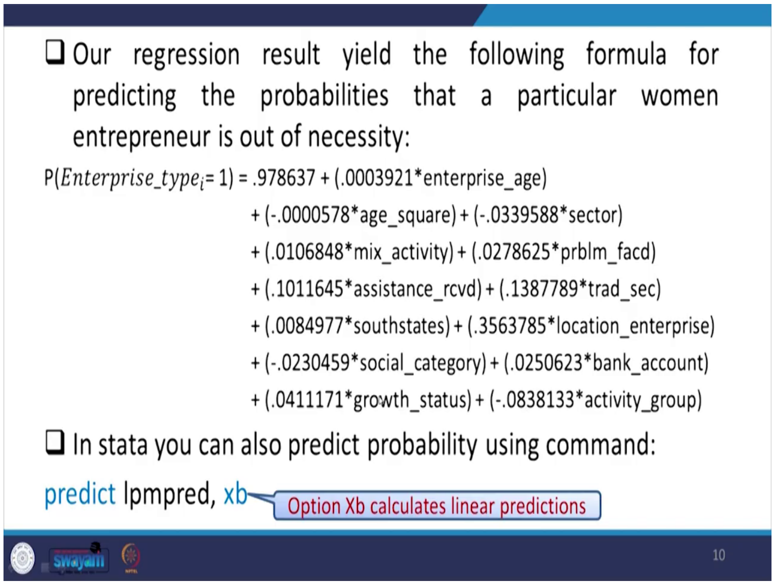
Properties

Variables	
Name	lpmpred
Label	Linear prediction
Type	float
Format	%9.0g
Value label	
Notes	
Data	
Filename	BCM_practice.dta
Label	
Notes	
Variables	16
Observations	6,492
Size	202.88K
Memory	64M

Ready   Vars: 1 of 16 Order: Dataset   Obs: 1,101 of 6,492   Filter: On   Mode: Browse   CAP   NUM

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)



- ❑ Our regression result yield the following formula for predicting the probabilities that a particular women entrepreneur is out of necessity:

$$\begin{aligned}P(Enterprise\_type_i = 1) = {} & .978637 + (.0003921*\text{enterprise\_age}) \\ & + (-.0000578*\text{age\_square}) + (-.0339588*\text{sector}) \\ & + (.0106848*\text{mix\_activity}) + (.0278625*\text{prblm\_facd}) \\ & + (.1011645*\text{assistance\_rcvd}) + (.1387789*\text{trad\_sec}) \\ & + (.0084977*\text{southstates}) + (.3563785*\text{location\_enterprise}) \\ & + (-.0230459*\text{social\_category}) + (.0250623*\text{bank\_account}) \\ & + (.0411171*\text{growth\_status}) + (-.0838133*\text{activity\_group})\end{aligned}$$

- ❑ In stata you can also predict probability using command:

```
predict lpmpred, xb
```

Option Xb calculates linear predictions

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

browse

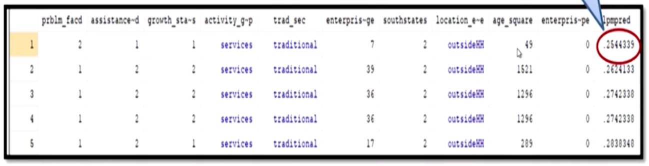
We can also sort the data by lpmpred:

sort *lpmpred*

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

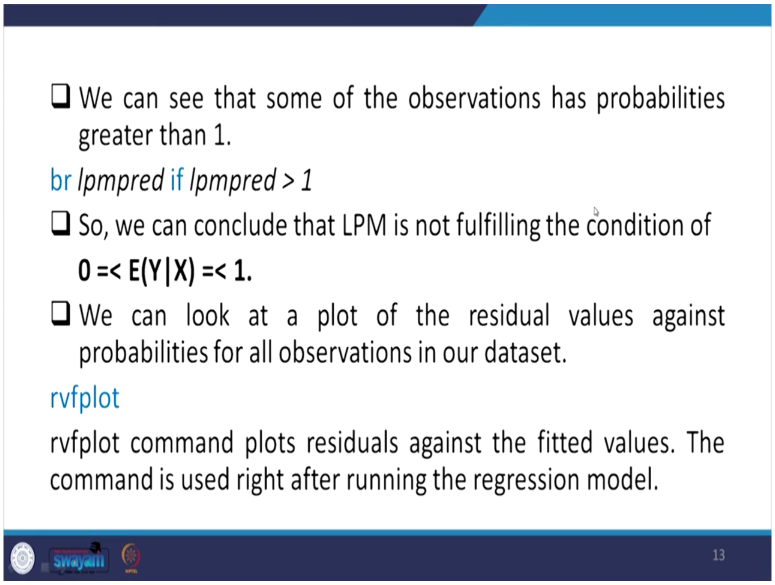
browse

This enterprise has 25.4% chances of being necessity



	prblm_facd	assistance~d	growth_sta~s	activity_g~p	trad_sec	enterpris~ge	southstates	location_e~e	age_square	enterpris~pe	lpmpred
1	2	1	1	services	traditional	7	2	outsideHH	49	0	.2544339
2	1	2	2	services	traditional	39	2	outsideHH	1521	0	.2624133
3	1	2	2	services	traditional	36	2	outsideHH	1296	0	.2742338
4	1	2	2	services	traditional	36	2	outsideHH	1296	0	.2742338
5	1	2	1	services	traditional	17	2	outsideHH	289	0	.2838348

- ❑ It is also assumed that predicted probabilities should lie between 0 and 1.
- ❑ By browsing the **lpmpred** variable we can look for values greater that one or less than zero.



□ We can see that some of the observations has probabilities greater than 1.

br *lpmpred* if *lpmpred > 1*

□ So, we can conclude that LPM is not fulfilling the condition of **0 =< E(Y|X) =< 1.**

□ We can look at a plot of the residual values against probabilities for all observations in our dataset.

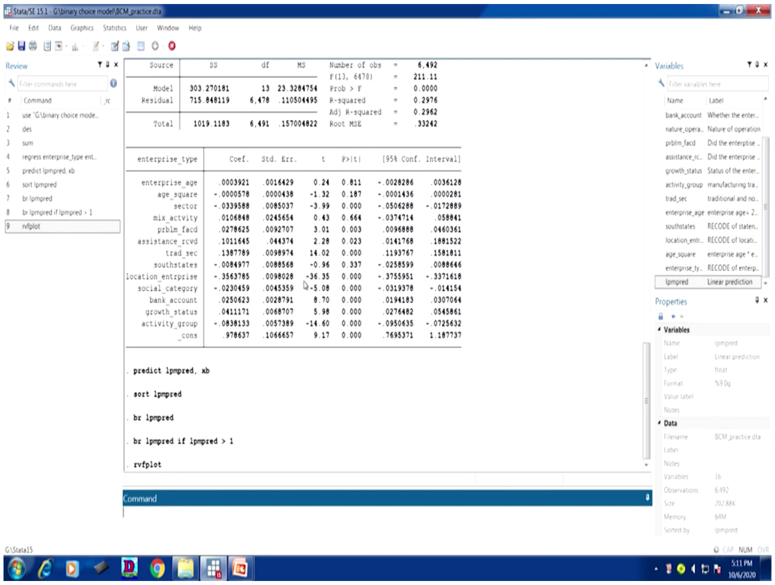
rvfplot

rvfplot command plots residuals against the fitted values. The command is used right after running the regression model.

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)



Stata/SE 15.1 - G:\binary choice model\BCM_practice.dta

Review

#	Command	_rc
1	use "G:\binary choice mode...	
2	des	
3	sum	
4	regress enterprise_type ent...	
5	predict lpmpred, xb	
6	sort lpmpred	
7	br lpmpred	
8	br lpmpred if lpmpred > 1	
9	rvfplot	

```
      Source |       SS           df       MS      Number of obs   =     6,492
-------------+----------------------------------   F(13, 6478)     =    211.11
       Model |  303.270181        13  23.3284754   Prob > F        =    0.0000
    Residual |  715.848119     6,478  .110504495   R-squared       =    0.2976
-------------+----------------------------------   Adj R-squared   =    0.2962
       Total |  1019.1183      6,491  .157004822   Root MSE        =    .33242

------------------------------------------------------------------------------
enterprise_type |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
 enterprise_age |   .0003921   .0016429     0.24   0.811    -.0028286    .0036128
     age_square |  -.0000578   .0000438    -1.32   0.187    -.0001436    .0000281
         sector |  -.0339588   .0085037    -3.99   0.000    -.0506288   -.0172889
    mix_actvity |   .0106848   .0245654     0.43   0.664    -.0374714     .058841
     prblm_facd |   .0278625   .0092707     3.01   0.003     .0096888    .0460361
assistance_rcvd |   .1011645    .044374     2.28   0.023     .0141768    .1881522
       trad_sec |   .1387789   .0098974    14.02   0.000     .1193767    .1581811
    southstates |  -.0084977   .0088568    -0.96   0.337    -.0258599    .0088646
location_entrprise |  -.3563785   .0098028   -36.35   0.000    -.3755951   -.3371618
social_category |  -.0230459   .0045359    -5.08   0.000    -.0319378    -.014154
   bank_account |   .0250623   .0028791     8.70   0.000     .0194183    .0307064
  growth_status |   .0411171   .0068707     5.98   0.000     .0276482    .0545861
 activity_group |  -.0838133   .0057389   -14.60   0.000    -.0950635   -.0725632
          _cons |    .978637   .1066657     9.17   0.000     .7695371    1.187737
------------------------------------------------------------------------------

. predict lpmpred, xb

. sort lpmpred

. br lpmpred

. br lpmpred if lpmpred > 1

. rvfplot
```

Variables

Name	Label
bank_account	Whether the enter...
nature_opera...	Nature of operation
prblm_facd	Did the enterptise ...
assistance_rc...	Did the enterprise ...
growth_status	Status of the enter...
activity_group	manufacturing tra...
trad_sec	traditional and no...
enterprise_age	enterprise age+ 2...
southstates	RECODE of staten...
location_entr...	RECODE of locati...
age_square	enterprise age * e...
enterprise_ty...	RECODE of enterp...
lpmpred	Linear prediction

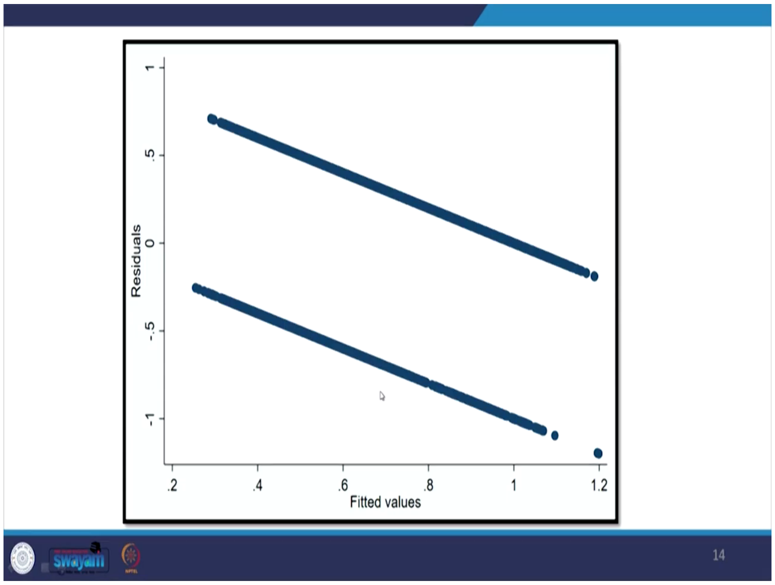
Properties

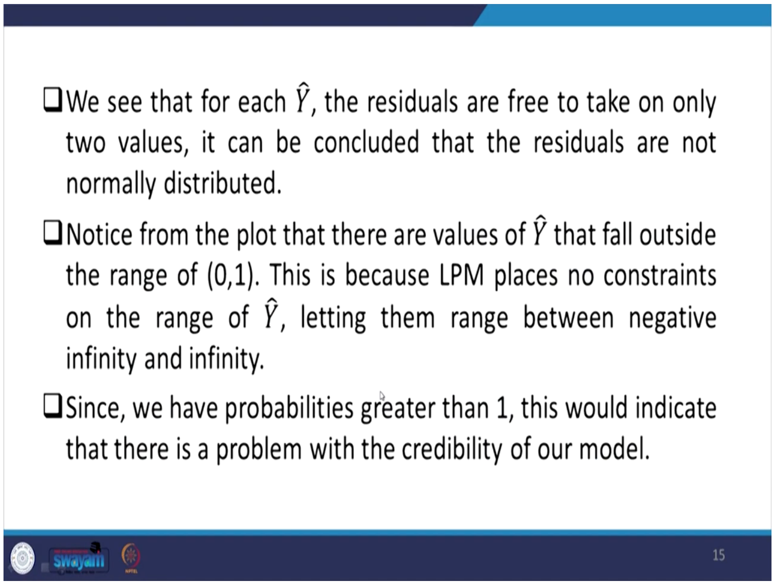
Variables

Name	lpmpred
Label	Linear prediction
Type	float
Format	%9.0g
Value label	
Notes	

Data

Filename	BCM_practice.dta
Label	
Notes	
Variables	16
Observations	6,492
Size	202.88K
Memory	64M
Sorted by	lpmpred





❑ We see that for each $\hat{Y}$, the residuals are free to take on only two values, it can be concluded that the residuals are not normally distributed.

❑ Notice from the plot that there are values of $\hat{Y}$ that fall outside the range of (0,1). This is because LPM places no constraints on the range of $\hat{Y}$, letting them range between negative infinity and infinity.

❑ Since, we have probabilities greater than 1, this would indicate that there is a problem with the credibility of our model.

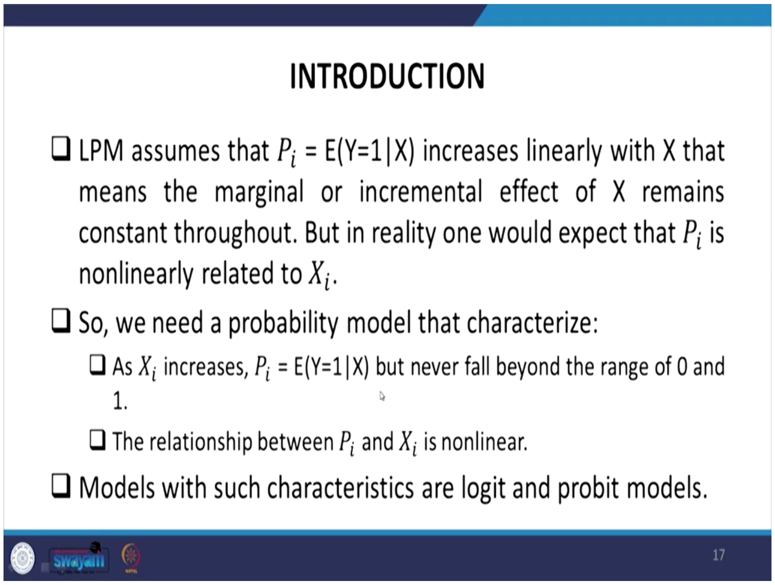
We can look at a plot also used to understand their residual values against probabilities of all the observations in our data set. So, rvf plot is going to give us that result. The residual is going to be defined. It has come.

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.

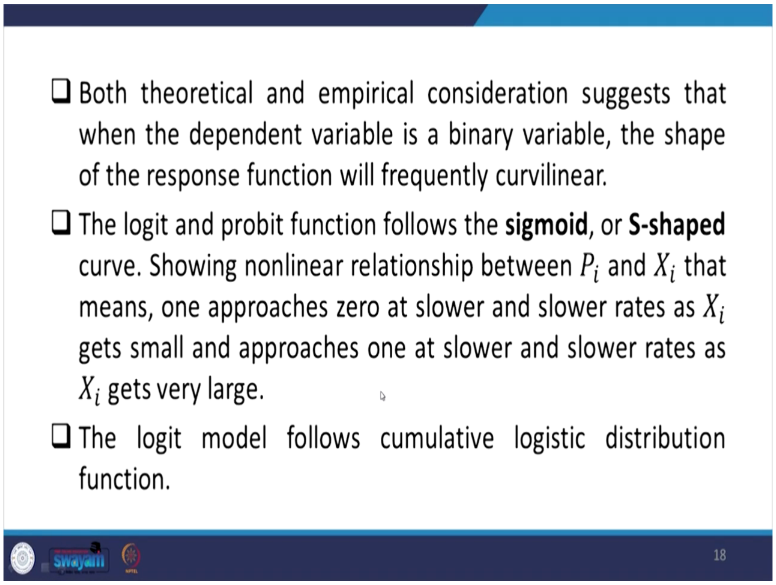
We are discussing another binary choice model, binary response model called Logit. Logit model is applied where the distribution follows a logistic distribution or logistic structure. So, we are going to discuss that in a short while.

(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.
  - ❑ The relationship between P_i and X_i is nonlinear.
- ❑ Models with such characteristics are logit and probit models.



- ❑ 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.

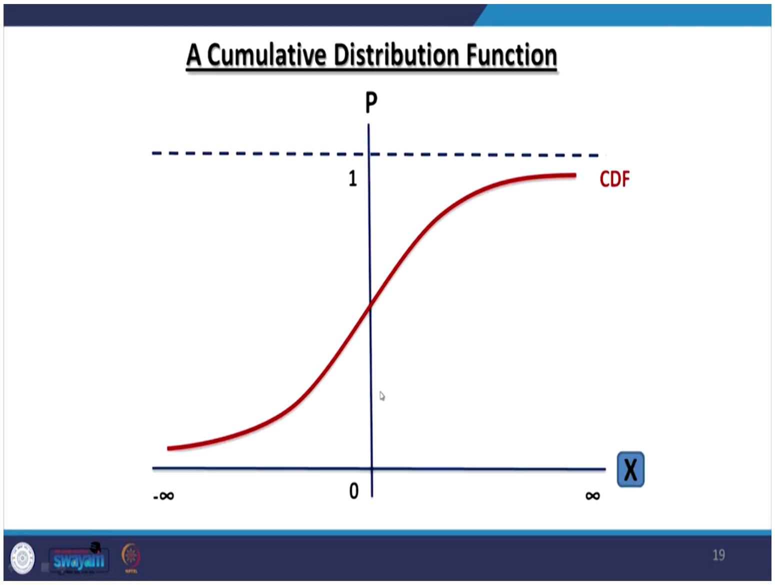
e

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.

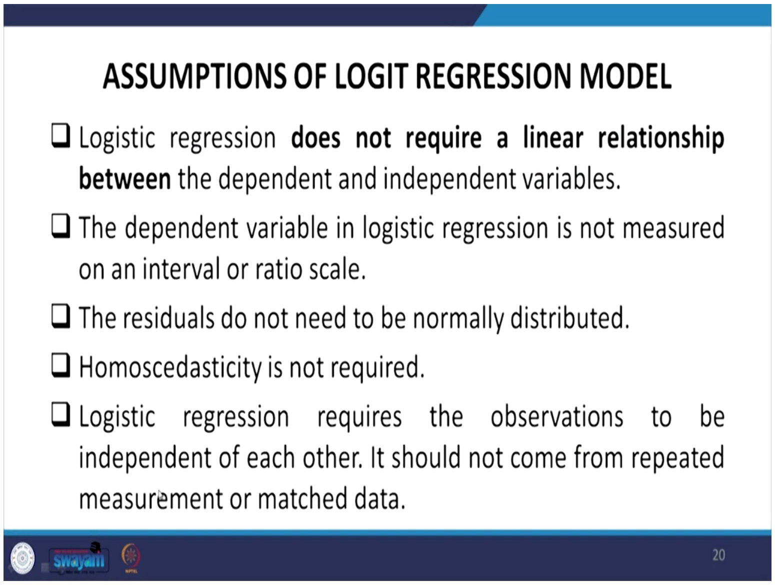
(Refer Slide Time: 20:06)



A Cumulative Distribution Function

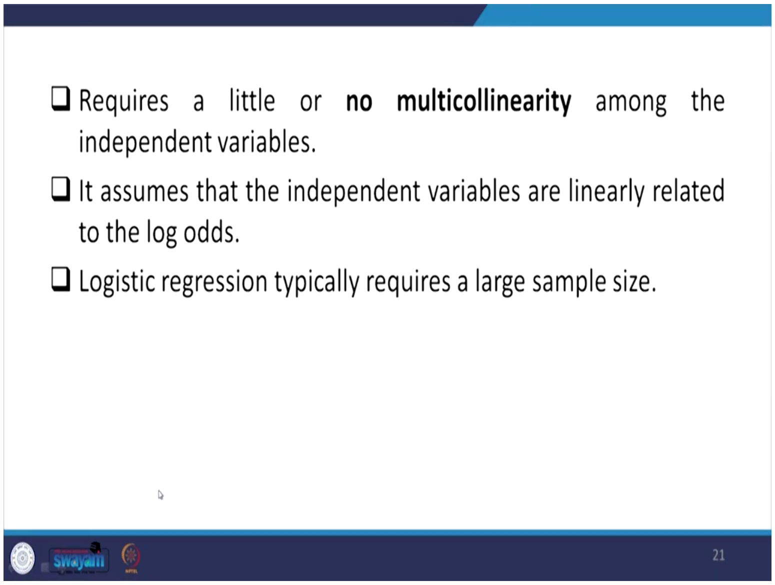
The Logit model follows a cumulative logistic distribution function. So, this looks like this. With the higher values, it approaches to a slower rate. Even at the lower value also approaches to a slower rate. So, that is all about the cumulative distribution function of a sigmoid curve or a Logit curve.

(Refer Slide Time: 20:26)



ASSUMPTIONS OF LOGIT REGRESSION MODEL

- ❑ Logistic regression **does not require a linear relationship between** the dependent and independent variables.
- ❑ The dependent variable in logistic regression is not measured on an interval or ratio scale.
- ❑ The residuals do not need to be normally distributed.
- ❑ Homoscedasticity is not required.
- ❑ Logistic regression requires the observations to be independent of each other. It should not come from repeated measurement or matched data.



- ❑ Requires a little or **no multicollinearity** among the independent variables.
- ❑ It assumes that the independent variables are linearly related to the log odds.
- ❑ Logistic regression typically requires a large sample size.

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.

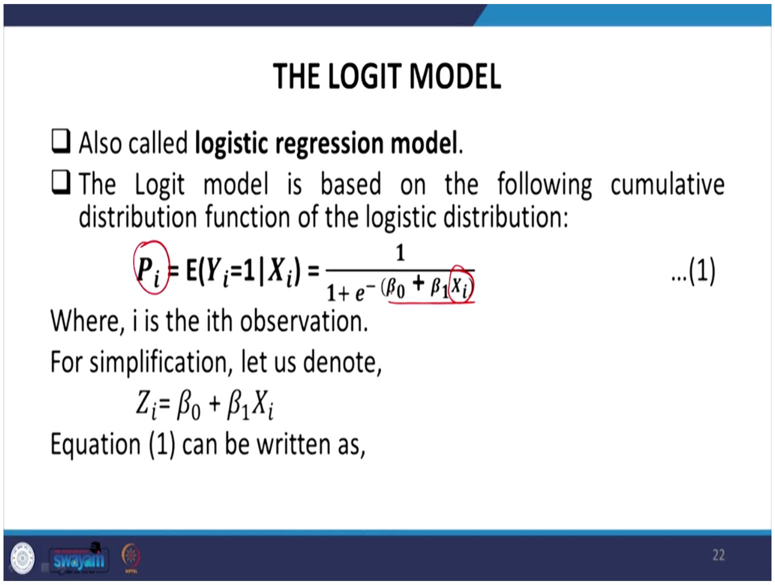
So far in the earlier models we only discussed linear relationship. So, Logit and even Probit not necessarily require linear relationship. That can be converted to a linear format for estimation but original they do not require linear relationship. The dependant variable in logistic regression is not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

(Refer Slide Time: 22:45)



THE LOGIT MODEL

❑ Also called **logistic regression model**.

❑ The Logit model is based on the following cumulative distribution function of the logistic distribution:

$$P_i = E(Y_i=1|X_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}} \qquad ...(1)$$

Where, i is the ith observation.

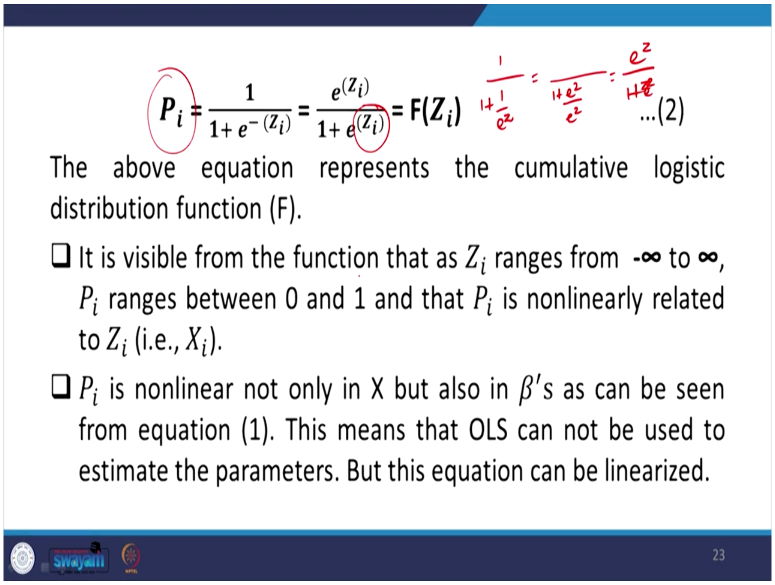
For simplification, let us denote,

$$Z_i = \beta_0 + \beta_1 X_i$$

Equation (1) can be written as,

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.

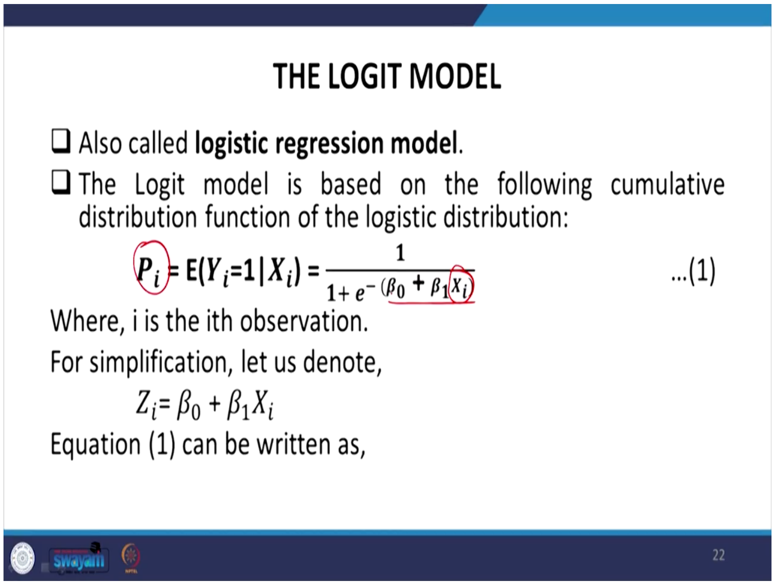
(Refer Slide Time: 23:56)



$$P_i = \frac{1}{1+e^{-(Z_i)}} = \frac{e^{(Z_i)}}{1+e^{(Z_i)}} = \mathrm{F}(Z_i) \qquad \frac{1}{1+\frac{1}{e^z}} = \frac{}{1+\frac{e^z}{e^z}} = \frac{e^z}{1+e^z} \qquad \ldots(2)$$

The above equation represents the cumulative logistic distribution function (F).

- ❑ It is visible from the function that as Z_i ranges from **-∞** to **∞**, P_i ranges between 0 and 1 and that P_i is nonlinearly related to Z_i (i.e., X_i).
- ❑ P_i is nonlinear not only in X but also in β's as can be seen from equation (1). This means that OLS can not be used to estimate the parameters. But this equation can be linearized.



THE LOGIT MODEL

- ❑ Also called **logistic regression model**.
- ❑ The Logit model is based on the following cumulative distribution function of the logistic distribution:

$$P_i = \mathrm{E}(Y_i = 1 | X_i) = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_i)}} \qquad \ldots(1)$$

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,

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.

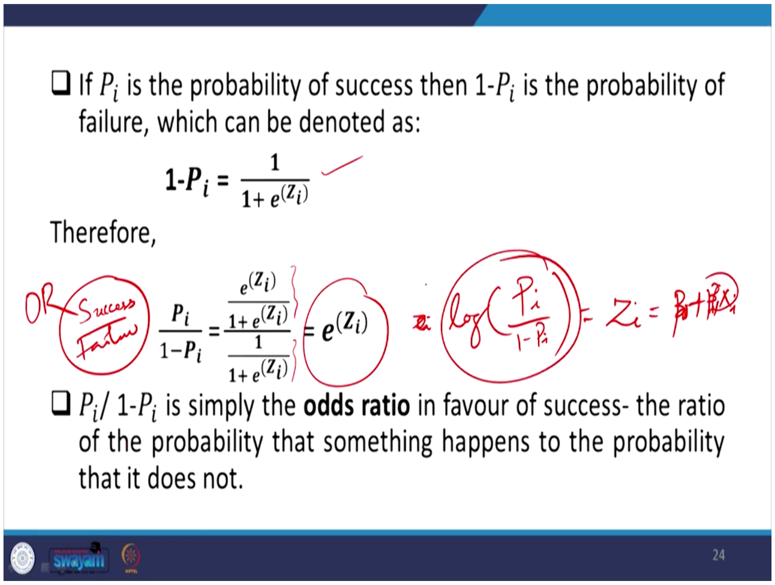
So, this is what e to the power Zi then 1 plus e to the power Zi which is nothing but called the distribution function of the Logistic model. So, this represents this F function or the cumulative logistic function. The probability at the any point can be defined through this particular equation so far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



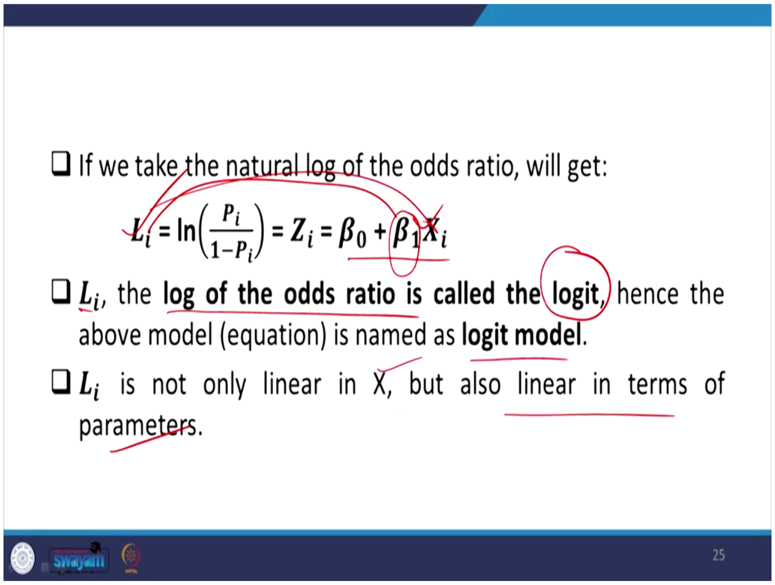
❑ If P_i is the probability of success then 1-P_i is the probability of failure, which can be denoted as:

$$1\text{-}P_i = \frac{1}{1+e^{(Z_i)}}$$

Therefore,

$$\text{OR} \; \frac{\text{Success}}{\text{Failure}} \quad \frac{P_i}{1-P_i} = \frac{\frac{e^{(Z_i)}}{1+e^{(Z_i)}}}{\frac{1}{1+e^{(Z_i)}}} = e^{(Z_i)} = \log\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_0 + \beta_1 X_i$$

❑ P_i/ 1-P_i is simply the **odds ratio** in favour of success- the ratio of the probability that something happens to the probability that it does not.



❑ If we take the natural log of the odds ratio, will get:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_0 + \beta_1 X_i$$

❑ L_i, the **log of the odds ratio is called the logit**, hence the above model (equation) is named as **logit model**.

❑ L_i is not only linear in X, but also linear in terms of parameters.

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.