

Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian institute of Technology, Kharagpur

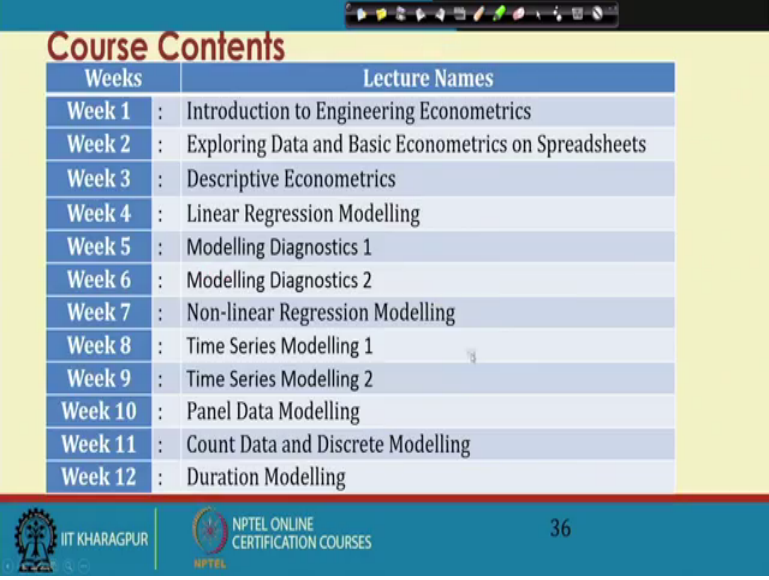
Lecture - 39
Extension of Dummy Regression Modelling- Dummy Dependent Variable Modelling

Hello everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today we will continue with non-linear regression modeling. As we have discussed earlier, non-linear regression modelling have two parts that too with respect to dummy modelling. So, the 1st part is a dummy modelling with respect to independent variable and the 2nd one is dummy modelling with respect to dependent variables.

So, that means in the first case dependent variable will be numeric in nature while at least one independent variable should be qualitative in nature and on the other side, the dependent variable will be qualitative in nature and no restriction to independent variables. In fact, we have discussed both the aspects and we have it, we have already highlighted a case where dependent variable is numeric in nature and that too one or more variables are there.

These are all independent variables which are qualitative in nature and in the 2nd group where we are using dummy dependent variable modelling, we have 3 different forms of model what we have discussed earlier. The 1st part is the linear probability model, the 2nd part is logit model and the 3rd part is the probit models. So, let us see what are these you know structures.

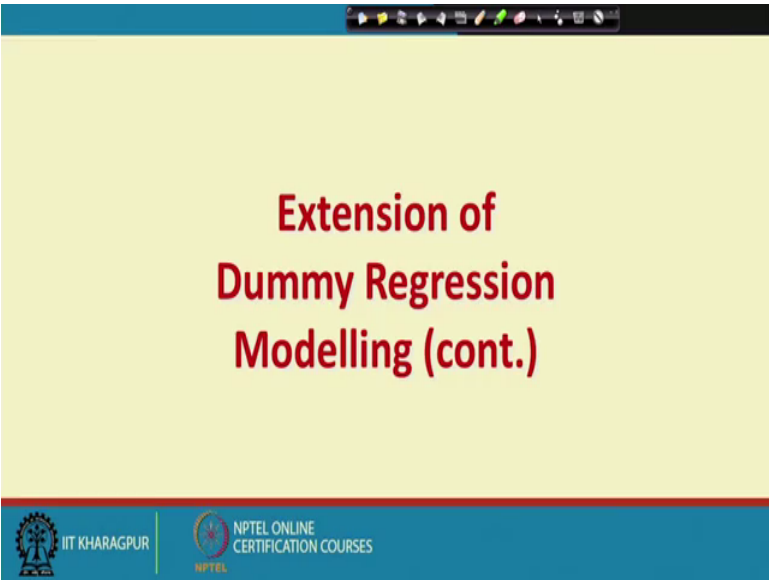
(Refer Slide Time: 02:06)



The slide displays a table titled "Course Contents" with two columns: "Weeks" and "Lecture Names". The table lists 12 weeks of content. At the bottom of the slide, there are logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES, along with the number 36.

Weeks	Lecture Names
Week 1	: Introduction to Engineering Econometrics
Week 2	: Exploring Data and Basic Econometrics on Spreadsheets
Week 3	: Descriptive Econometrics
Week 4	: Linear Regression Modelling
Week 5	: Modelling Diagnostics 1
Week 6	: Modelling Diagnostics 2
Week 7	: Non-linear Regression Modelling
Week 8	: Time Series Modelling 1
Week 9	: Time Series Modelling 2
Week 10	: Panel Data Modelling
Week 11	: Count Data and Discrete Modelling
Week 12	: Duration Modelling

(Refer Slide Time: 02:08)



The slide features the text "Extension of Dummy Regression Modelling (cont.)" in a large, bold, red font, centered on a light yellow background. At the bottom, there are logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES.

Extension of Dummy Regression Modelling (cont.)

Then, we will move into this you know discussion about the linear probability model, logit model and probit model. So, what we have discussed earlier it is with respect to linear probability model.

(Refer Slide Time: 02:19)

Linear Probability Model

- Obvious possibility is to use traditional linear regression model

$$\pi = \alpha + \beta x$$

But this has problems

- Distribution of dependent variable hardly normal
- Predicted probabilities cannot be less than 0, greater than 1

The slide features a yellow background with a blue header and footer. The equation $\pi = \alpha + \beta x$ is written in black with red circles around π , α , and βx . Red arrows point from the text 'But this has problems' to the equation. The footer includes the IIT Kharagpur logo and the text 'NPTEL ONLINE CERTIFICATION COURSES'.

All these 3 forms of models are connected with the probability and so far as a probability is concerned, we have two extremes 0 and 1 and if naught, then the information relating to dependent variable will be in between 0 1.

So, that means it will be mostly in ratio type and that too that too the information will be positive in nature. That is the kind of you know requirement while analyzing the linear probability model, logit model and probit model. So, obviously we have two different cluster. In the 1st cluster, it is the probability model which is linear in nature and in the 2nd group logit and probit it is a probability model which is non-linear in nature. So, in the 1st case this the modeling will be simple like you know y equal to α plus βx like this.

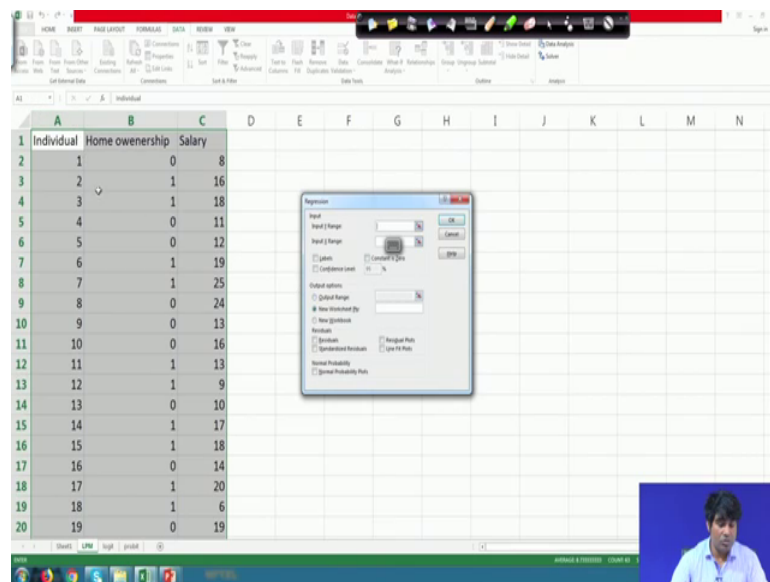
So, like this here this one is the dependent variable and x is the independent variable. So, in the case of linear probability models, the dependent variable information will be either 0 or 1. That is the structure of linear probability model and then, x you know may be 1 x or you know more x which is not necessarily qualitative, it may be qualitative, it may be a numeric in nature, but dependent variable must be qualitative and it will be in between 0 to 1.

So, how will you transfer into 0 to 1 that is the kind of you know example which we like to cite and that too really find out a problem where this model can be fitted. So, that means all kinds of you know engineering problems, this model cannot be applied. The

use of this model is a very kind of you know restricted area or restricted environment where the dependent variable must be qualitative in nature and the information for the qualitative variable will be only between 0 and 1. So, that means kind of you know yes, no, male, female, these are the kind of you know situation we have to apply, but in this kind of you know models, we like to use the concept you know probability.

So, that means instead of saying you know yes, no, male, female, so here the particular information will be derived through a kind of you know probability. So, in order to dis, you know understand the particular you know models so let us start with a particular example and then, we can discuss this model the 0, the model will be like this.

(Refer Slide Time: 05:22)



So, let us assume that you know this is a kind of you know spreadsheet where we have we have the data here and we have taken actually 20 observations and this is what the 20 observations and we like to discuss you know linear probability model and 20 observation means it is an information about 20 individuals and they are in a job and we like to collect actually 2 information corresponding to dependent variable and independent variable and that too the use of you know linear probability models or binary choice model.

So, the first variable which we have here is the home ownership and the second variable is with respect to salary. So, that means the sample structure will be like this. We have to you know touch upon a particular individual and ask whether he is in job or not and then,

whether he is staying in a rented house or own house. That is what the ownership is all about. So, that means here the hypotheses that are simple understanding is that you know if a guy will be having high salary, he may have the opportunity to stay in the you know own house. If the salary is not sufficient, he may be staying in rented house. So, that is the kind of you know game.

So, that means technically having house or not having house is the kind of you know dependent variable. It is kind of you know yes no type of situation, but ultimately it is the kind of you know transfer into a probability format only. So, the moment you will say yes no, then obviously in place of yes, you can put 1 and in place of you know no, you can put simply 0. So, that means simply we will visit the customers or the particular individual ask whether he is in job or not and if he or she will be in a job.

So, then the first question will be what should be your salary structure and then, second one is whether you are staying in rented house or you know own house. So, as a result we have a clear cut information here. So, that means in the first individual the, in the case of in ownership it is 0. So, that means he is possibly staying in rented house and not in own house and he said our structure will be 8 lakhs per annum.

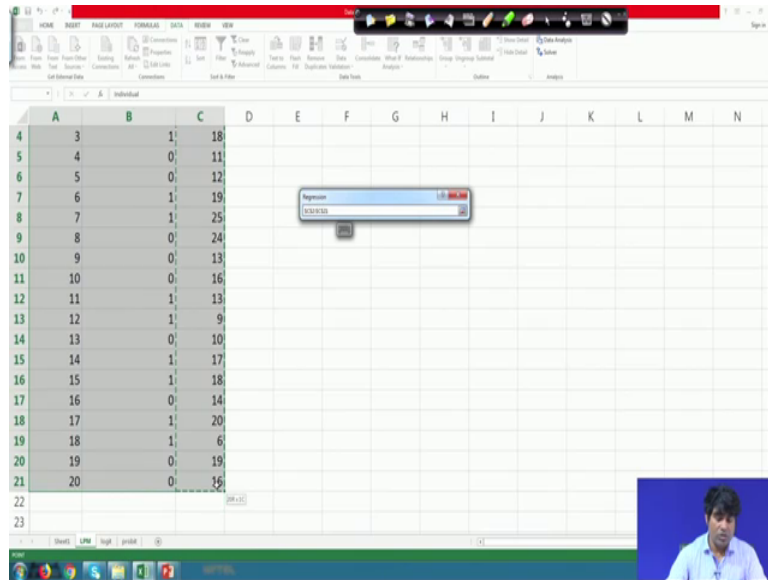
So, the figures of salaries are in lakhs and then, lakhs per annum and we like to know whether salary you know is very useful component to decide to have you know own house or to stay in the rented house. So, likewise we have actually 20 individuals. So, every time we are asking whether they are staying in the own house or rented house and what is the salary structure and here the salary structure itself is an independent variable and having a house or not having house, that means staying in own house or staying in a rented house will be the dependent variable and that too the transformation will be 0 to 1.

So, after doing the transformation, that means technically the actual information will be yes no. So, then it will be 1 here where there is you know, then you will put a 0 and after doing this, then we have to actually transfer the data in a structured form and then, we will go for the estimation. So, the estimation process is you know usually you know same like you know like the case of you know bivariate you know regressions or simple you know multiple regression.

So, here we can still use actually data analysis package and click this once and then, we choose the regression and put after putting, ok. So, then you will see here. So, our target

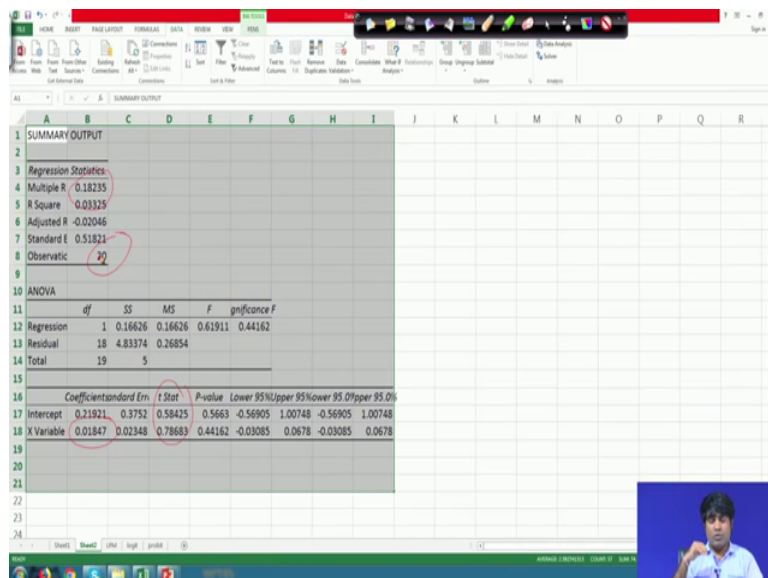
is here to give an indication about the dependent variable. So, this is what the dependent variable indication which we will highlight all the samples and then, we can give the input range starting with you know 8 lakhs to 16 lakhs.

(Refer Slide Time: 09:52)



Then, we fill up the you know structure and after that you just put actually.

(Refer Slide Time: 09:59)



So, by default we will have a result here and this is what the results, right. So, as usual this is repeat simple regression estimation. Only thing we need to actually structure the data and then, go for the estimation and structure of the data itself has lots of you know

interpretation and then, we will predict the kind of you know situation whether salaries infact you know to decide whether a guy will stay in a rented house or in a kind of you know own house, right. So, that means technically, it is the game which with respect to yes no and then, that will be transferred into 0 1 and then, finally we like to regress.

Then, as usual we like to check whether the coefficient of you know salary is significant one while deciding to stay in a own house or the rented house. So, as a result we have the results here and of course, you know the results are not so good here, but by the way the coefficients are you know very low and in the kind of you know R square is also very low, but by the way it gives you a kind of you know positive indication. That means, you know the impact is positive.

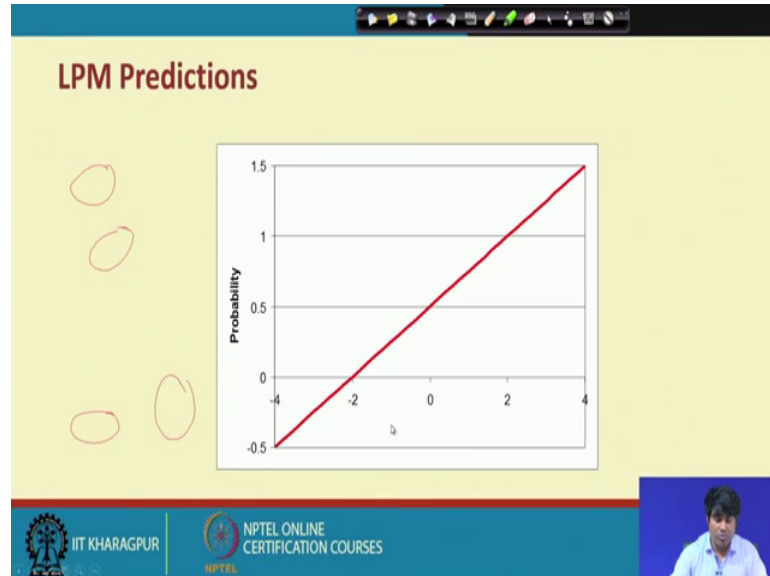
So, that means technically the simple interpretation will be like this. So, you know having you know more salary, then the guy can stay in the rented house and having low salary, the guy may be staying in means if the salary structure will be high, then he will be staying in own house and when the salary structure will be low, then he will be staying in the rented house. So, obviously the theory says that you know there is a positive relationship between the two.

So, that means a salary and housing structures will be a positive link ah? Having high salary, the staying will be in the own house and low salary will be on a rented house and that is what we are getting here and in fact, the particular coefficient is not statistically significant. Even F is not statistically significant and then, the R square is also not actually too high. The reason is that we have only low samples. You know technically it is only 19 sample and as not exactly it is 20 samples.

19 is you know degree of freedom and here is the total observation that is 20 and if you can increase the sample size, then obviously there is a high chance that you know the particular you know variable will be statistically significant. So, ultimately our M is not to just you know you know check whether the particular variable is statistically significant or not. Of course, that is the one of the requirement of this modeling, but ultimately here our idea is just to know how it works as far as the linear probability model is concerned, where the dependent variable is the dependent means and you know dummy type and the information when the dependent variable will be in between 0 to 1, 0 and 1 not 0 to 1 and then, independent variable will be numeric in nature and as a result

we will have the estimation like this and the kind of you know structure will be starting like this.

(Refer Slide Time: 13:45)



So, usually the linear probability models will be in a straight line kind of you know structure. So, it will be in a kind of you know increasing a slope where the indication is that you know you know salary structures and you know person saving house, own house is actually positively linked.

So, this graph itself is giving such kind of you know indication because this slope is actually coming positive. So, that means you know linear probability models that is LPM is a kind of you know structure where you can easily predict certain things, where you know the information will be qualitative in nature. In fact, this particular model can be extended to logit model and probit model.

So, that means the same problem. So, that means here the issue is that you know we like to decide or we like to know whether salary is a factor you know to decide whether customers or they kind of in a you know kind of you know persons who can stay in own house or rented house. The same problems can be tested through linear probability model, logit model and the probit models.

So, now what is happening in the case of you know linear probability model, this particular sample is enough to go ahead with the estimation even if it is not statistically

significant, but model is working. So, that means you know the particular information is the correctly specified to test this particular you know inference.

So far as a significance level is concerned, you can increase the sample size and get the results, but it is coming you know as per the expectation and that too we expect positive relationship and the coefficient is also coming you know you know positive in nature. So, as a result the linear probability model is a very effective tool to decide, to decide whether a person will be staying in a rented house or own house left with respect to you know salary structure. Let us see the same problems can be tested through logit models.

(Refer Slide Time: 16:13)

Logistic Regression Model

- Instead, use logistic transformation (logit) of probability, log of the odds

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

Odds ratio = $\frac{PS}{PF}$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Here the logit model will be you know the structure will be like this log p upon 1 minus p equal to alpha plus beta x. So, that is the kind of you know things we like to check. What is happening here the same structure that mean; that means, if you go through the linear probability models here, this is actually a the pi that is the dependent variables lying between 0, either 0 or 1 and the functional form will be alpha plus beta x. That is the right hand side of a model in fact, this is also same here.

The right hand side of the game is more or less seems compared to linear quality model and if you see the first one and this one, then you will find the left hand side person is actually different. So, here the transformation you know transformation is the log p upon 1 minus p. That is the item which is actually called as you know odd ratio. So, that means p by 1 minus p is the odd ratio, where p is the probability of success and 1 minus

p is the probability of failure. So, ultimately what we need to have is to find out p value first. So, like the previous case we have actually observed p value which is actually type of you know yes no and that to the transformation will be 0 1. Only since probability value two extremes are 0 and linear probability for it is like that only. So, we can easily run the model and get to know the kind of you know inference.

Now, the same information we like to handle through this logit model. In that case, your dependent variable structure will be somehow you know different. So, here we simply means we use the same p , but instead of you know direct p , we will use $\log p$ by $1 - p$. That means, it is nothing but log upon odd ratio which is actually the ratio between probability of success to probability of failure.

So, that means we will use a component called as you know odd ratio which is actually which is actually the ratio between probability of success by probability of failure. That means, technically p by $1 - p$. So, what is actually requirement you know is that, we need to go for you know data transformations or data restructuring. So, that means the first item requirement is p , then we will go for $1 - p$.

So, just you know $1 - p$ will get the data because total probability exactly equal to 1. So, whatever p will be, find that if that is the indication of the success by default $1 - p$ will be the indication about the failure. So, that means technically how we have to deal with this problem. So, ultimately if we will go to this you know data set technically let us go to this is data set, where we have a linear probability model. In this case what is happening is, so we have actually this is the probability value.

So, that means technically the actual information will be yes no and we transfer into 1 0 where 1 represents yes and 0 represents no. So, now by default this will be this will be the indication of you know probability, right. So, either probability will be 0 or probability of equal to 1. So, now when you use actually a linear probability model, then this much of information is sufficient however, if you use in logit model, then only p is not enough. So, at the same times you have to calculate actually $1 - p$ and then, the kind you know odd ratio. So, obviously the odd ratio is p by $1 - p$. So, also as a result this sample is not sufficient to help to calculate the you know odd ratio because it is the ratio between p by $1 - p$.

(Refer Slide Time: 20:27)

Individual	Home ownership	Salary
1	0	8
2	1	16
3	1	18
4	0	11
5	0	12
6	1	19
7	1	25
8	0	24
9	0	13
10	0	16
11	1	13
12	1	9
13	0	10
14	1	17
15	1	18
16	0	14
17	1	20
18	1	6
19	0	19

So, from p equal to 1 by default $1 - p$ equal to 0. So, if that is the case, then the whole ratio will be equal to infinite. So, now \log of infinite is undefined. So, as a result that means technically the \log of p by $1 - p$ is the odd ratio where you know this is that you know you first calculate p and then, you calculate $1 - p$ and then, finally p by $1 - p$, which is the kind of you know requirement.

So, now what is happening if we equal to 1 that is one extreme, then $1 - p$ equal to 0 and when p stands to 0, then $1 - p$ stands to 1. So, now when p equal to 1, then $1 - p$ equal to 0. So, now in this case p by $1 - p$ is equal to 1 by 0, ok. In this case if you put actually 0 here, then divided by you know $1 - 0$ means it is 1. So, that means the order ratio is, odd ratio here is 0 and odd ratio here is the infinite.

So, now ultimately the transformation will be instead of you know 0 1, the transformation will be 0 and infinite. So, that means if you directly use this data, as a result you are not in a position to estimate the model because the information content while you know having you know $1 - p$ equal to 1 that will be infinite. So, as a result we need actually different kind of you know structuring or you know transformation, so that you know we can use this information to estimate the model and to get the kind of you know inference as per our you know you know a problem requirement or the kind of you know engineering econometrics requirement.

So, ultimately what we like to do? We like to see, what is the structural change, you have to apply here so that you know without you know changing the overall structure of the model, we can still run the model and then, test this particular you know hypothesis. So, as a result we need actually transformational mechanism for that. So, we can go to another problem, the similar kind of you know information, but the processing, a processing structure will be different. So, as a result we come to this you know second data set. So, this is a data set here and in this data set.

(Refer Slide Time: 23:25)

Income	N	n
6	40	8
8	50	12
10	60	18
13	80	28
15	100	45
20	70	36
25	65	39
30	50	33
35	40	30
40	25	20
45	20	15

So, we have actually same informations. So, this is in fact we have actually 12 informations, sample size is $2n$ and then, what is happening this is the income level and this is capital N , this is small n .

So, that means technically instead of you know targeting an individual, we will target a group of individuals let us say a particular you know organization and then, we identify 12 such organization and then, we target you know couple of you know people in that particular organization, then ask what is the salary structure and whether they are staying in own house or in a rented house means simply we are asked that any way that they have their own house or they have not there you know own house.

So, you know obviously when you actually you know touch upon you know 40 peoples, let us say the first sample is 40 year, then obviously few will be actually having you know own house and few may not have their own house. So, now if the p component

will be the, p component will be here n by n , so that is the usual structure of you know probability calculation. So, when you calculate probability will be n by n . So, there is a high chance that the a you know the particular transformation will be in between 0 1 , it may not be 0 , it may not be 1 . So, now we need the data in between 0 to 1 to apply the logit model.

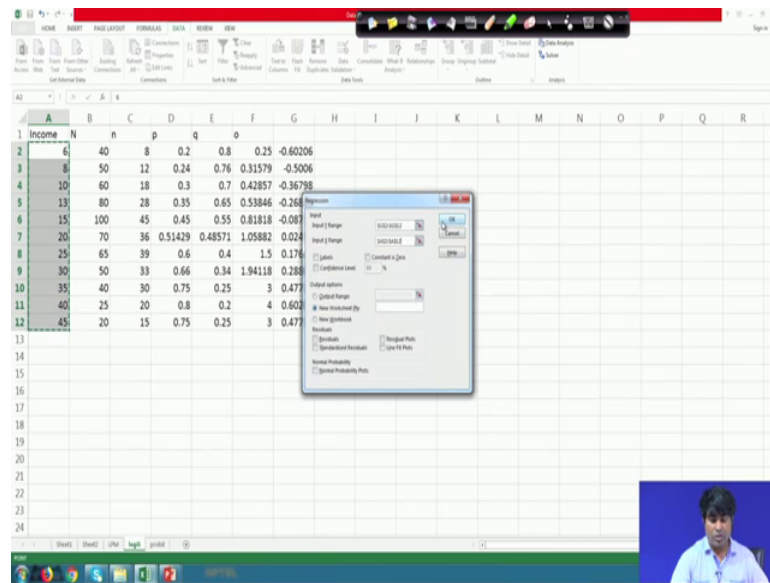
So, as a result what will you do? We first you know pick up in organization or the kind of you know area, then you pick up a couple of you know individuals there and as you know with the structure is that you know out of the you know all these you know individuals, you must have their you know own house, so that you know we can actually get p value you know as per your you know requirement. Otherwise suppose you know you know all people having their own house or you know all people have not there in own house, then this kind of you know something may not be useful here to you know estimate the particular you know logit model.

So, that means sample, specific sample is high requirement here and then, we have to link and you know estimate as per our you know requirement. So, that means technically it is the group sampling which is the requirement to estimate the logit model. So, instead you know like linear probability model, you just you know randomly pick up individuals and generate the data, but here you have to be you know identify the organizations or the area or the kind of you know locations, then you pick up a group out of which you know few of them having their own house and few of them not having their you know own house.

That is by default. So, as a result we are in a position to calculate p here and the moment you get p value which is actually in between 0 to 1 and by default we have 1 minus p which is again positive, then we will find a ratio which is also coming positive, then after that we can go for you know estimation. For instance, let us see here.

So, this is the data set here. So, what we can do here, we can calculate p here, this is what actually we can have here.

(Refer Slide Time: 27:14)



This is actually let us say p indications. So, now the p indication will be 8 by 40, and then, you can just scroll it and you will have the structure is like this, ok. So, then ultimately this is actually p value and you see here the p value range is actually between you know 0 to 1. So, it is not exactly 0. It is not exactly 1. So, then you will calculate 1 minus p. So, let us say it is you know q and then, which is actually the difference between 1 minus p so, that is the first one is the 0.2. So, you can by default it will be 0.8. So, you really see here the difference.

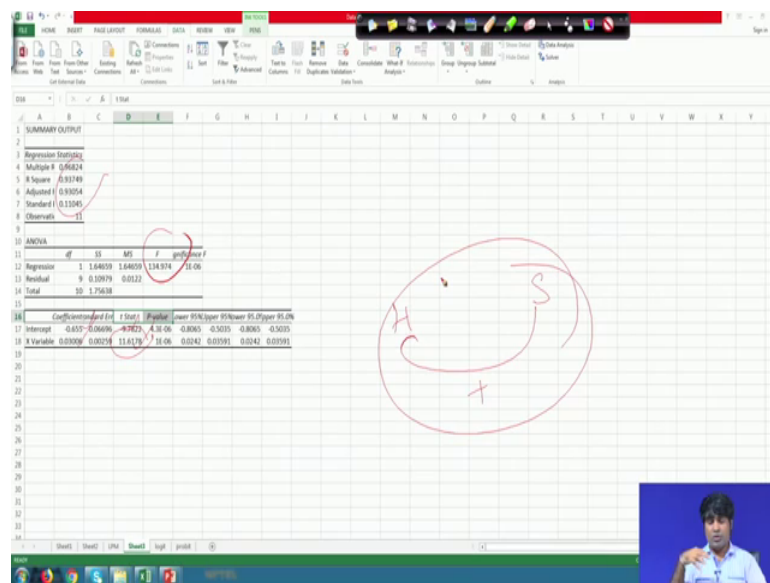
So, now if you add these two, then you will get 1. If you add these two, you will get 1 because total products that will be equal to 1. So, now this is the p information and this is the q information. So, what we need? Actually we needed you know transformations. So, what is the transformation? That is actually you know p by q. That is the odd ratio. So, on the odd ratio let us say that it is war which is nothing, but actually p you know divided by q.

So, that will be having, sorry this will be equal to p divided by q. So, you will get a ratio here. So, that will be 0.25 and then, you just you know generate the series. So, we have a series like this, ok. So, then what will you do? That is called as you know p by 1 minus p and what will you do, we look up for log up on you know odd ratio.

So, as a result you can go for another round of you know transformation which is nothing, but you know log transformation also, put you know our log here. So, this is the

log transformation. So, now you can just give any indications you will find the log transformation here. So, this got the log transformation. So, this is what actually p log of you know odd ratio. So, now this information is used as you know dependent variable and the independent variable will be in these one's. So, now what will you do? You again go to the data analysis and choose regression. So, now what you can do is, you know reset. So, there are two dependent variables is here only. So, this one, and the independent variable will be in independent variable will be income and then, run the models.

(Refer Slide Time: 30:09)



So, this is what the estimations. Of course, this gives you know better result compared to the previous one here you know if you compare with the linear probability model. So, the model result is not so good you know like this it is a very poor 0.18 and the coefficients are you know not statistically significant know if the same problems if you actually run through logit model, then it has actually huge impart and that too the effects are you know very clear here. In fact, as usual these coefficients are also statistically significant here and not only significant, it is a positive coefficient and R square is very high and F is also very high and both you know intercept and slope are statistically significant and the slope is positive.

So, that means a salary structure and you know having a house, they are actually positive building. That is what we have got the inference here. Now, if you compare the first

problem and the second problem, we find you know there is a drastic improvement. So, that means this kind of you know problem is having actually you know better way of analyzed, you know analysis or you know you know some kind of you know good inference you know it to come to the particularly you know conclusions while you know you know checking the issue of you know salary structure and kind of you know peoples own house and that means, technically we have 3 different kind of you know structure and the same problems can be checked through linear probability model, logit model and probit model and then, you can compare and where it is actually having good result, we can keep and the remaining one can be discarded.

Then, finally it will give you the kind of indication that yes you know it is a kind of non-linear format and that too it is a situation where actually the dependent independent variable is influencing the dependent variables, where the dependent variable is a qualitative one and that too it is in the form of you know probability and again the information will be in between 0 to 1, not exactly 0, not exactly 1 like the case of you know linear probability model.

So, like linear probability model, logit model is another kind of you know dummy models that is qualitative as regression response modelling, where you know we can predict the independent variable to dependent variable which is actually qualitative in nature and the third one is the probit model, which is a similar kind of you know structure, but a transformation mechanism will be different which we will discuss in the next class. With this we will stop here.

Thank you very much.