

**Exploring Survey Data on Health Care**  
**Prof. Pratap C. Mohanty**  
**Department of Humanities and Social Sciences**  
**Indian Institute of Technology, Roorkee**

**Lecture - 27**  
**Regression Models of Qualitative Dependent Variables**

Welcome participants to my NPTEL MOOC module on Exploring Healthcare Survey Data. We are in on the week explaining influential statistics in healthcare. In the previous lecture, we discussed regression models using the quantitative dependent variables. Here we are tending towards explaining the dependent variable which is qualitative.




The regression models are not usually these models are also referred to as some generalized form of the least square model. We will discuss all those things now without consuming more time. Regression models that we have discussed so far implicitly assumed that the dependent variable Y is quantitative whereas, the exponential variable can take either form of variables may be a mixture, or maybe all quantitative maybe all qualitative.

Many economic or social phenomena of interest; however, concern variables that are not continuous or perhaps not even quantitative. But in this lecture we will discuss an analysis of the model when the dependent variable itself is of qualitative nature, you can just refer to this example for your own understanding.

(Refer Slide Time: 01:41)

### Introduction


- Regression models that we have discussed so far implicitly assumed that the dependent variable Y is quantitative, whereas the explanatory variable can take either form quantitative or qualitative or a mixture of both.
- Many economic or social phenomena of interest, however, concern variables that are not continuous or perhaps not even quantitative.
  - E.g. What determines the nature of treatment (traditional vs modern) in case of ailment.
- But in this lecture we will discuss, analysis of model when dependent variable itself is **qualitative in nature**.

2

(Refer Slide Time: 01:42)

**When Y is Quantitative**, the objective of a regression model is to estimate its expected, or mean value given the values of the regressors.

**When Y is Qualitative**, the objective of the model is to find the probability of something happening. qualitative response regression models are often known as **probability models**.




3

When the dependent variable is quantitative the objective of a regression in a model is to estimate it is expected or mean value given the values of the regressors. When the dependent variable that is Y is qualitative the objective of the model is to find the probability of something happening, qualitative response regression models are often known as probability models.

(Refer Slide Time: 02:11)

**Types of dependent variables and their estimation method :**

Dependent variable	Independent variable	Estimation Method
Continuous (Quantitative)	Quantitative/Qualitative	Ordinary Least Squares (OLS)
Binary(Qualitative)	Quantitative/Qualitative	LPM/Logit/ Probit Model
Categorical (Qualitative)	Quantitative/Qualitative	Multinomial Logit/Probit
Ordered Categorical(Qualitative)	Quantitative/Qualitative	Cumulative Logit/Probit
Repeated Binary (Qualitative)	Quantitative/Qualitative	Panel Logit/Probit



4

The types of dependent variables and their estimation methods we have made a comparison and you can understand very clearly what methods and what context is usually applied. If your dependent variable is continuous that is quantitative in nature and the independent variable may be mixed or in of the qualitative or quantitative, you are supposed to apply the OLS regression model.

Where your dependent variable is binary, but qualitative binary in an independent variable may be of a mixed kind. You are supposed to apply the linear probability model logit model or probit model. There are, there might be, not necessarily the dependent variable to be binary it might be categorical if it is categorically usually discussed as qualitative data, irrespective of the independent variables, multinomial logit regression is applied.

If your dependent variable is ordered; ordered categorical, but order like education is the variable educational argument. In that case, you are supposed to apply the cumulative logit model or cumulative probit model these are called cumulative models if your data that is dependent variable is binary, but repeated binary your responses are in fact, repeated. In that case, as we already discussed from the beginning of the data panel logit or probit is applied.

Since data is a kind of longitudinal or a kind of repeated one irrespective of the kind of independent variable you carry. First of all, we will discuss the starting of binary choice models. I am not discussing all the things in detail I will just discuss with you the basic overview of the theory from the beginning, then I will suggest you to go through my previous lecture.


Last year I floated this year also that is running that is on exploring handling large scale data set that is all also of aid credit. You will get detailed guidance and their practical applications as well.

(Refer Slide Time: 04:38)

### Binary Choice Models

- ❑ Consider a model:  
$$Y = \beta_0 + \beta_1 X + \epsilon$$
$$E(Y) = \beta_0 + \beta_1 X \quad E(\epsilon) = 0$$
- ❑ The response variable 'y' can take only two values say,
  - 1 if attribute is present
  - 0 if absent
- ❑ The regressand 'Y' is a **binary or dichotomous** variable.
- ❑ The binary variable 'Y' be defined by Bernoulli Distribution with  $P(y=1) = p$  and  $P(y=0) = 1-p$ .
- ❑ By the definition of the expected value of the random variable:  
$$E(Y) = 1(P_i) + 0(1-P_i) = P_i$$

The mean response is interpreted as the probability of success i.e  $Y=1$ , when the independent variable takes on the values  $X$ .



5

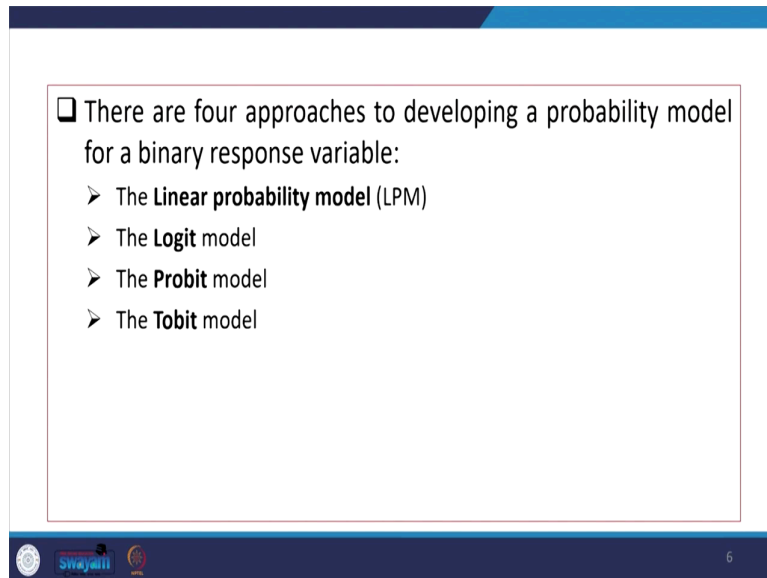
At this moment, I am just clarifying if it is a binary choice model the choice function is binary that is why if it is linearized with beta naught, beta 1 X plus epsilon the expected value of Y is in fact, defined as beta naught plus beta 1 X with the assumption that the estimated or expected value of the error term is equal to 0.

The response variable Y can take only two values for say binary as 1 if its attribute is present, else it is 0, that is, why the regression Y is called a binary or dichotomous variable which we just said. it has a probability with the distribution called the Bernoulli distribution.

When we have a binary response the distribution follows called Bernoulli distribution with p is equal to 1 that is small p and then p if it is equal if the dependent variable is with Y equal to 0 then other that is 1 minus p is the answer.

By the definition of the expected value of the random variable expected Y is, therefore, would be equal to 1 time Pi plus 0 that is 1; 0 0 times 1 minus Pi. So, 1 minus Pi is the probability of 0 as the occurrence Pi is the probability of one is the occurrence. Finally, the expected value of the dependent variable is nothing, but the probabilities. The mean response is interpreted as the probability of success that is Y equal to 1 when the independent variable takes on the values of x.

(Refer Slide Time: 06:07)



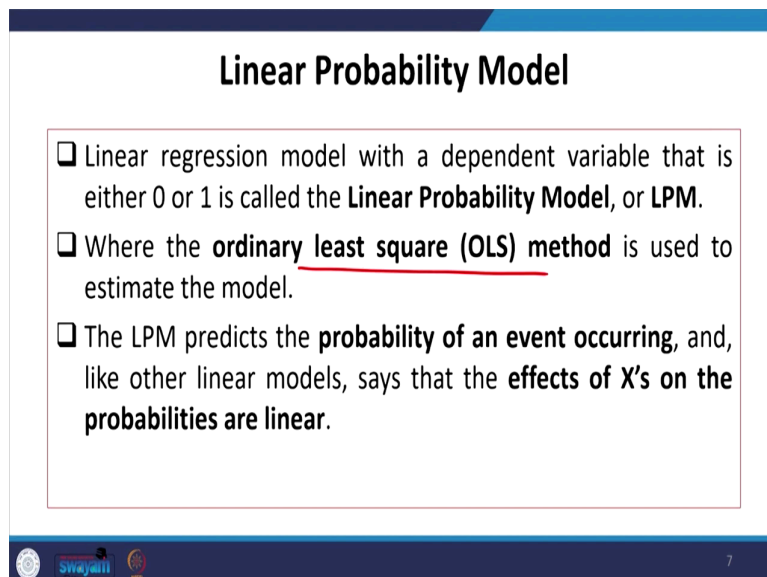
❑ There are four approaches to developing a probability model for a binary response variable:

- The **Linear probability model (LPM)**
- The **Logit** model
- The **Probit** model
- The **Tobit** model

6

There are four approaches to developing a probability model for the binary response kind of variables. Broadly they are discussed as LPM Linear Probability Model the logit model, the probit model, and Tobit model.

(Refer Slide Time: 06:25)



### Linear Probability Model

❑ Linear regression model with a dependent variable that is either 0 or 1 is called the **Linear Probability Model, or LPM**.

❑ Where the ordinary least square (OLS) method is used to estimate the model.

❑ The LPM predicts the **probability of an event occurring**, and, like other linear models, says that the **effects of X's on the probabilities are linear**.

7

Linear probability model LPM that you have to refer the book also to my previous module as I have already suggested it has some linearity in its responses where the ordinary least square method is actually applied. OLS is applied, the LPM predicts the probability of an event

occurring and like other linear models say that the effects of  $X_i$  on the probabilities are in fact linear.

(Refer Slide Time: 06:54)

□ Consider a linear regression model:  
 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ , Where  $Y = 1$  or  $Y = 0$ ,  
 Expected value of dependent variable  $E(Y_i)$  is conditional probability that individual has the attribute.

The Probability Distribution of $\varepsilon_i$			
$Y_i$	Probability	$P_i \cdot E(P_i)$	$\varepsilon_i$
1	$P_i$	$1 \cdot P_i$	$1 - \beta_0 - \beta_1 X_i$
0	$1 - P_i$	$0 \cdot (1 - P_i)$	$-\beta_0 - \beta_1 X_i$

Therefore,  $E(Y) = 1(P_i) + 0(1 - P_i) = P_i = E(Y_i = \beta_0 + \beta_1 X_i)$  for  $E(\varepsilon_i) = 0$

We are just explaining here to the model as we already explained I am not going to repeat I am just explaining the information given in the table. The probability distribution of the error term is equal to if Y is equal to 1 that is dependent variable is equal to 1 the probabilities are in fact, called P i. Otherwise if it is 0 the probability is 1 minus P i.

So what are the expected probabilities? Then one times P i if 1 is the occurrence in Y and 0 times of this. So, the error is basically 1 minus 1 minus the expected values. So, 1 minus that is the expected value we have already mentioned that this one 1 minus the expected value 1 Y 1 since Y i is 1.

So, 1 minus this 1 is the error term in the case of the one value of Y and otherwise 0 minus this. So, this is the expected value of the error term therefore, the expected Y is equal to 1 time P i plus 0 times 1 minus P i equal to P I, which is equal to Y i the expected value of this 1 for each of the expected value of the error term which is assumed to be 0.

(Refer Slide Time: 08:21)

### LPM Practical in STATA

- For this purpose, we are using a sample of NSS 75<sup>th</sup> round data on healthcare for the year 2017-18.
- We are interested in knowing what factors affects the nature of treatment(traditional vs modern) during ailment case.
- The dependent variable here is a dummy variable categorized in traditional treatment and modern treatment.
- A mixture of qualitative and quantitative variables are used as regressors.

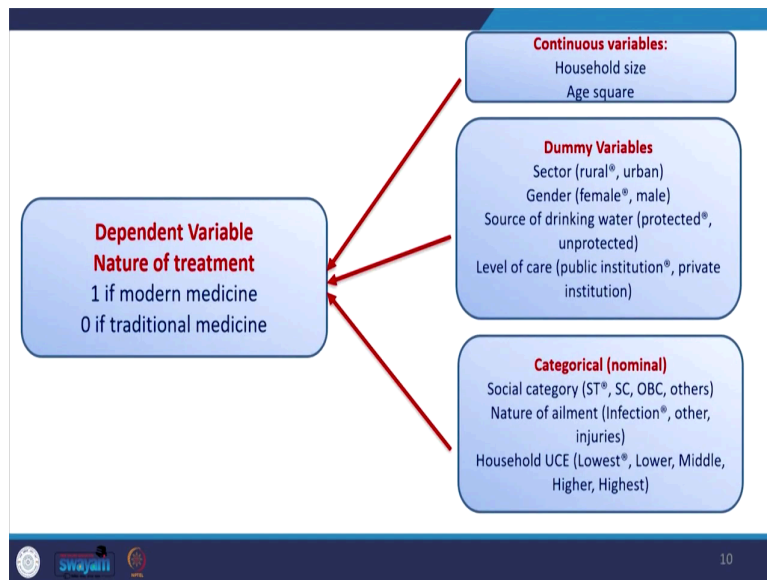
9

LPM and the practical in Stata were also explained in my previous model. for this purpose we are using a sample of NSS 75th round data on health care, this is the difference compared to my last year module on the previous module on handling large scale data using Stata.

This year we have included the NSS 75th round on health care data. We are interested in knowing what factors affect the nature of treatment that is traditional versus modern during the case of an ailment. If the ailment person is reporting access to some forms of medicines nonmedicinal treatment or nature of treatment those are traditional or modern.

we have categorized into two the dependent variable here is a dummy one that is yes or no. A mixture of qualitative and quantitative variables is used as the regressors.

(Refer Slide Time: 09:31)



Here are the dependent and independent variables. dependent variables we have just defined as 1 is for modern medicine, 0 is for traditional medicine. And the independent variable we have taken for explanation. We have already discussed independent variables in our regression model in the previous lecture. Here we have to consider some continuous variables, some dummy variables, and some categorical variables. You can read between these lines for your better understanding.

(Refer Slide Time: 09:58)

### Estimation of LPM using OLS *Regress*

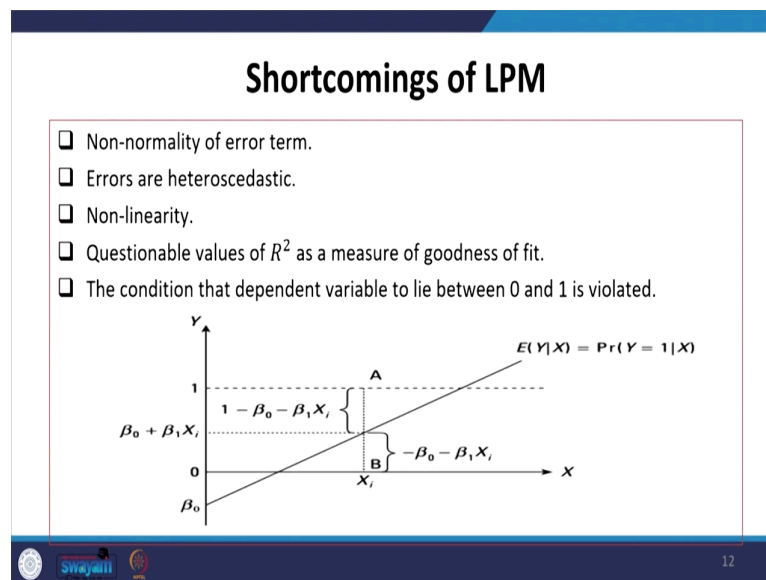
- ❑ Before going for data analysis, our first task is to understand the data.
- ❑ Examine descriptive statistics (As we did in previous lecture).
  - describe, summarize etc.
- ❑ *Regress* trad\_modern HouseholdSize square\_age i.Sector i.male\_female i.public\_privatehospital i.drinkingwater i.SocialGroup i.nature\_ailment i.UMPCE



the estimation of LPM using a simple OLS technique is taken. So, as we know when OLS is applied, when the OLS result is derived, simply apply the command called regress. So, simple regress with the first variable which is the dependent variable followed by the independent variable some i dot commands are given.

We have already explained that i dot commands are assigned or added just to identify whether that variable is in fact, categorical or not. Whether this has considered any base category for comparison and it impacts on the medicine on the choice of treatment. So, this is what is made we will also do our practical at the end. And I am just clarifying each of the concepts, and you can also run like these using the practice data set that is going to be uploaded.

(Refer Slide Time: 11:06)



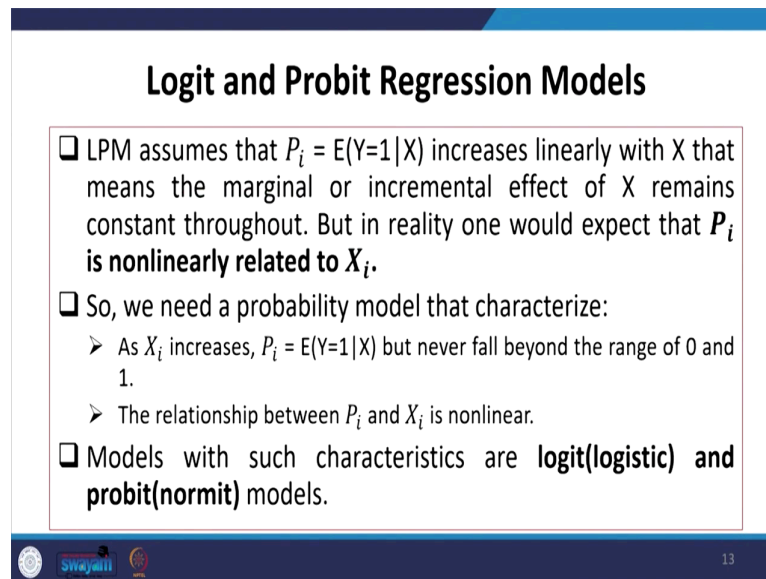
The shortcoming of this LPM are that the distribution that has been considered is not normal. Because it is following a Bernoulli distribution with 1 and 0, it is not normal. So, it is not avoiding not the errors are not estimated to be 0. So, errors are basically heteroscedastic that is important you may be asked the question if the distribution is not normal. Then what would be the error distribution? Does it have homoscedasticity or heteroscedasticity whether it violates or not?

Similarly, the non-linearity aspect is also identified since the data is not following a linear pattern, questionable values of R square, of course, because your distribution is not normal.

So, the goodness of fit in question the condition that the dependent variable to lie between 0 and 1 is in fact, violated.

So, the 0 and 1 that we have said may not be actually derived correctly. So the rest I have already explained in my previous model you can just follow.

(Refer Slide Time: 12:22)



**Logit and Probit Regression Models**

- ❑ 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(logistic)** and **probit(normit)** models.

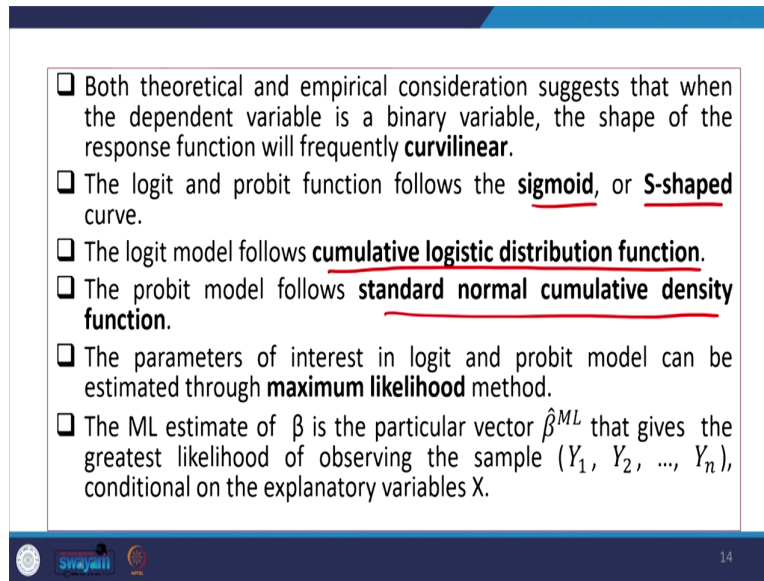
swajani 13

Now, coming to other two important and widely applied regression models in this context are called logit and probit regression models. LPM assumes with 1 and 0  $P_i$  is equal to the expected value of  $Y$  if it is 1 given the conditional values increase. When that  $P_i$  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 non-linearly related to the regressors.

So, we need a probability model that characterizes the  $X_i$  with the  $P_i$  values but never falls beyond the range of 0 and 1. The relations between  $P_i$  and  $X_i$  is non-linear models with such characteristics called logit and probit. If normal distribution in case of the normal distribution is followed.

A probit model is also called normit model, normit distribution model and here it is called logistic distribution models. Both theoretical and empirical considerations suggest that when the dependent variable is a binary one the shape of the response function will frequently be curvilinear. The logit and probit function follows the sigmoid curve.

(Refer Slide Time: 13:53)



- ❑ 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.
- ❑ The logit model follows **cumulative logistic distribution function**.
- ❑ The probit model follows **standard normal cumulative density function**.
- ❑ The parameters of interest in logit and probit model can be estimated through **maximum likelihood** method.
- ❑ The ML estimate of  $\beta$  is the particular vector  $\hat{\beta}^{ML}$  that gives the greatest likelihood of observing the sample  $(Y_1, Y_2, \dots, Y_n)$ , conditional on the explanatory variables  $X$ .

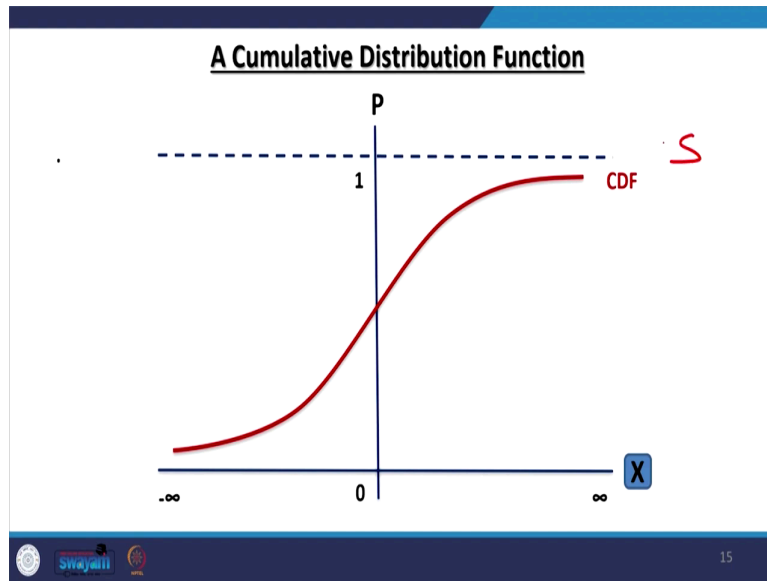
Sigmoid you might be asked a question that which distribution it follows. In LPM we have already said that that is following a Bernoulli distribution, here it is s-shaped as a sigmoid curve the logit model follows a cumulative logistic distribution function. Again the curve that is explained is called the cumulative logistic distribution function.

The probit model follows the standard normal cumulative density function which is the difference between logit and probit. Standard normal, cumulative density function where in case of logit it is called cumulative logistic distribution function. The parameters of interest in logit and probit can be estimated through the maximum likelihood estimation method.

I think you might have heard about different estimators one is called the ordinary least square method. Another method is called the maximum likelihood method.

What is the likelihood of being included in that particular model? The maximum likelihood estimator of beta is the particular vector that is also written as estimated value beta hat ML that gives the greatest likelihood of observing the sample conditional on the explanatory variables.

(Refer Slide Time: 15:17)



(Refer Slide Time: 15:27)

### Logit Regression Model

**Assumptions :**

- Logistic regression **does not require a linear relationship** between the dependent and independent variables.
- The **residuals do not need to be normally distributed.**
- Logistic regression requires the **observations to be independent of each other.** It should not come from repeated measurement or matched data.
- It assumes that the **independent variables are linearly related to the log odds.**
- Logistic regression typically requires a **large sample size.**

This is what it looks like this is a S-shaped curve called a sigmoid curve. This is a kind of logistic map the logistic model have certain assumptions that it does not require a linear relationship between the dependent and independent variable. That is why we said if OLS is violated then we may go for some form of non in the non-linear model.

But again these are called transform linear models. These are also called generalized forms of model they are also called GLM. In some cases of GLM we will explain these to some extent

in one of the modules, In one of the lectures the residuals do not need to be normally distributed.

It is a distribution is s-shaped this is sigmoid kind. So, not necessarily following a normally distributed one this is quite important to note. Logistic regression requires the observations to be independent of each other. It should not come from repeated measurement or mass match data this assumes that the independent variables are linear related to the log of odds.

If the odds are taken as a log; that means, the distribution comes out to be linear. So, we will show that also in our derivation. Logistic regression typically requires large sample data that has to be understood.

(Refer Slide Time: 16:42)

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

$$P_i = \text{Prob}(Y_i=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} = \frac{1}{1 + e^{-Z_i}} = \frac{e^{Z_i}}{1 + e^{Z_i}}, \text{ where } Z_i = \beta_0 + \beta_1 X_i$$

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

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

It is visible from the function that as  $Z_i$  ranges from  $-\infty$  to  $\infty$ ,  $P_i$  ranges between 0 and 1.

Logit, the log of odd ratio, is not linear in  $X$ , but also linear in the parameters

Linear estimators (OLS) are not applicable.

➤ As if  $Y=1$  then  $\ln\left(\frac{1}{0}\right)$  and if  $Y=0$  then  $\ln\left(\frac{0}{1}\right)$  are both meaningless.

Here is the model when we take the log of this one basically you just go through this I have explained very clearly in my previous module on handling large-scale data, this is the logistic function. It has the expected value the exponential values, X exponential opens the standard normal distribution, but it is exponentiated.

So, then exponentiated then another one actually the function is, in fact, presented like this. And it is probable and it is odd of the probabilities or 1 minus P i is represented here. Finally, the logit is actually called the transformation of the function of P i to that of 1 minus P i this is also called the odd ratio.

This is when we take a log of all those things we are finally, getting a linear function. So, the log of the odd is nothing, but the OLS nothing, but in fact, the linear regression model. So, it is visible from the function that the  $Z_i$  is ranging from minus infinite to plus infinite that you can just read between the line.

So, logit that is basically a log of odd ratio is not linear in  $X$ , but also linear in the parameter the way we discuss in the assumption of OLS regression. So, linear with the parameter and non-linear with the variables. So, the ordinary least square method is not applied when you have this kind of function.

(Refer Slide Time: 18:32)

### Marginal Effects in Logit Model

- ❑ The coefficients of the logit function is quite difficult to interpret since it follows a logistic distribution function.
- ❑ As a results we compute the **marginal effects**.
- ❑ Marginal effects gives the derivative of the probability that the dependent variable equals one with respect to a particular conditioning variable.

$$\frac{\partial P(Y=1)}{\partial X} = \frac{\partial F(\hat{\beta}_0 + \hat{\beta}_1 X)}{\partial X} \hat{\beta}_1 = \frac{\hat{\beta}_1 e^{-(\hat{\beta}_0 + \hat{\beta}_1 X)}}{[1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 X)}]^2}$$

The derivative is non-linear and depends on the value of  $X$ .

whenever you derive a logit model it is always suggested to derive marginal effects because logit is not have with continuous variables. by ordinal least square regression we get all marginal changes marginal effect of each conditional variable or control variables on it is the dependent variable, but since in logit, it is a sporadic change from 1 to 0.

It follows a path different than that of a continuous since the dependent variable is not continuous. We are supposed to further take the marginal effect. So, the marginal effect basically the derivative again is taken like this. So,  $dP$  by  $dX$  and you can just follow between the lines and how it is interpreted.

(Refer Slide Time: 19:27)

## Probit Regression Model

- ❑ Assumptions :
  - Probit model **does not require a linear relationship** between the dependent and independent variables.
  - In the probit model, the **error term is normally distributed.**
- ❑ We assume that the model takes form:  
$$P(Y = 1 | X) = F(X\beta)$$
  - P denotes probability.
  - F denotes cumulative distribution function (CDF) of the standard normal distribution.
  - Parameters  $\beta$ 's are typically estimated by maximum likelihood method.

19

The next one is called the probit regression model, probit regression model. This does not require a linear relationship between the dependent and independent variables probit model the error term is normally distributed that is most important. But in the case of logit, it is not normally distributed.

the probability function is Y is equal to 1 given X. So, in sort, it is this cumulative frequency are written as X beta rest of the detail the is given here. This is the distribution follows as CDF Cumulative Distribution Function.

(Refer Slide Time: 20:08)

□  $Y^* = X\beta + \varepsilon$   
Where  $Y^*$  is unobserved and  $\varepsilon \sim N(0, \sigma^2)$

□ The  $Y$  can be viewed as an indicator for whether this latent variable is positive:  
 $Y = 1$ , if  $Y^* > 0 = X\beta + \varepsilon > 0$  i.e.  $\varepsilon > -X\beta$   
0 otherwise.

$P(Y^* > 0 | X) = P(Y = 1 | X) = P(\varepsilon > -X\beta)$   
 $= P\left(\frac{\varepsilon}{\sigma} > \frac{-X\beta}{\sigma}\right)$

swayam 20

And the probabilities, in this case, are defined we have already said that it should follow it is error distribution to be normally distributed. And the probability is having greater than 0 values, given different  $Y$  is greater than 0 values given different control variables.

It suggested that  $p$  is equal to  $Y$  is equal to 1 given  $X$  should equal to the probability of the error distribution and the error distribution should be at least greater than that of it is estimated values. Therefore, this is how it is defined you can just follow it therefore, the probabilities will be presented with this range.



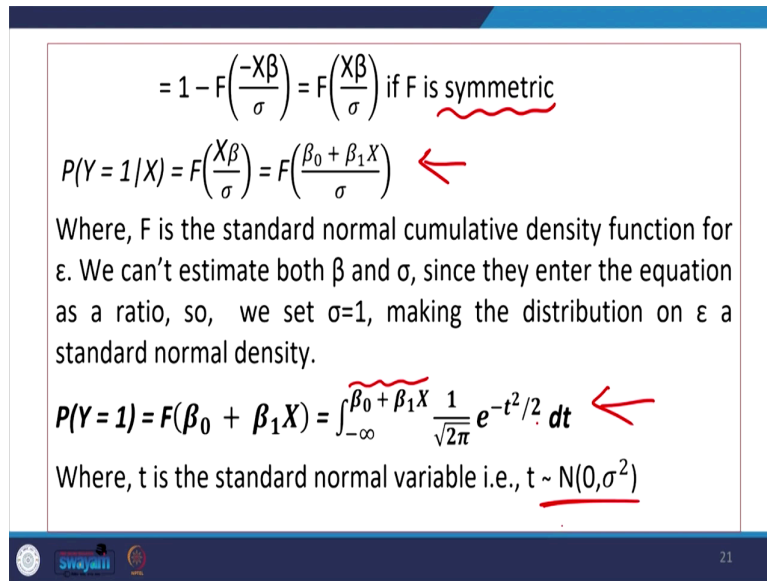
(Refer Slide Time: 20:58)

$$= 1 - F\left(\frac{-X\beta}{\sigma}\right) = F\left(\frac{X\beta}{\sigma}\right) \text{ if } F \text{ is symmetric}$$
$$P(Y = 1|X) = F\left(\frac{X\beta}{\sigma}\right) = F\left(\frac{\beta_0 + \beta_1 X}{\sigma}\right)$$

Where, F is the standard normal cumulative density function for  $\epsilon$ . We can't estimate both  $\beta$  and  $\sigma$ , since they enter the equation as a ratio, so, we set  $\sigma=1$ , making the distribution on  $\epsilon$  a standard normal density.

$$P(Y = 1) = F(\beta_0 + \beta_1 X) = \int_{-\infty}^{\beta_0 + \beta_1 X} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Where, t is the standard normal variable i.e.,  $t \sim N(0, \sigma^2)$

The slide contains mathematical derivations and text. The first equation shows the symmetry property of the standard normal cumulative distribution function. The second equation shows the probability of Y=1 given X, which is equal to the cumulative distribution function of a normal distribution with mean beta\_0 + beta\_1 X and standard deviation sigma. The text explains that sigma is set to 1 for estimation purposes. The third equation shows the integral form of the cumulative distribution function. The text below explains that t is a standard normal variable. There are red wavy lines under the 'if F is symmetric' text and the 't ~ N(0, sigma^2)' text. There are also red arrows pointing to the second and third equations.

Now, accordingly, we arrive at the density function you can just see once how density function is defined it is basically defined as symmetry. The, therefore, the error distribution is defined to be normal. So, it follows the normal distribution with these functions.


We know that normal distribution that varies from minus infinity to the estimated value of the beta one of 1 upon square root of 2 pi times it is its exponential value of the Z value or t values here with it is marginal changes. So, t is the standard normal variable with the function is varies from 0 with 0 mean and sigma square standard deviation or sigma square as the variance.

(Refer Slide Time: 21:47)

□ Now we can obtain information on probability  $P(Y=1)$  and  $\beta$ 's:

$$F^{-1}(P) = \beta_0 + \beta_1 X$$

The inverse transformation is called the **probit**, which gives the linear predictor as a function of the probability.



22

Now, we can obtain the information and probabilities rest you just can follow I will come to the practical aspect of it. Similarly, since it has probabilities and it is extreme values with 1 and 0 it is always suggested to go for marginal effect.


(Refer Slide Time: 22:09)

### Marginal Effects of the Probit Model

□ The drawback of probit model is that the coefficients are more difficult to interpret, hence they are less used.

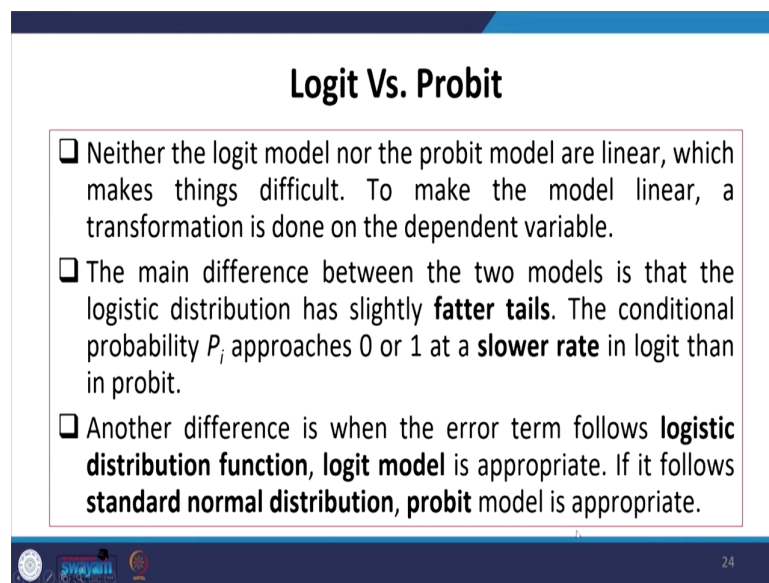
□ The marginal effect of changing  $X$  on  $\hat{P}$ , the probability of getting  $Y = 1$ .

□ For the probit model,

$$P(Y = 1|X) = F(\hat{\beta}_0 + \hat{\beta}_1 X)$$
$$\frac{\partial P(Y = 1|X)}{\partial X} = F(\hat{\beta}_0 + \hat{\beta}_1 X) \hat{\beta}_1$$


23

(Refer Slide Time: 22:13)



### Logit Vs. Probit

- ❑ Neither the logit model nor the probit model are linear, which makes things difficult. To make the model linear, a transformation is done on the dependent variable.
- ❑ The main difference between the two models is that the logistic distribution has slightly **fatter tails**. The conditional probability  $P_i$  approaches 0 or 1 at a **slower rate** in logit than in probit.
- ❑ Another difference is when the error term follows **logistic distribution function**, **logit model** is appropriate. If it follows **standard normal distribution**, **probit model** is appropriate.

24

Again we take the derivative of the probabilities function with respect to it is X and accordingly, it is interpreted. Now it is important to know the difference between logit and probit. Neither the logit model nor the probit model is linear which makes things difficult to make the model linear transformations are taken usually log transformation is taken.

The main difference between the two models is that logit distribution has slightly fatter tails the tails are fatter. Whereas, in the case of the conditional probability P the conditional probabilities  $P_i$  approaches if 0 or 1 at a slower rate in logit. So, in the probit model, we will have fatter tails, but in the case of logit, it has a slower rate.

(Refer Slide Time: 23:06)

### Caution for Comparing Models

- ❑ Though the models (LPM, logit and probit) do the same thing (estimating probabilities), one has to be careful in interpreting the coefficients estimated.
- ❑ The coefficients are not directly comparable.
- ❑ But there are certain ways to compare the models. By multiplying coefficients by certain value:

$$\begin{array}{ll} \beta_{logit} = 4 \beta_{LPM} & \beta_{probit} = 0.625 \beta_{logit} \\ \beta_{LPM} = 0.25 \beta_{logit} & \beta_{probit} = 2.5 \beta_{LPM} \\ \beta_{logit} = 1.6 \beta_{probit} & \end{array}$$

25

No, I think I have said it a little differently I will show it here maybe I think in my previous model I have said let me show you once again yes. The logistic distribution would have the tail should be fatter. Whereas, in case of probit, it is more is less fat.

Another difference is the is when the error term follows the logistic distribution function logit is more appropriate. Whereas, when there are distribution is normal or normally distributed cumulative density function is normal then probit is most appropriate. These two points are quite important.

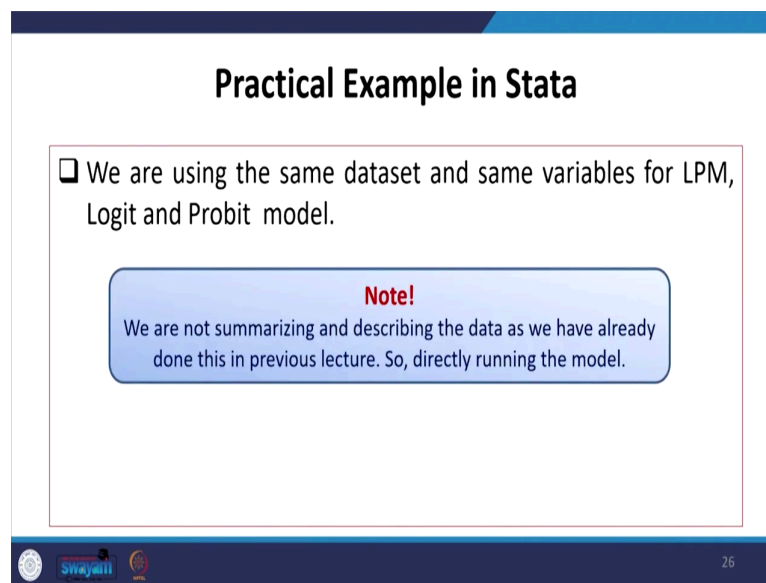
Now, some questions for comparing models so I have said in earlier models how these are estimated, what are the comparison points beta value in case of logit is 4 times the beta value of the LPM model linear probability model? Beta value in probit is 0.625 times of it is logit similarly other values you can just compare and for your record, these are already estimated and made the comparison.

(Refer Slide Time: 24:17)

### Practical Example in Stata

- ❑ We are using the same dataset and same variables for LPM, Logit and Probit model.

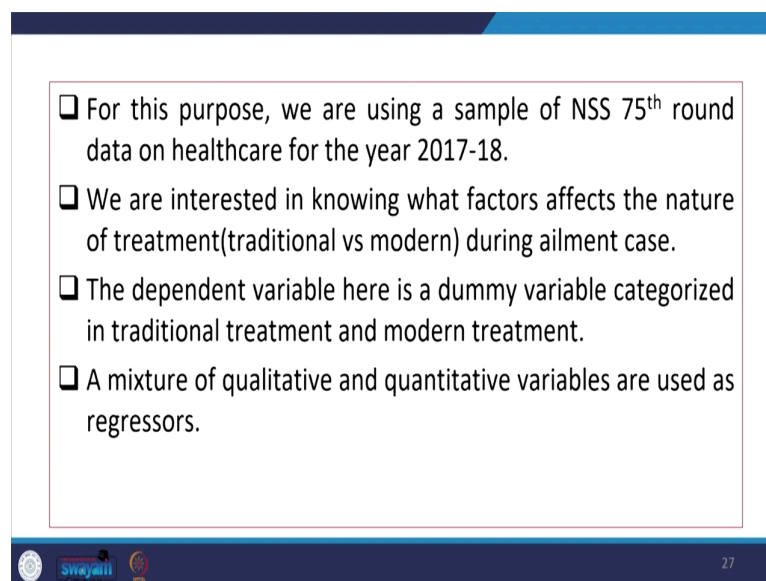
**Note!**  
We are not summarizing and describing the data as we have already done this in previous lecture. So, directly running the model.



26

(Refer Slide Time: 24:24)

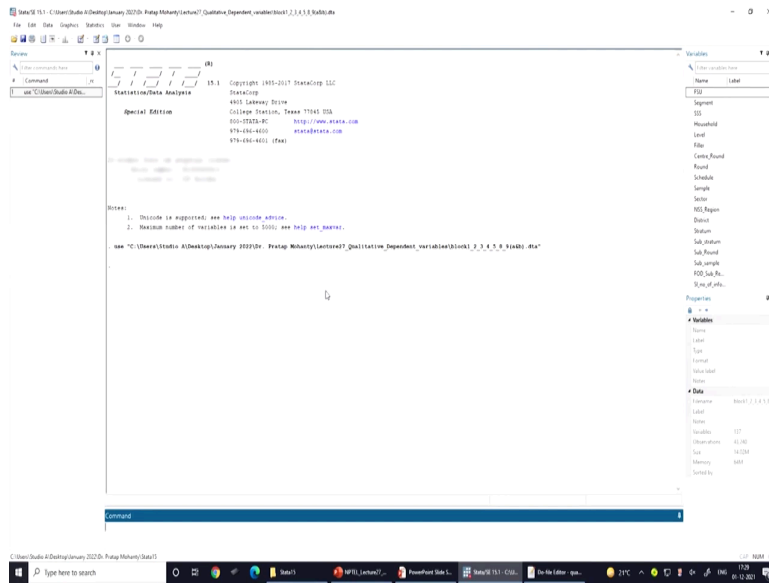
- ❑ For this purpose, we are using a sample of NSS 75<sup>th</sup> round data on healthcare for the year 2017-18.
- ❑ We are interested in knowing what factors affects the nature of treatment(traditional vs modern) during ailment case.
- ❑ The dependent variable here is a dummy variable categorized in traditional treatment and modern treatment.
- ❑ A mixture of qualitative and quantitative variables are used as regressors.



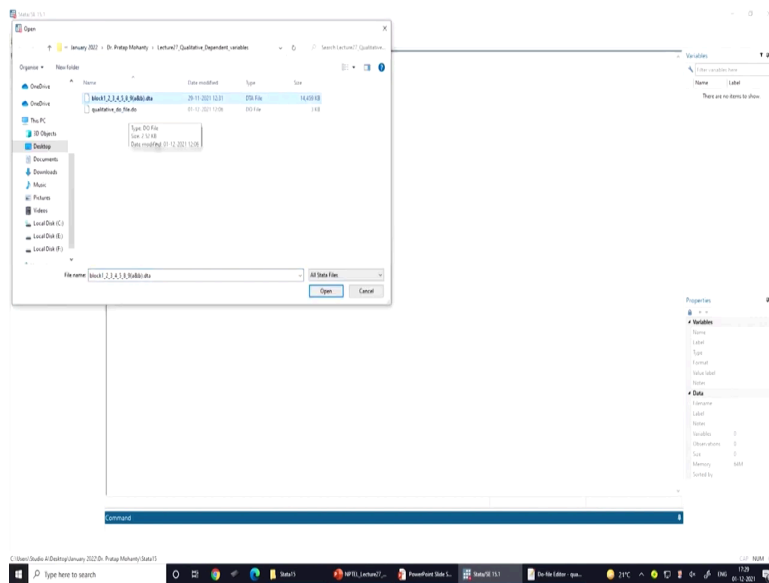
27

Now, we are going to operate with our practical example that we have taken the NSS 75th round data and we will also explain you.

(Refer Slide Time: 24:29)



(Refer Slide Time: 24:36)



Here is your stata we are going to open the data on your screen. The sample data that we are also going to share with you these are the MOS data 1, 2, 3, 4, 5, 8, 9 of NSS 75th already the fine tuned data after merging.

(Refer Slide Time: 24:48)

```
1 regress treat_model Kwambhiddile square_age i.Sector i.male_female i.public_privatehospital i.dischargepayer i.SocialGroup i.nature_aidment i.DMPC
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

Now, we will also open the do file. We have made just for your reference you can we have already discussed several times about how to develop a do file and how it helps a lot in running the model. We will first regress. Why regress is required?.

Because if you regress with the limited dependent variable that is dependent variable is categorical the result we are deriving is basically called LPM first one we will regress the LPM result on your screen. Therefore, that is simply taking the command as we take in wireless regression.

(Refer Slide Time: 25:27)

Stata 15.1 - C:\Users\Student\Desktop\January 2022\Stata\Pragya Mahapatra\January2022\Quantitative\level\Bmschick\_2\_3\_4\_5\_8\Bmschick.dta

```
use "C:\Users\Student\Desktop\January 2022\Stata\Pragya Mahapatra\January2022\Quantitative\level\Bmschick_2_3_4_5_8\Bmschick.dta"
do "C:\Users\Student\Desktop\January 2022\Stata\Pragya Mahapatra\January2022\Quantitative\level\Bmschick_2_3_4_5_8\Bmschick.dta"

logit treat_anders Bmschick15sqr equate_age i.Sector i.male_female i.public_privateshospitals i.drinkingwater i.SocialGroup i.nature_alltime i.DMPCE
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Bmschick15sqr	.0004775	.0004387	1.08	0.278	-.0003864	.0013389
equate_age	3.08e-06	5.42e-07	5.49	0.000	1.86e-06	4.18e-06
Sector						
urban	.0037407	.0031404	1.19	0.231	-.0026416	.0089717
male_female						
female	-.0027620	.001961	-1.40	0.000	-.0056601	-.0018639
public_privateshospitals						
private_hospitals	-.0168055	.0020643	-8.08	0.000	-.0209454	-.0126655
drinkingwater						
unprotected_water	-.0115778	.0026686	-4.34	0.000	-.0168244	-.0063312
SocialGroup						
SC	.0047385	.0047029	1.00	0.317	-.0042879	.0134849
DMC	.0048436	.0043049	1.12	0.261	-.0032711	.0128854
Other	.0043031	.0046124	0.93	0.347	-.0047457	.0144320
nature_alltime						
river	-.0023614	.0023829	-0.99	0.000	-.0045387	-.0011841
vegetation	-.0028427	.0025023	-1.13	0.000	-.0053664	-.0013190
DMPCE						
1	.0047704	.0030504	1.56	0.119	-.0012963	.0128314
2	.0034495	.0033269	1.03	0.302	-.0034842	.0127460
3	.0028496	.0032624	0.87	0.382	-.0036264	.0120742

Here on the screen this is the one. This gives the result of the LPM, since our dependent variable is limited. Now we are again explaining another one logit. A logit, but another aspect is we are given conditional logit conditional logit Xi command is given on the next do file command. We have mentioned is Xi, Xi command is basically giving you some conditional information about it is about it is reference categories.

(Refer Slide Time: 26:10)

Stata 15.1 - C:\Users\Student\Desktop\January 2022\Stata\Pragya Mahapatra\January2022\Quantitative\level\Bmschick\_2\_3\_4\_5\_8\Bmschick.dta

```
xi logit treat_anders Bmschick15sqr equate_age i.Sector i.male_female i.public_privateshospitals i.drinkingwater i.SocialGroup i.nature_alltime i.DMPCE
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Bmschick15sqr	.0149330	.0137504	1.07	0.284	-.0106111	.0394974
equate_age	.0004975	.0004549	1.09	0.000	.0000556	.0009393
Sector						
urban	.0037407	.0031404	1.19	0.231	-.0026416	.0089717
male_female						
female	-.0027620	.001961	-1.40	0.000	-.0056601	-.0018639
public_privateshospitals						
private_hospitals	-.0168055	.0020643	-8.08	0.000	-.0209454	-.0126655
drinkingwater						
unprotected_water	-.0115778	.0026686	-4.34	0.000	-.0168244	-.0063312
SocialGroup						
SC	.0047385	.0047029	1.00	0.317	-.0042879	.0134849
DMC	.0048436	.0043049	1.12	0.261	-.0032711	.0128854
Other	.0043031	.0046124	0.93	0.347	-.0047457	.0144320
nature_alltime						
river	-.0023614	.0023829	-0.99	0.000	-.0045387	-.0011841
vegetation	-.0028427	.0025023	-1.13	0.000	-.0053664	-.0013190
DMPCE						
1	.0047704	.0030504	1.56	0.119	-.0012963	.0128314
2	.0034495	.0033269	1.03	0.302	-.0034842	.0127460
3	.0028496	.0032624	0.87	0.382	-.0036264	.0120742



Now we are just running it over here the result is on your screen now this is derived it is here. We have derived the logit with it is conditional variables like for example, the conditional you can just read here sector on 2. One is rural and two is urban what is the interpretation of it the coefficient is given not the odd ratio you should note it very clearly. These are not odd ratio these are simply called coefficient and it is significance value is given as P values, P values column are there.

It says that the probability of the likelihood of the patients or the persons with element located in urban areas are less are more likely to access modern medicines. The 0 and 1 we have taken, the 0 and 1 in our dependent variable traditional versus modern.

Similarly, you can also follow others. Second one is female as compared to male, then public versus private etcetera rest of the details you can find from our paper their paper link we already shared to you earlier. Now, we are also trying to give logit estimation with it is simple this is simple logistic regression equation is it is you can also derive and find out here this is result is displayed.

(Refer Slide Time: 27:44)

	coef	std. err.	z	P >  z	[95% Conf. Interval]
-----+-----					
Constant	1.103944	.0277756	3.71	0.000	.9899463 1.2512266
female	-.0257616	.0258793	-0.99	0.000	-.0809559 .0304250
public_private	-.0013756	.0298229	-0.46	0.000	-.0552869 .0525351
urban	-.0257616	.0258793	-0.99	0.000	-.0809559 .0304250
socialclass					
SC	1.209447	.1547937	7.82	0.000	.8976839 1.5586104
OC	1.132076	.1320260	8.58	0.000	.8698750 1.4044770
Other	1.199719	.1961385	6.12	0.000	.8017759 1.5976621
patient_attended					
other	-.5405758	.0489764	-11.05	0.000	-.6370344 -.4441172
attended	-.5398676	.0797970	-6.77	0.000	-.6885159 -.3912193
ORFID					
1	1.128700	.0260263	4.34	0.000	1.0764461 1.1839855
2	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
3	1.089175	.0602059	18.09	0.000	1.0691527 1.1094000
4	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
5	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
6	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
7	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
8	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
9	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
10	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
11	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
12	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
13	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
14	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
15	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
16	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
17	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
18	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
19	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
20	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
21	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
22	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
23	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
24	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
25	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
26	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
27	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
28	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
29	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
30	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
31	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
32	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
33	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
34	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
35	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
36	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
37	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
38	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
39	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
40	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
41	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
42	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
43	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
44	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
45	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
46	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
47	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
48	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
49	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508
50	1.101971	.1306429	8.43	0.000	1.1174330 1.4852508

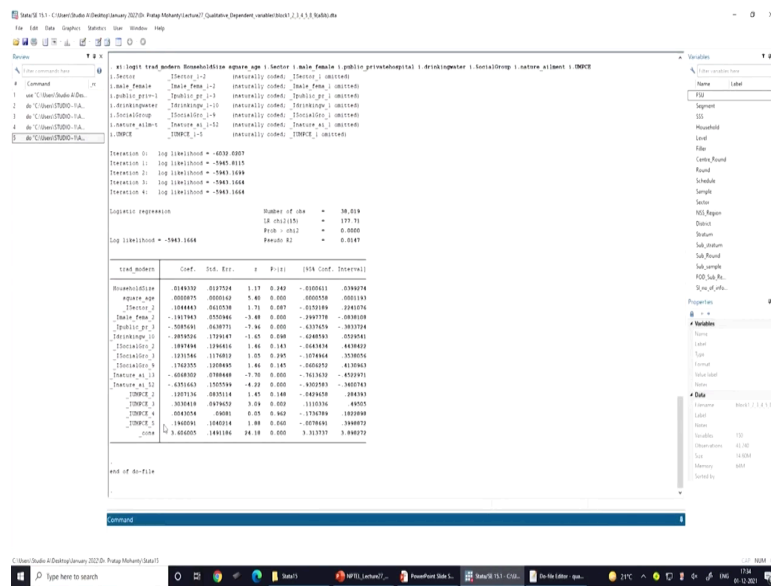
So, basically simple one it takes the average without having like look at the here. When you have let me tell you the command here logistic regression, when you give logit it gives you the coefficient, when you take logistic command it gives you the odd ratio.

So, here is the odd ratio on the first column of the result is your odd ratio. So, odd ratio if it is greater than 1; that means, your result is in favor if it is less than 1 it is against. So, like P and 1 minus P and whether they are significant or not you can interpret accordingly. So, then we will have margin command as well.

So, margin at it means after running the logit model when you run logit command you are getting coefficient not the odd ratio. But when you have applied the command with logistic you have got the odd ratio in case of odd ratio no margin effect is required.

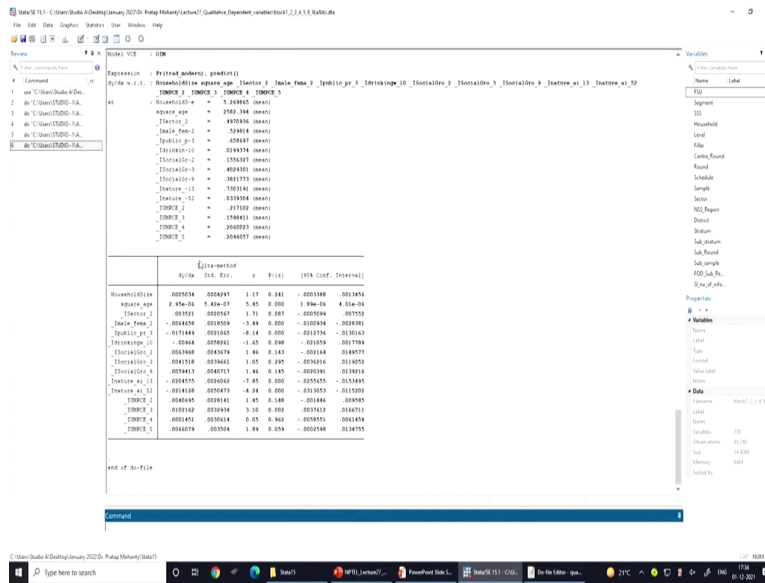
If you have got the value as coefficient you have to run the margin command that is has to be interpreted better. So, once again we will run the logit and show you followed by it is margin effect X i command. Once we have taken Xi logit. So, we are just operating it over here that I can show it to you how it works.

(Refer Slide Time: 29:22)



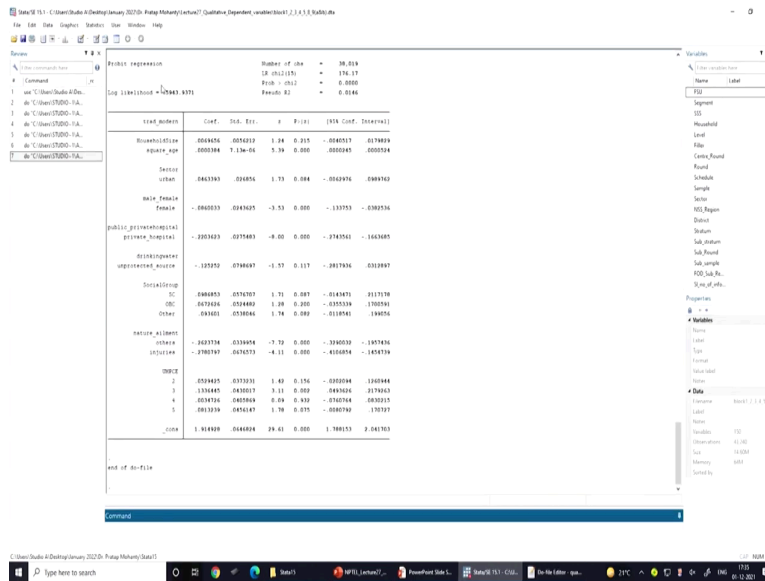
So, this is in fact, our coefficient values this is what is the logit command now followed by that we will run the margin command margin with dY by dX and at it means. So, it means we are deriving the result has already been derived.

(Refer Slide Time: 29:42)



So, this  $dY$  by  $dX$  the first order derivative and its value and its significance level is derived. but the interpretation are similar to that of the coefficient, but these gives most appropriate values than that of the logit coefficients. So, margin effects has to be derived alright. Then similarly you can run the probit model probit the probit coefficient is here probit and it is dependent and independent variable. Here are the results.

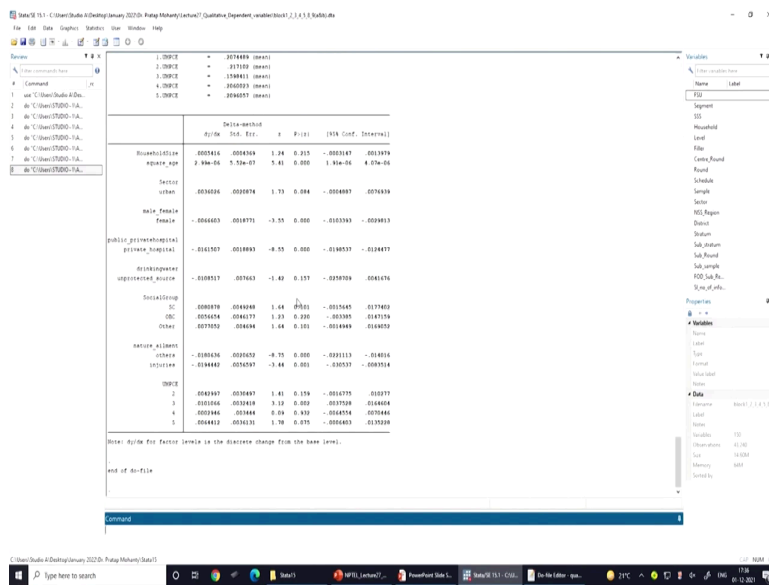
(Refer Slide Time: 30:14)



Now probit regression on the top is written the in the probit regression when your distribution is symmetric your distribution cumulative distribution function is following normal distribution or error distribution normal then probit is most applied.

And accordingly, we have derived the coefficient when you have got the coefficient I already suggested to you that you are supposed to run marginal margin probit as well. So, margins we are also deriving here alright when you have got the margin values then you can check over here. I think we have not, we will run it once again and the result is on your screen.

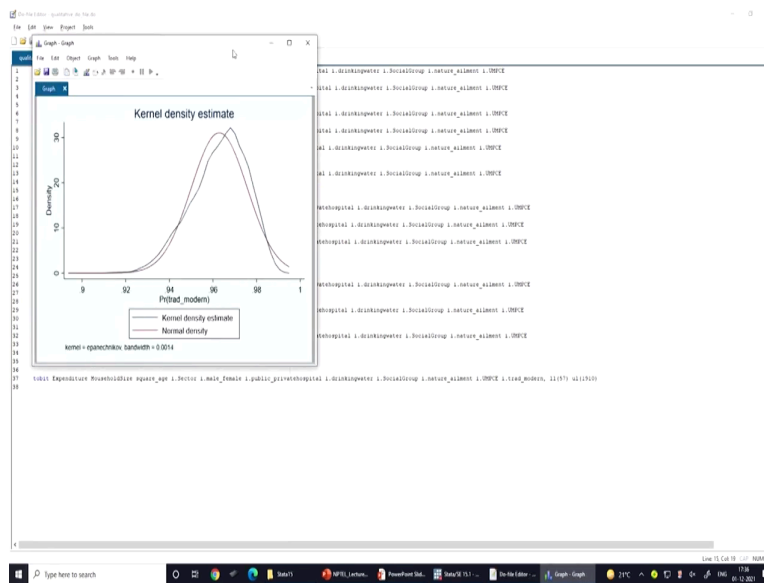
(Refer Slide Time: 31:01)



So, margin at means is already derived and another interesting part in probit is that you can predict since you have probabilities. So, prediction of the variables are possible you can predict for a different time. So, first we will draw the logit probit regression. Then we will predict then that prediction will give us whether the distribution is actually normal or not normal.

So, we run the probit once again then we will run the predictor. Predict of the residuals are stands for the residuals or the error term. Now, we will see whether they are actually distributed normally or not; k density function is going to give us result over here.

(Refer Slide Time: 31:47)

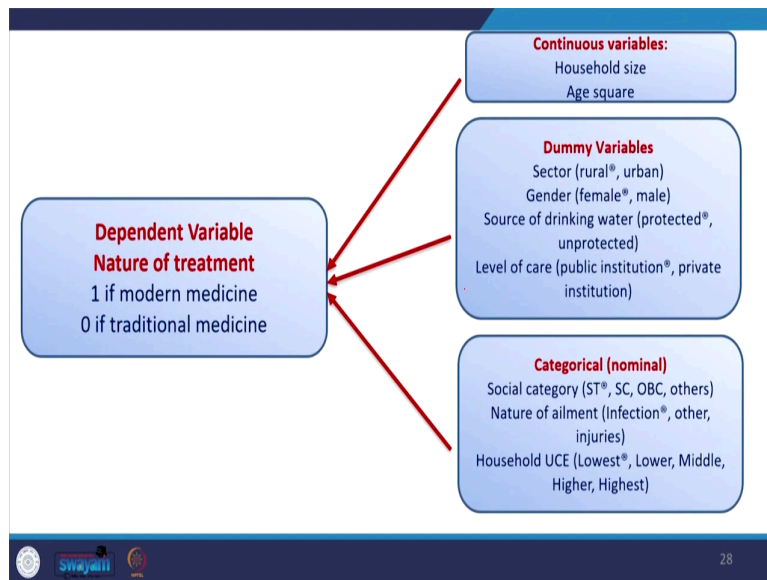


So, here this same session it is almost assimilating with the normal trend. So, in that case probit model could be applied. If it is much deviated and more asymmetric then in that case you need not apply a probit rather logit is more suitable.

So, other aspects are like how to compare all those things like comparing logit, then probit, then your LPM together. How these values which we are already shown what proportion of logit coefficient is comparable to the coefficient of LPM. These all rest of the commands are actually going to help you because we can store all the commands together and compare their values.

So, at this moment I am not running I am just keeping it with you I am sure that you will be enjoying and running those things. And my if you are having some difficulties I will suggest that you refer my previous NPTEL module that is still this time it is also being floated that is on handling large scale large handling large scale data with stata. The videos are available in YouTube as well all the videos with that title you can search and find out the details. So, I am not going to explain all those things those are care for your interest.

(Refer Slide Time: 33:15)



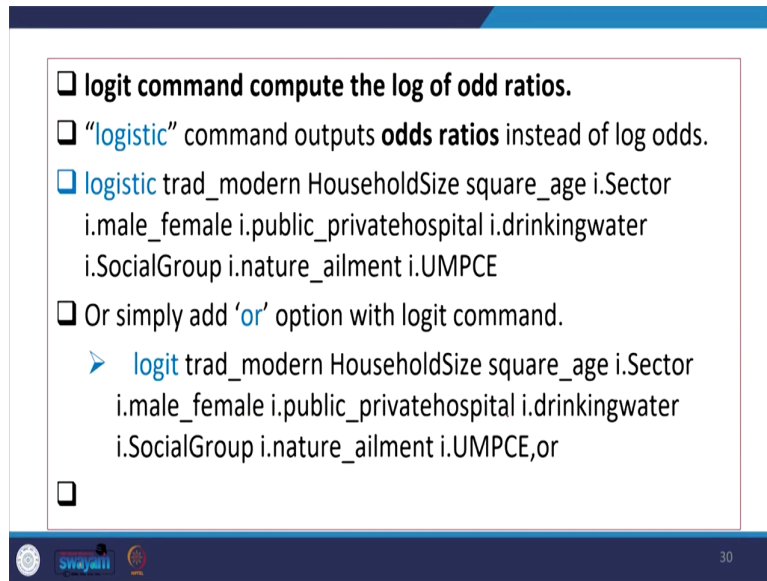
(Refer Slide Time: 33:17)

## Logit Model

- ❑ `logit/xi:logit trad_modern HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE`
- ❑ The `i.` before some variables denotes that these variables are **dummy** or **categorical variables**, and that it should be included in the model as a **series of indicator variables**.
- ❑ If you don't specify stata which category to take as reference, by default it takes first category as a reference category.
- ❑ If you want to change the reference category. For example, if a variable has three categories and you want second category as reference, you can specify it while running command:
  - `logit trad_modern ib2. SocialGroup`

At the bottom of the slide, there are logos for Swajati and a page number 29.

(Refer Slide Time: 33:24)



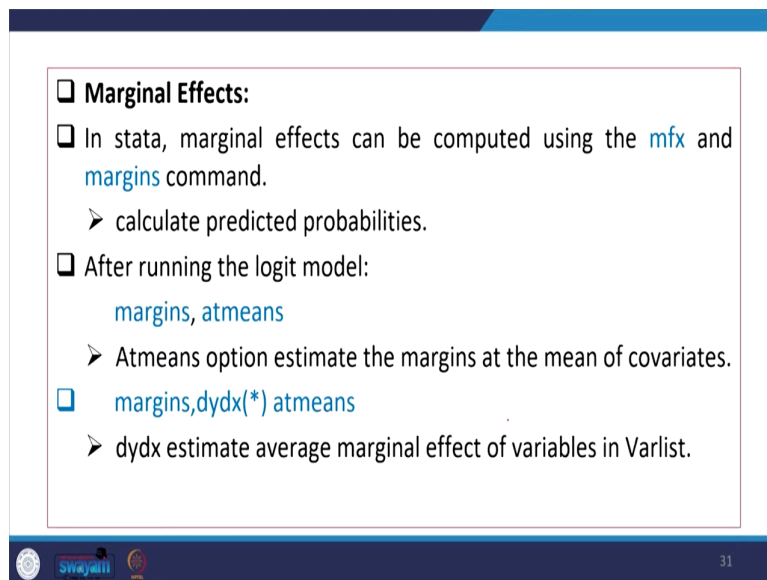
Slide 30 content:

- ❑ **logit** command compute the log of odd ratios.
- ❑ “**logistic**” command outputs **odds ratios** instead of log odds.
- ❑ **logistic** trad\_modern HouseholdSize square\_age i.Sector i.male\_female i.public\_privatehospital i.drinkingwater i.SocialGroup i.nature\_ailment i.UMPCE
- ❑ Or simply add ‘**or**’ option with logit command.
  - **logit** trad\_modern HouseholdSize square\_age i.Sector i.male\_female i.public\_privatehospital i.drinkingwater i.SocialGroup i.nature\_ailment i.UMPCE,or
- ❑

Slide footer: Swajati 30

So, we have all explained all those things including mixture of qualitative and quantitative variables. Logit model with these commands we have already discussed. I am not discussing you need to go through on your own.

(Refer Slide Time: 33:28)



Slide 31 content:




- ❑ **Marginal Effects:**
- ❑ In stata, marginal effects can be computed using the **mf** and **margins** command.
  - calculate predicted probabilities.
- ❑ After running the logit model:
  - margins, atmeans**
  - Atmeans option estimate the margins at the mean of covariates.
- ❑ **margins, dydx(\*) atmeans**
  - dydx estimate average marginal effect of variables in Varlist.

Slide footer: Swajati 31

(Refer Slide Time: 33:31)

## Probit Model




- ❑ `probit` trad\_modern HouseholdSize square\_age i.Sector i.male\_female i.public\_privatehospital i.drinkingwater i.SocialGroup i.nature\_ailment i.UMPCE
- ❑ The basic `probit` commands report **coefficient estimates** and the **underlying standard errors**. These coefficients are the **index coefficients** and what we can only say is the **direction of the effect** and **partial effects** on the Probit index/score. They do not correspond to the average partial effects.



32

(Refer Slide Time: 33:33)

- ❑ We can only check for the sign (whether positively or negatively related) and significance of the coefficient.
- ❑ In analysing binary choice models the parameter of interest are not the index coefficients, rather the **marginal/ partial effects**.
- ❑ The command for this task is `margins`.
- ❑ First run the probit model then:
  - `margins, dydx(*) atmeans`



33




(Refer Slide Time: 33:35)

- ❑ The choice of logit and probit model is on the researcher. But there exist one difference between logit and probit model that is, when the residual follows the **logistic distribution** we apply **logit model** and when the residual follows the **standard normal distribution** we apply **probit model**.
- ❑ We run the probit model and then predict the values of residuals:
  - `predict r`

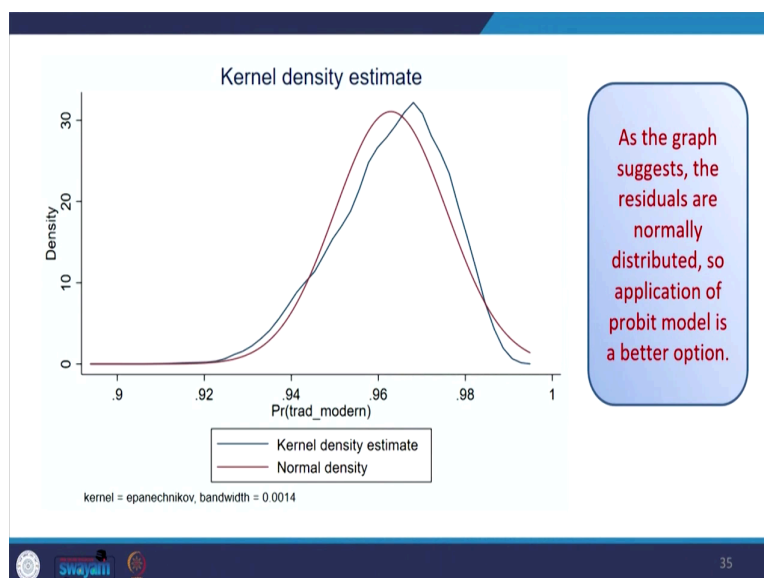
To check the distribution of residual, we use kernel density function:

- `kdensity r, normal`



I am deliberately skipping marginal effects and all their details probit and their marginal effects I have already explained. So, no need to explain it further we have already run the command called `predict r`. Then we explained about k density function and what is the necessity of k density function I already discussed.

(Refer Slide Time: 33:47)



This is what the result looked like and I think you must have understood.

(Refer Slide Time: 33:52)

### Comparison of Models and Parameters

- `quietly regress trad_modern HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE`
- `estimates store blpm`
- `quietly logit trad_modern HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE`
- `estimates store blogit`

36

Now, how to make comparison of all those models and their parameters. these are the command we have also given step by step and I am quite sure with these command you can able to estimate and compare.

(Refer Slide Time: 34:01)

- `quietly probit trad_modern HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE`
- `estimates store bprobit`
- `estimates table blpm blogit bprobit, star (.01 .05 .10)`
  - ☐ Star option is specified to get significance along with coefficients.
- `estimates table blpm blogit bprobit, t stats(N ll)`
  - ☐ t value and other stats like number of observation and log likelihood is specified.

37

(Refer Slide Time: 34:04)

- ❑ `quietly` command is used to suppress terminal output.
- ❑ `Estimates store` command stores the estimation result of different models.
- ❑ `Estimates table` command organizes estimation results from one or more models in a single formatted table.
- ❑ Similarly we can also compare marginal effects of three models.

38

(Refer Slide Time: 34:09)

**Marginal effects:**

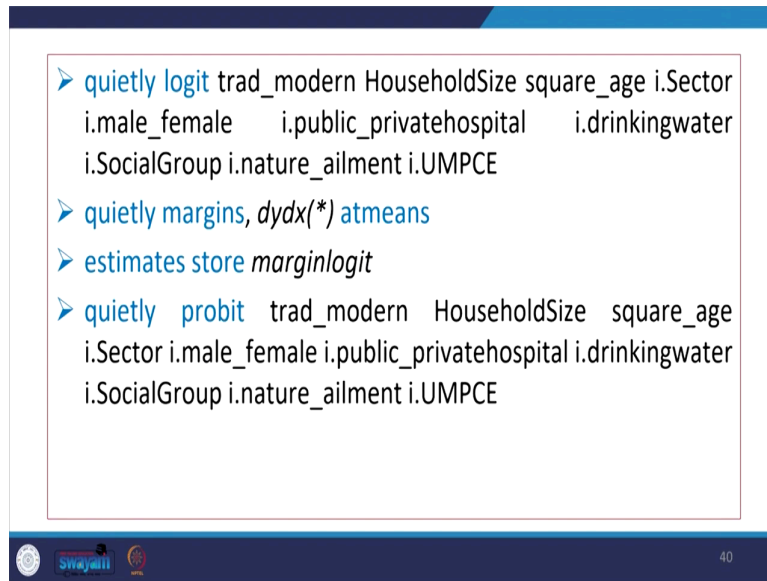
- `quietly regress trad_modern HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE`
- `quietly margins, dydx(*) atmeans`

This command gives similar values as coefficient of OLS model.

- `estimates store marginlpm`

39

(Refer Slide Time: 34:11)



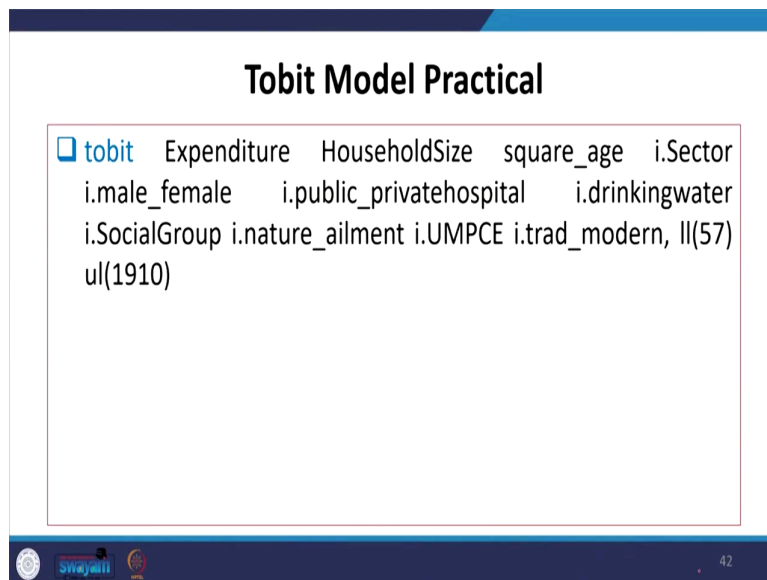
Slide 40 displays four Stata commands for logit and probit models. The commands are:

- `quietly logit trad_modern HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE`
- `quietly margins, dydx(*) atmeans`
- `estimates store marginlogit`
- `quietly probit trad_modern HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE`

The slide footer includes the Swayam logo and the number 40.

If you have still difficulties my videos are already floated in YouTube. I am sure you can able to understand very clearly at this moment I am not repeating. Because this is simply consuming time and with this less span of time you guys can able to get all those details.

(Refer Slide Time: 34:27)



Slide 42 is titled "Tobit Model Practical" and shows the following Stata command:

```
▣ tobit Expenditure HouseholdSize square_age i.Sector i.male_female i.public_privatehospital i.drinkingwater i.SocialGroup i.nature_ailment i.UMPCE i.trad_modern, ll(57) ul(1910)
```

The slide footer includes the Swayam logo and the number 42.

Then comes tobit, tobit model is applied when we have the data that is having continuous data the dependent variable is continuous. But those who are not following normal distribution or not properly normal in the sense it is not symmetric.

When it is not symmetric either it will be positively skewed or negatively skewed. That means, there are some what is called data at the extreme points, maybe at the lower extreme, at the upper extreme. So, in that case tobit is taken instead of OLS if you run the OLS the since distribution is not normal OLS it is not applied.

Your expenditure is the variable others we are not going to care for it. So, it simply you take tobit and run the regression I am sure you will get the result. So, like some other details like tobit sensor how it is different, how it is truncated look from the lower side or from the upper side.

You have to read it from my previous lecture, previous module and I am not emphasizing because we have already exceeded the time of this particular lecture. So, with this I so, stop here I am sure you will understand and apply in your day to day research work and develop research paper. If you have difficulties do not hesitate and raise your queries we will be most happy to address it.

Thank you.