**Business Analytics for Management Decision**
**Prof. Rudra P Pradhan**
**Vinod Gupta School of Management**
**Indian Institute of Technology, Kharagpur**

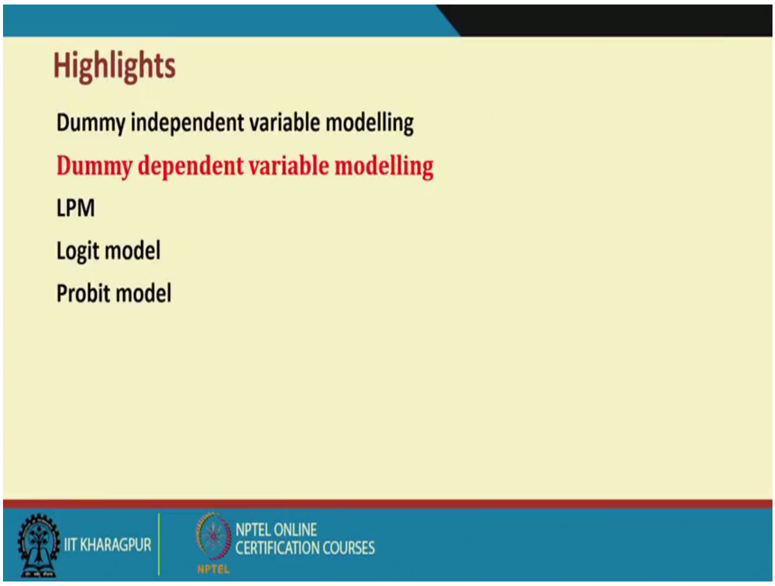
# Lecture – 32
# Predictive Analytics: Dummy Modelling (Contd.)

Hello everybody and this is Rudra Pradhan here. Welcome you all to BMD lecture series. Today we will continue predictive analytics and that too dummy modelling part 2.

So, in the last lectures we have discussed this dummy modelling concept and that too we are going through predictive kind of you know structure, where one or you know many independent variables may be categorical in nature, but again in real life scenario you will find plenty of you know managerial problems or business problems, where you will find dependent variable structure will be categorical.

So, as a result here you know the kind of you know predictive structure will be very challenging and again it will be also more interesting like you know dummy independent modelling. So, dummy dependent modelling is also very interesting kind of you know structure through, which you can actually go for you know hardcore management kind of you know decision and then the kind of you know management predictions. So, in order to know detail about the particular you know structure.

(Refer Slide Time: 01:23)



Highlights

Dummy independent variable modelling

Dummy dependent variable modelling

LPM

Logit model

Probit model

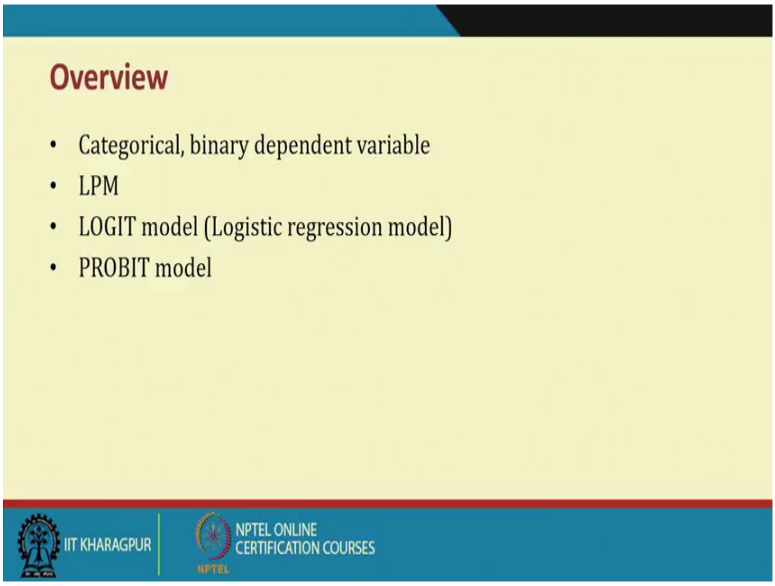
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

.

So, let me start with the kind of you know discussion today. And in the dummy dependent modelling structures, we like to focus on 3 things. And typically there are 3 different you know models which you frequently use in the kind of you know dummy modelling that to dummy dependent structure, that is called as you know Linear Probability Model, Logit model and Probit model.

So, all are you know more or less same, but there is a kind of you know structural kind of you know structure through, which we have to actually understand and then connect as per the problem requirement. So, the linear probability model is nothing, but actually the is a kind of you know linear regression modelling and logit and probit is a kind of you know cluster called as you know general non-linear regression modeling. In that to in a kind of you know dummy set of.

(Refer Slide Time: 01:13)



Overview

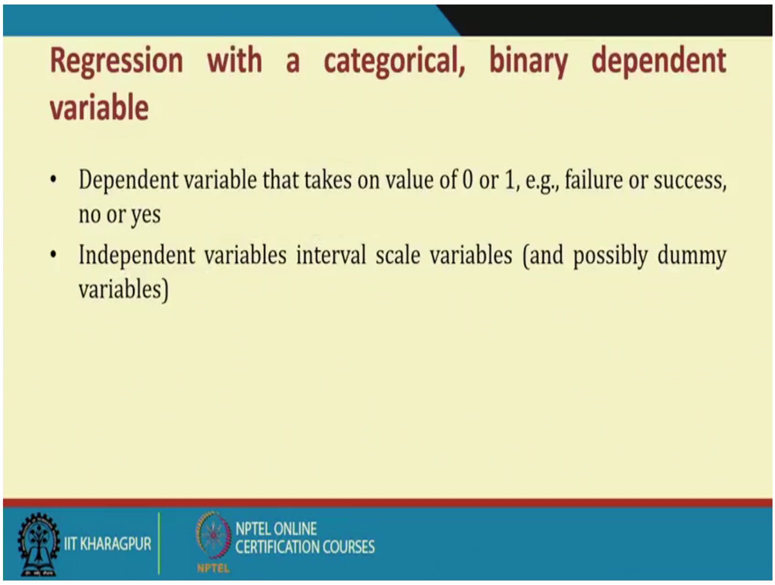
- Categorical, binary dependent variable
- LPM
- LOGIT model (Logistic regression model)
- PROBIT model

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

.

So, these are you know kind of you know structure through which you have to discuss.

(Refer Slide Time: 02:17)



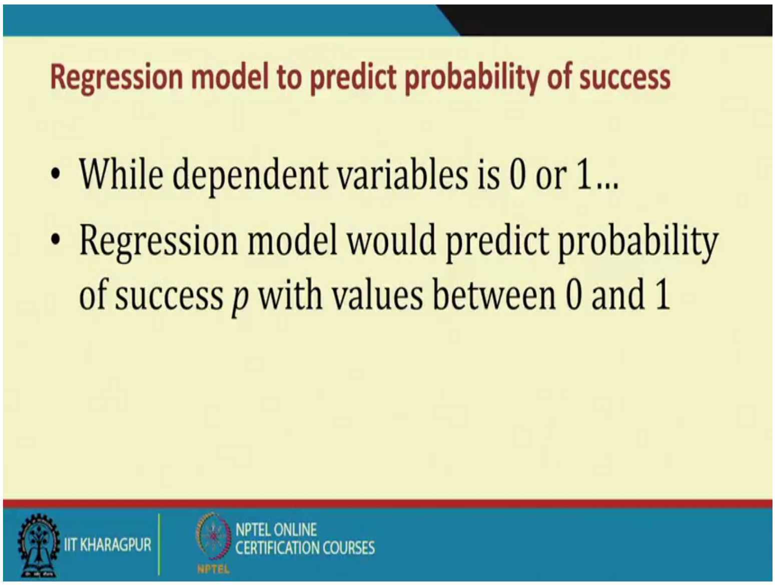
Regression with a categorical, binary dependent variable

- Dependent variable that takes on value of 0 or 1, e.g., failure or success, no or yes
- Independent variables interval scale variables (and possibly dummy variables)

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

So, let us first understand the dummy dependent structure all together. So, most of the cases either it is a binary representation like a failure success, yes no, true false, kind of you know situation and in that context, we use a kind of you know probability model and that to it is called as a you know the linear probability model or sometimes it is called as you know binary choice model. And some in some occasions the dependent variable structure may be categorical so; that means, it may have the option of you know more such kind of you know categorizations.

(Refer Slide Time: 02:57)



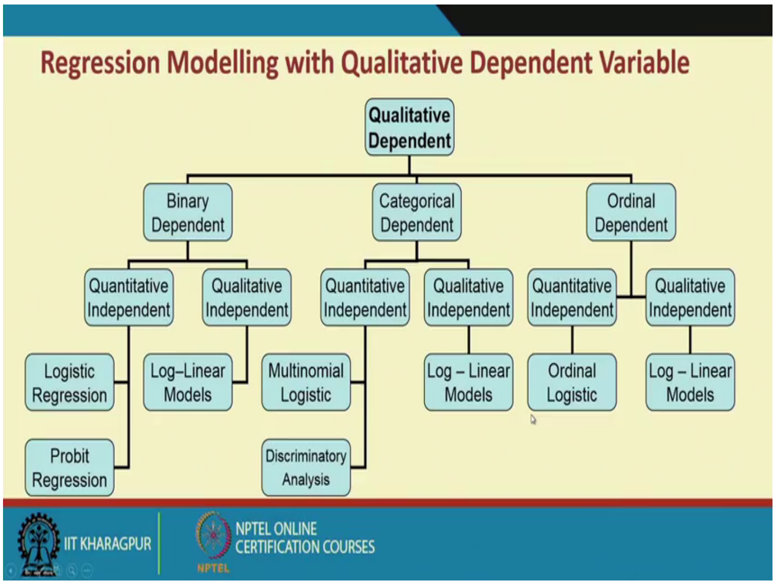
Regression model to predict probability of success

- While dependent variables is 0 or 1…
- Regression model would predict probability of success $p$ with values between 0 and 1

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

So, now in order to know the kind of you know details. So, let us see here's how is the a kind of you know predictive structure or the kind of you know modelling structures, when the dependent variable structure will be a binary or some kind of you know categorical.

(Refer Slide Time: 03:15)


Regression Modelling with Qualitative Dependent Variable

.

So, accordingly so we have actually a different structure all together and this is the you know complete structure of you know dummy dependent modeling. We start with actually qualitative dependent structure and that to dependent variable structures we are not highlighting here. Because we are you know stately focusing on you know dummy independent modeling. And in the binary dependent it can be with respect to quantitative independent and with respect to qualitative independent.

Then accordingly in the case of quantitative independ it will go with logistic regressions or you know probit regressions, which is our actually todays you know discussion. And in other sides if it is a qualitative independent, then this will go with you know log linear models.

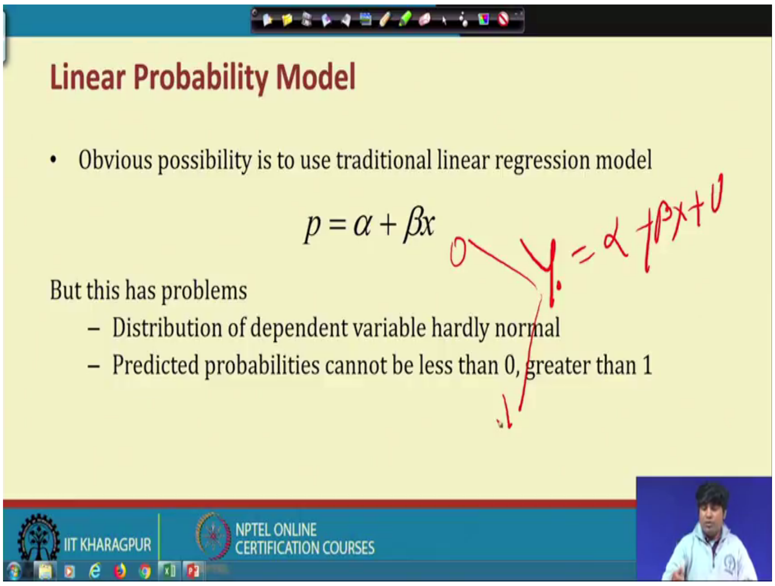
And again in the categorical dependent it may be quant with you with respect to quantitative independent or with respect to qualitative independent. In the case of quantitative independent, it can be with you know multinomial logistic or it can be go with you know discriminant discriminatory analysis, but in the case of qualitative independent it will go with the log linear models.

Again in the case of you know ordinal dependent we have actually 2 options and that to you know it may be with you know ordinal logistics or it will be with you know log linear models.

So; that means, actually the kind of you know dummy dependent modelling we have actually plenty of you know different structure, through which you can do the kind of you know modelling and do the kind of you know predictive structure as per the you know management requirement.

So, by default after looking this chart we can understand that you know this has lots of you know utility or flexibility to why you know and that is highly required for the kind of you know management or the kind of you know business environment.

(Refer Slide Time: 05:13)



## Linear Probability Model

- Obvious possibility is to use traditional linear regression model

$$p = \alpha + \beta x$$

$$y = \alpha + \beta x + u$$

0

1

But this has problems

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

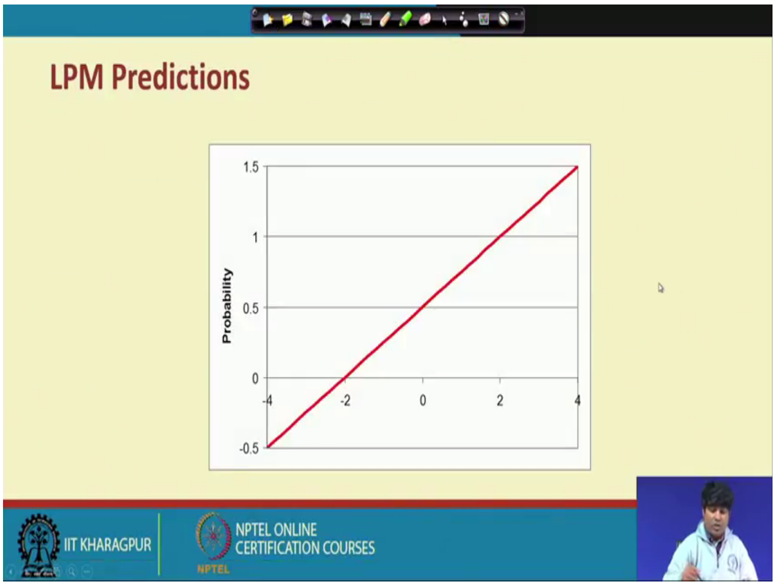
So, let us see how is this particular you know dummy independent variable modelling structures and then we will connect with some of the sample problems. And this standard model you know start with you know dummy dependent modelling is called as a linear probability model and in that context.

So, your y is nothing, but called as you know and I can actually write here y equal to simply a alpha plus beta x plus u. So, this is actually the original model, but here every times y will be in between 0 1 so; that means, the variable and the problem means a problem will be in such a way that the y information will be in between 0 and 1 only.

So, that means, either yes no or something kind of you know like male, female kind of you know situation. So, it is only you know vary between 2 kind of you know alternative situations.

And then it will be a predicted with some of the independent variables. So, now, the particular you know modelling structure connected with you know probability and as a results we have actually 3 different models all together this is linear probability model otherwise called as a binary choice model, BCM and then we have a logit model and probit model.

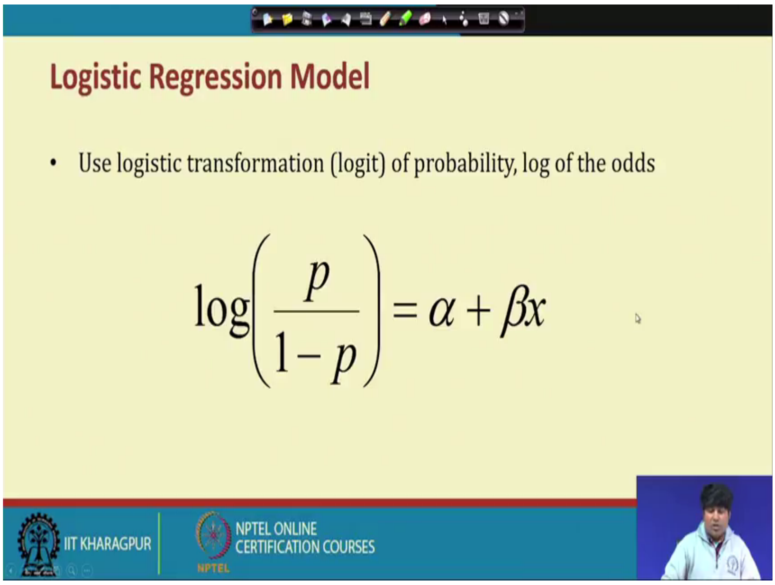
(Refer Slide Time: 06:34)



LPM Predictions

So, now it so let me first you know connect with this you know linear probability model and this is the sample case. So, the since it is a linear kind of you know modelling structure. So, the kind of you know prediction will be a more or less you know straight line.

(Refer Slide Time: 06:46)



Logistic Regression Model

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

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

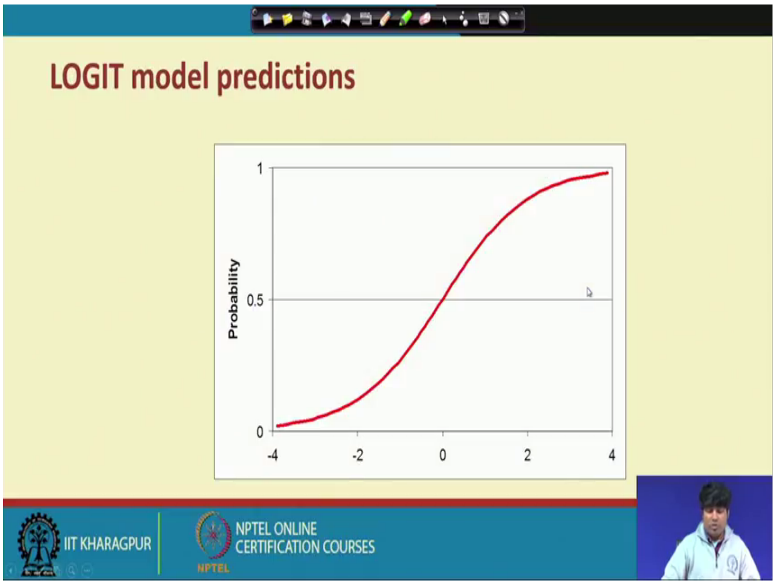
And which is actually slightly different in the context of you know logistic regression modeling. And the logistic regression modelling the typical modelling framework is a log of v by one minus p which is equal to alpha plus beta x. So that means, technically on their case it is p equal to alpha plus beta x here log upon p by 1 minus p, that is actually called as you know odd ratio.

And which is actually ratio between probability of success to probability of failure and that need to be predicted with the independent variables. And that is what you know a little bit you know more advance and little bit more complex compared to linear quality model.

And this is followed by logistic functions and technically. So, the y variable is means technical p is nothing, but 1 1 by 1 plus e to the power minus that is what the logistic function all together. So, the usual the graphical if you go by graphical plotting then logic model behavior will be like this it follows a logistic function all together compared to linear probability model it is a straight line equations right.

So, only thing is actually you have to check, whether you know the particular model is fitted under this you know you know problem kind of you know situation or not.

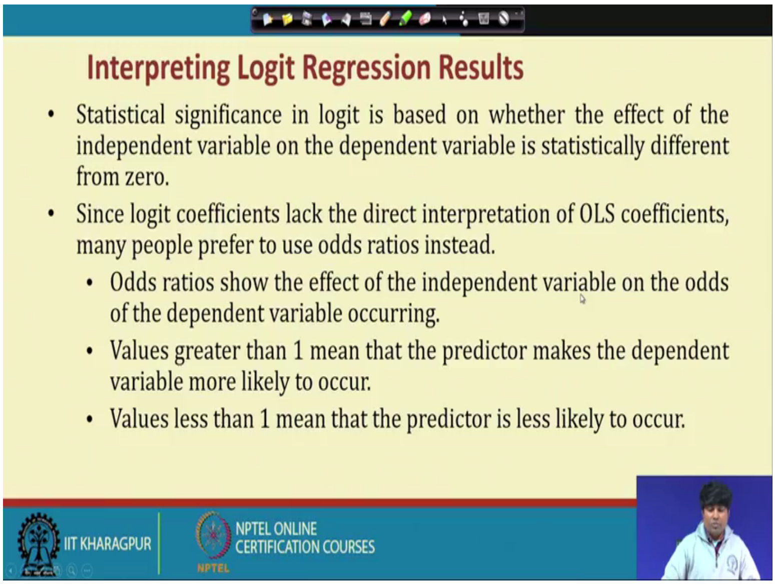
(Refer Slide Time: 08:04)



## LOGIT model predictions

So, we have to first check and then apply as per the requirement.

(Refer Slide Time: 08:08)



## Interpreting Logit Regression Results

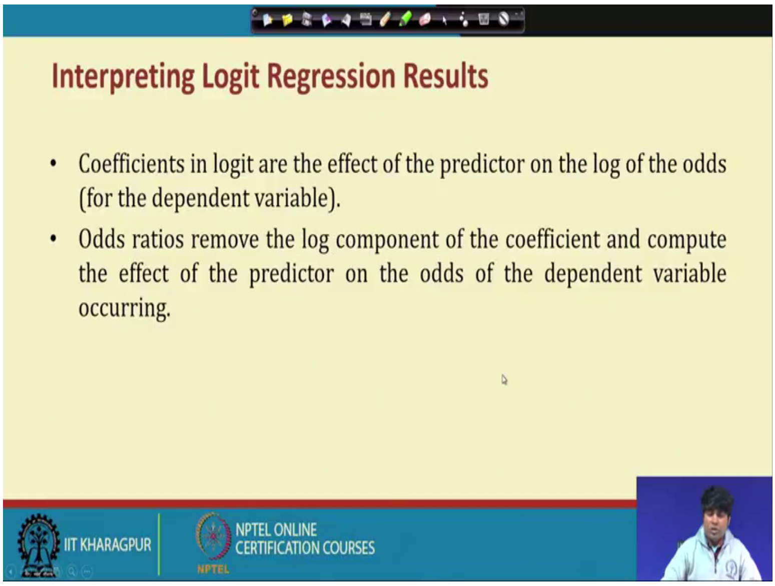
- Statistical significance in logit is based on whether the effect of the independent variable on the dependent variable is statistically different from zero.
- Since logit coefficients lack the direct interpretation of OLS coefficients, many people prefer to use odds ratios instead.
  - Odds ratios show the effect of the independent variable on the odds of the dependent variable occurring.
  - Values greater than 1 mean that the predictor makes the dependent variable more likely to occur.
  - Values less than 1 mean that the predictor is less likely to occur.

However, so interpretation of logistic regression model is not so straightforward, because the effect of a independent variable on dependent variable is (Refer Time: 08:24) means technically it is a similar kind of you know testing procedure, but the interpretation is slightly different directly cannot be use like you know a usual procedure of you know OLS estimations.

Here the kind of you know interpretation depends upon you know Odd ratio which is nothing, but p by 1 minus p. So, after doing the transformation when we will get the estimated values so the what we can do actually to get the odd ratio? It you find out you know e to the power that particular you know estimated coefficient, then that will give you the component of you know odd ratio.

So, having higher value of you know odd ratio particularly you know greater than 1. So, the prediction will be more likely to happens and if it is less than 1 it will be less likely to happen.

(Refer Slide Time: 09:15)



Interpreting Logit Regression Results

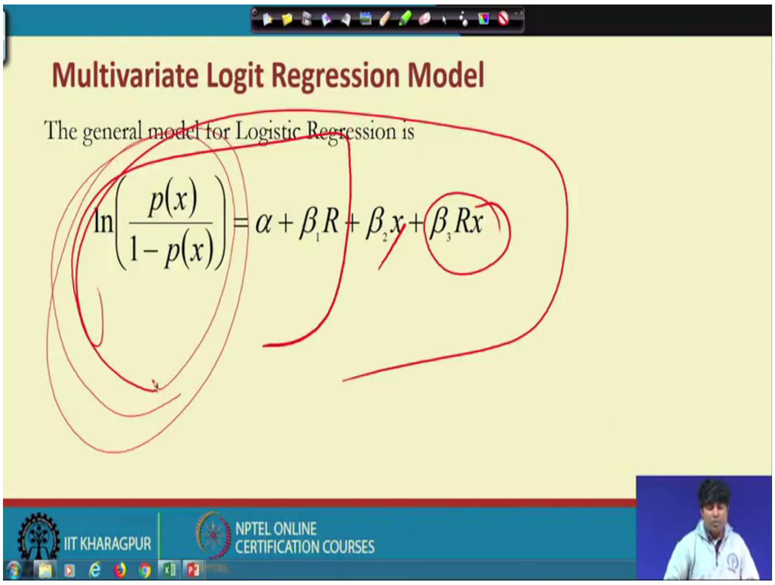
- Coefficients in logit are the effect of the predictor on the log of the odds (for the dependent variable).
- Odds ratios remove the log component of the coefficient and compute the effect of the predictor on the odds of the dependent variable occurring.

.

So, this is how the interpretation site. So, far as you know logistic regression is concerned. Otherwise the kind of you know structuring the kind of you know estimation process. So, what will you do actually in the excel you can do the or kind of you know operation and as usual you can estimate the a kind of you know models or else you can go to the particular you know software and give the option about this, but the typical models automatically you will get the estimated output.

So, that means, technically in the kind of you know logistic structure it is the odd ratio, which give you the model structure about the a kind of you know in means it is a kind of you know prediction structure and the kind of you know interpretation.

(Refer Slide Time: 09:45)



## Multivariate Logit Regression Model

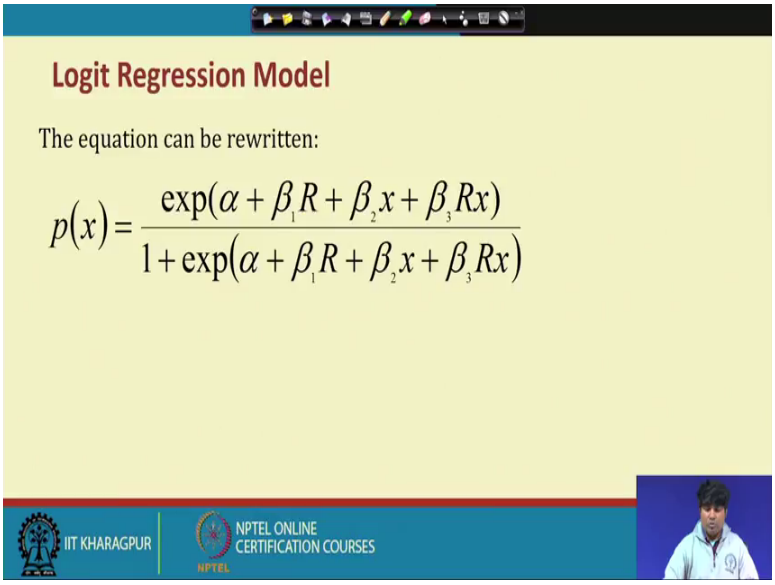
The general model for Logistic Regression is

$$\ln\left(\frac{p(x)}{1-p(x)}\right)=\alpha+\beta_1 R+\beta_2 x+\beta_3 Rx$$

And in the kind of you know multivariate structure. So, this is the same framework, but we can add one after another variable and in this context. So, this is actually 1 variables and followed by followed by a this is another variables and this is a interactive effect, but it can start with you know single variables or it can go with you know multiple variables. So, it is not an issue, but the structure is more or less same only part only you know difference is only the kind of you know right hand side right.

So, whatever a dependent variable information you have that will be converted into this particular, you know format then after that the estimation process will be more or less same and a while interpreting. So, you have to take care of the structure of you know what we called as you know odd ratio.

(Refer Slide Time: 10:45)



## Logit Regression Model

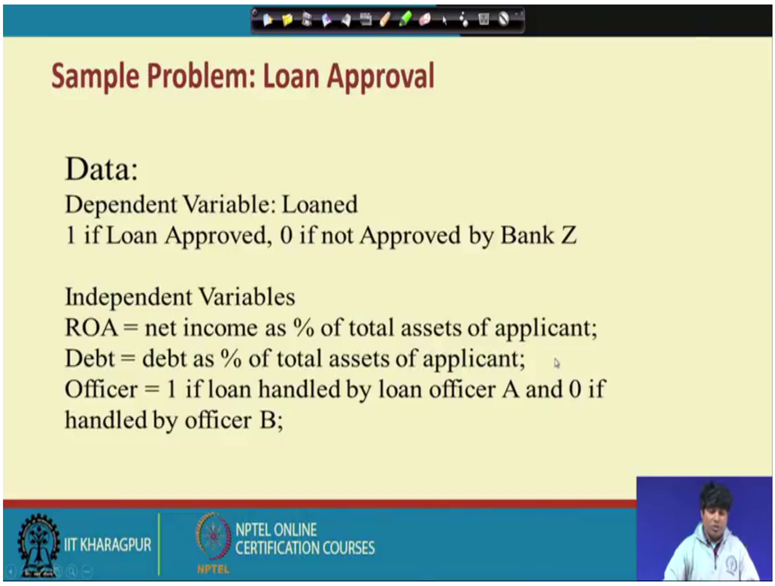
The equation can be rewritten:

$$p(x)=\frac{\exp(\alpha+\beta_1 R+\beta_2 x+\beta_3 Rx)}{1+\exp(\alpha+\beta_1 R+\beta_2 x+\beta_3 Rx)}$$

So, likewise, you can actually you know replace this one. So, this is actually what I have mentioned earlier the kind of you know logistic function and the same model which we can actually put in a different way where probability of x can be predicted with you know 1 plus 1 by e to the power minus z means; if you rearrange then the model will be it can be written like this.

So, this is what actually the kind of you know logistic structure.

(Refer Slide Time: 11:14)



## Sample Problem: Loan Approval

Data:
Dependent Variable: Loaned
1 if Loan Approved, 0 if not Approved by Bank Z

Independent Variables
ROA = net income as % of total assets of applicant;
Debt = debt as % of total assets of applicant;
Officer = 1 if loan handled by loan officer A and 0 if handled by officer B;

And let me highlight with you a simple example and this is a kind of you know example. Here you know bank loan problem you know issue and when somebody will apply for the loan. So, it may be approved it may not be approved.

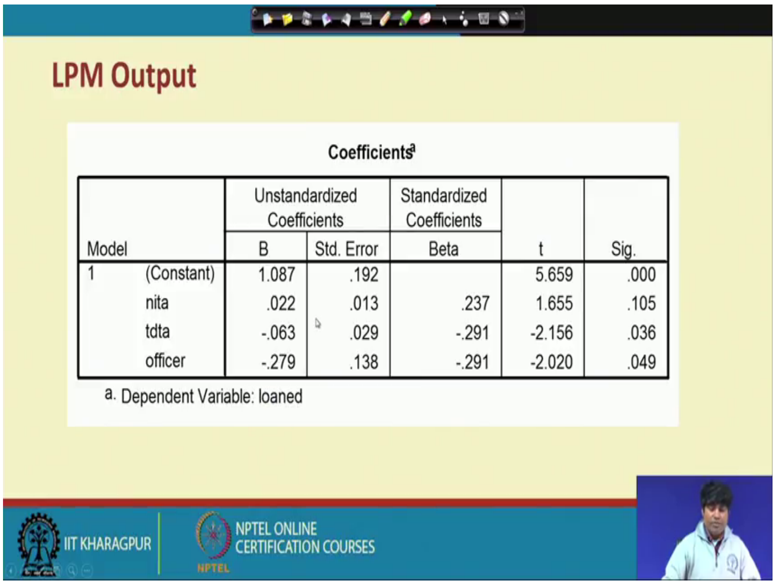
So, now, we have a spreadsheet and where's we have a plenty of you know data out of which some are you know approved case some are not approved case. And it is actually predicted through couple of you know independent variables. And your dependent variable is actually loan approved or not approved that is a yes no type of things.

So, if it is approved then yes will put 1 and no case we will put 0 and it is predicted with independent variable that is return on assets net income as a percent of total assets of a particular applicant and debt as a percent of percentage of total assets of that applicant and officers who is in charge for handling this kind of you know loans.

So, this is a what actually the kind of you know problem background and then again we will see the kind of you know output compared to the kind of you know linear probability model and the logit model.

And this is what the linear probability model output and as usual actually since with respect to this you know we have a spread sheet just you have to connect as usual you know regression structure.

(Refer Slide Time: 12:39)



LPM Output

Coefficients[a]

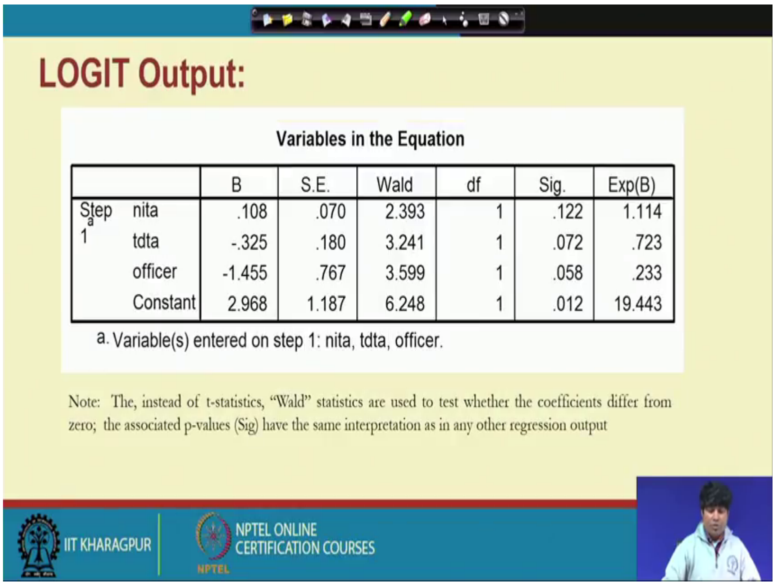
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 1.087 | .192 | | 5.659 | .000 |
| | nita | .022 | .013 | .237 | 1.655 | .105 |
| | tdta | -.063 | .029 | -.291 | -2.156 | .036 |
| | officer | -.279 | .138 | -.291 | -2.020 | .049 |

a. Dependent Variable: loaned

So, by default you will get the output and here’s the you know on the kind of you know assets is coming actually positive and the debt part is coming negative and the officer in charge is coming actually negative.

So, as a result, but in you know you know, but most of the cases it is coming actually significant so that means; so this is you know slightly to go for the kind of you know prediction.

(Refer Slide Time: 13:06)



LOGIT Output:

Variables in the Equation

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | nita | .108 | .070 | 2.393 | 1 | .122 | 1.114 |
| | tdta | -.325 | .180 | 3.241 | 1 | .072 | .723 |
| | officer | -1.455 | .767 | 3.599 | 1 | .058 | .233 |
| | Constant | 2.968 | 1.187 | 6.248 | 1 | .012 | 19.443 |

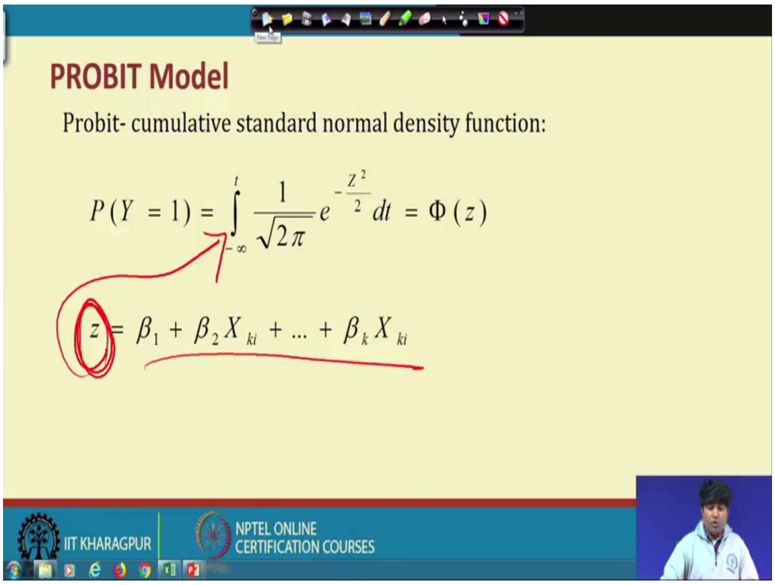
a. Variable(s) entered on step 1: nita, tdta, officer.

Note: The, instead of t-statistics, “Wald” statistics are used to test whether the coefficients differ from zero; the associated p-values (Sig) have the same interpretation as in any other regression output

Now, compared to linear probability models so the logit model output will be like this again the if you check the kind of you know sign of the particular you know coefficients, if whether you use you know linear faulty model or you know logit model. So, the sign is actually more or less same. So, k whatever results are coming here this said the results are also coming in a similar way only thing here we have to actually with respect to these estimates. So, we have to find out the you know e value of these particular estimates. So, that you know this will give you the kind of you know odd ratio and the order ratio will be finally, useful used for the kind of you know prediction.

So; that means, in this case it has a high impact and this will be followed by second highest impact and this will be followed by third highest impart.

(Refer Slide Time: 13:56)



## PROBIT Model

Probit- cumulative standard normal density function:

$$P(Y=1) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dt = \Phi(z)$$

$$z = \beta_1 + \beta_2 X_{ki} + \ldots + \beta_k X_{ki}$$

So, likewise actually logit model can be frequently used to predict the kind of you know situation and to solve some of the problems, where the dependent variable structure is a kind of you know categorical or the kind of you know means; typically binary in nature it is a 0 1 kind of you know situation, but now compared to actually linear probability model and binary a sorry logit model what is happening in the case of linear probability model? So, the framework requires actually either 0 and 1.

So, yes and no, but you know in the case of you know logit model. So, the dependent side component requirement is log p upon 1 minus p that is the odd ratio. So, now, what is happening you know?

So, while doing this a kind of you know sampling and the kind of you know data analysis be it be careful how you have to deal with this particular you know conversion otherwise? If you put simply 0 and 1 then the odd ratio might give you the value of you 0 and infinite. Which can give some kind of you know problem in the estimation process, which will be address with I mean we will address here with a typical you know examples in the meantime.

So, the third model is called as you know probit model and in this case it will be followed by normal density functions and the particular model is like this. And if you compare this you know correct corresponding to the you know logit model.

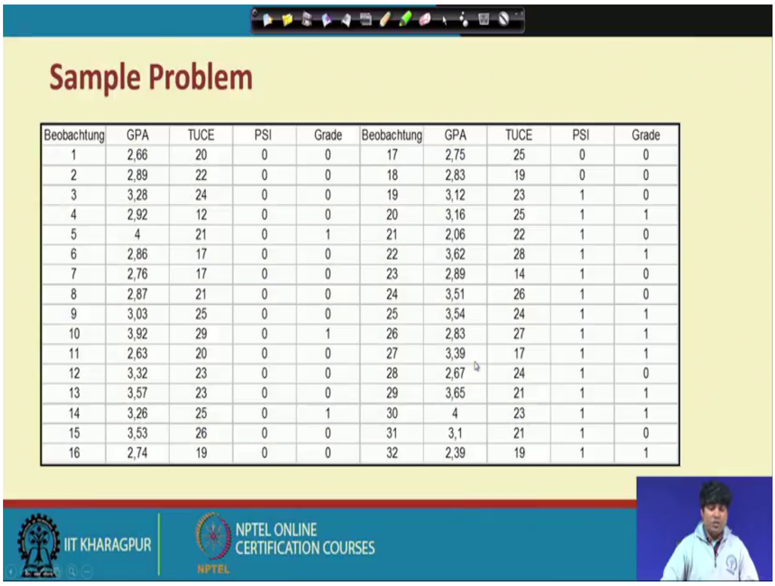
So, it is more or less sense again the right hand side is the z component, which is directly followed by this you know normal density functions or that that will be derived from the normal institution you know structures. Once you get this particular you know structure then you have to connect with the as per the you know problem requirement or the kind of you know prediction requirement.

So, if you go through go through the kind of you know kind of you know structure. So, you will find in the case of you know let us see here, this is actually the this is actually linear probability LPM structure and then this is the kind of you know logistic structures and then finally, it is the kind of you know kind of you know of probit structures.

So, every case actually the right hand side part is more or less you know similar kind of inner structure only thing is in the left hand side. So, the you know the information's will be transferred into different kind of you know structure, 1 which is simply 0 1 structure, another 1 is with you know with respect to odd ratio and third one is with respect to you know z transformations.

Then after that the estimation process will be more or less same. We can actually connect the probit model with your standard examples and in this case.

(Refer Slide Time: 16:49)

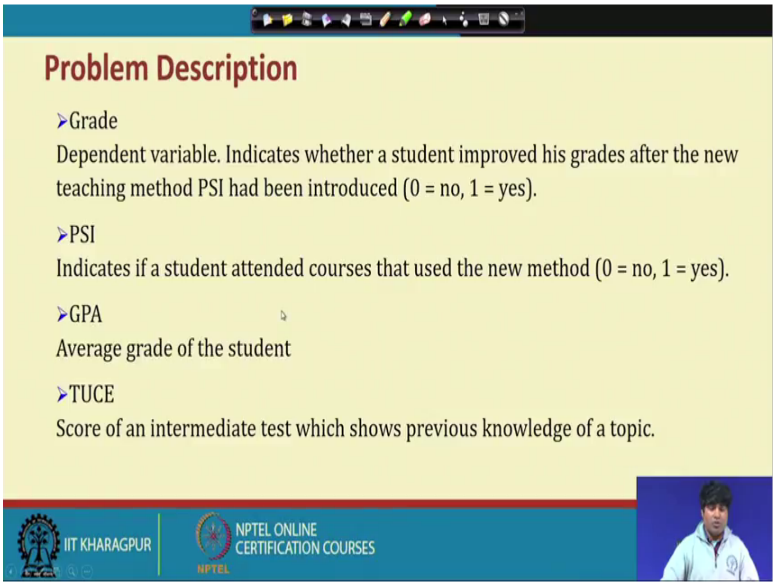


Sample Problem

| Beobachtung | GPA | TUCE | PSI | Grade | Beobachtung | GPA | TUCE | PSI | Grade |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2,66 | 20 | 0 | 0 | 17 | 2,75 | 25 | 0 | 0 |
| 2 | 2,89 | 22 | 0 | 0 | 18 | 2,83 | 19 | 0 | 0 |
| 3 | 3,28 | 24 | 0 | 0 | 19 | 3,12 | 23 | 1 | 0 |
| 4 | 2,92 | 12 | 0 | 0 | 20 | 3,16 | 25 | 1 | 1 |
| 5 | 4 | 21 | 0 | 1 | 21 | 2,06 | 22 | 1 | 0 |
| 6 | 2,86 | 17 | 0 | 0 | 22 | 3,62 | 28 | 1 | 1 |
| 7 | 2,76 | 17 | 0 | 0 | 23 | 2,89 | 14 | 1 | 0 |
| 8 | 2,87 | 21 | 0 | 0 | 24 | 3,51 | 26 | 1 | 0 |
| 9 | 3,03 | 25 | 0 | 0 | 25 | 3,54 | 24 | 1 | 1 |
| 10 | 3,92 | 29 | 0 | 1 | 26 | 2,83 | 27 | 1 | 1 |
| 11 | 2,63 | 20 | 0 | 0 | 27 | 3,39 | 17 | 1 | 1 |
| 12 | 3,32 | 23 | 0 | 0 | 28 | 2,67 | 24 | 1 | 0 |
| 13 | 3,57 | 23 | 0 | 0 | 29 | 3,65 | 21 | 1 | 1 |
| 14 | 3,26 | 25 | 0 | 1 | 30 | 4 | 23 | 1 | 1 |
| 15 | 3,53 | 26 | 0 | 0 | 31 | 3,1 | 21 | 1 | 0 |
| 16 | 2,74 | 19 | 0 | 0 | 32 | 2,39 | 19 | 1 | 1 |

So, here the spreadsheets show that you know great change happenings with respect to 3 different indicators. So, the problem understands you know the problem or structure will be like this.

(Refer Slide Time: 17:00)



## Problem Description

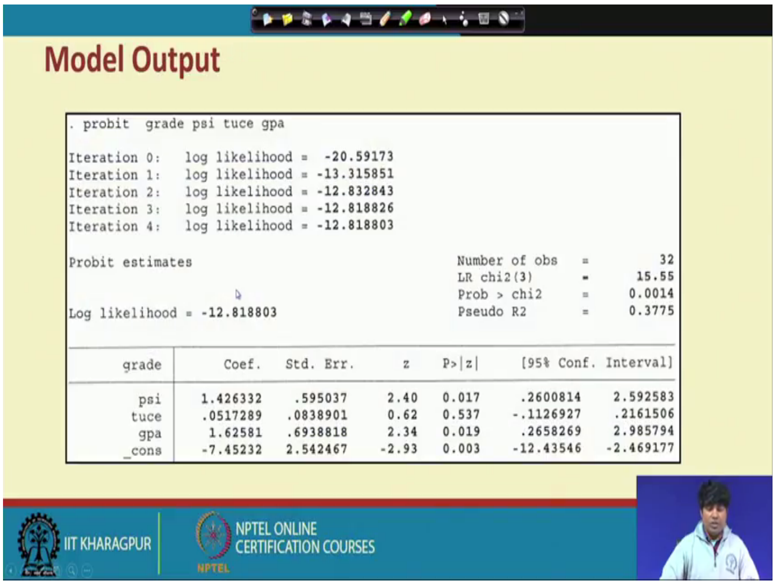
- Grade
  Dependent variable. Indicates whether a student improved his grades after the new teaching method PSI had been introduced (0 = no, 1 = yes).
- PSI
  Indicates if a student attended courses that used the new method (0 = no, 1 = yes).
- GPA
  Average grade of the student
- TUCE
  Score of an intermediate test which shows previous knowledge of a topic.

.

So, the dependent variable indicate indicates whether a student improved his grade after the new teaching method PSI has been introduced. And it will be having actually yes no kind of you know situation. And it is reflected by or predicted by 3 indicators PSI indicates if a student, you know attended cross courses that using the new method.

And the GPA grade point you know average of the students and then this score of an intermediate test with such previous knowledge of the topic. So, these are the 3 variables and through which we have to predict the grade, whether it is change after this particular you know new mechanism or not.

(Refer Slide Time: 17:44)



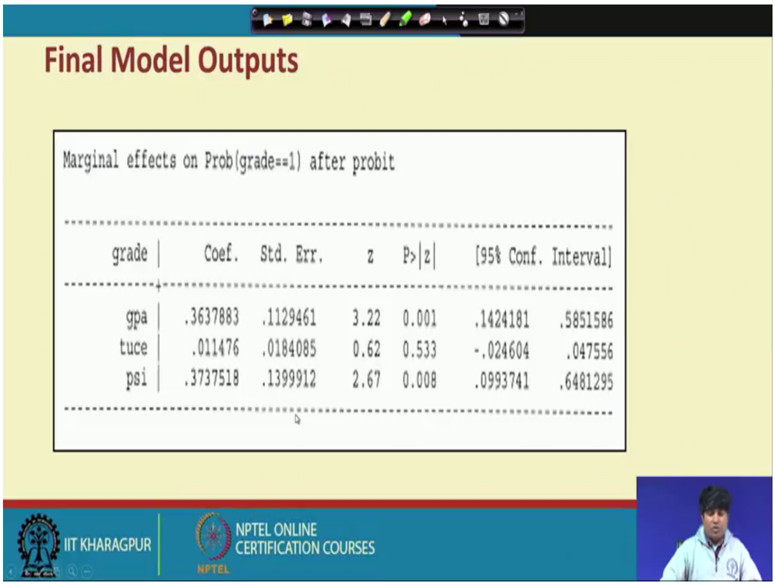
Model Output

```
. probit  grade psi tuce gpa

Iteration 0:   log likelihood =  -20.59173
Iteration 1:   log likelihood = -13.315851
Iteration 2:   log likelihood = -12.832843
Iteration 3:   log likelihood = -12.818826
Iteration 4:   log likelihood = -12.818803

Probit estimates                                  Number of obs   =         32
                                                  LR chi2(3)      =      15.55
                                                  Prob > chi2     =     0.0014
Log likelihood = -12.818803                       Pseudo R2       =     0.3775

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         psi |   1.426332    .595037     2.40   0.017     .2600814    2.592583
        tuce |   .0517289   .0838901     0.62   0.537    -.1126927    .2161506
         gpa |    1.62581   .6938818     2.34   0.019     .2658269    2.985794
       _cons |   -7.45232   2.542467    -2.93   0.003    -12.43546   -2.469177
```

.

So, accordingly after the estimations, so, the as usual you know like logit models. So, the probit model result is also coming similar kind of you know structures. So, here's most of the variables are coming in a significant impact for this you know grade change.

(Refer Slide Time: 17:59)



Final Model Outputs

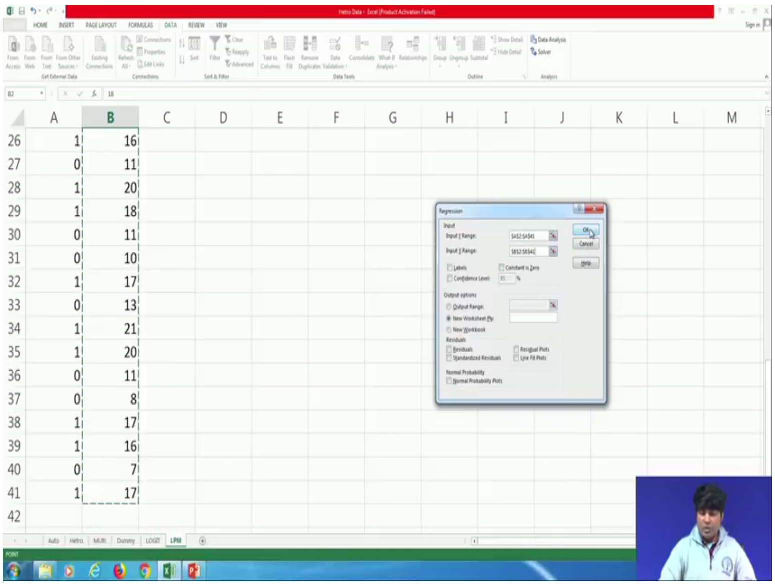
```
Marginal effects on Prob(grade==1) after probit

---------------------------------------------------------------------------
    grade |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+----------------------------------------------------------------
      gpa |  .3637883   .1129461     3.22   0.001     .1424181    .5851586
     tuce |   .011476   .0184085     0.62   0.533     -.024604     .047556
      psi |  .3737518   .1399912     2.67   0.008     .0993741    .6481295
---------------------------------------------------------------------------
```

.

Again you know it is you know with you know standardized framework. So, the model output is coming a little bit more attractive way. So, what is happening here's? You know all are having you know a positive signal to the grade change.

So, these are the various you know what I have done actually. So, I have connected all these 3 models with standard examples. Let me show you the typical you know structure through how it actually works in the real life scenario.

(Refer Slide Time: 18:27)



|  | A | B |
|---|---|---|
| 26 | 1 | 16 |
| 27 | 0 | 11 |
| 28 | 1 | 20 |
| 29 | 1 | 18 |
| 30 | 0 | 11 |
| 31 | 0 | 10 |
| 32 | 1 | 17 |
| 33 | 0 | 13 |
| 34 | 1 | 21 |
| 35 | 1 | 20 |
| 36 | 0 | 11 |
| 37 | 0 | 8 |
| 38 | 1 | 17 |
| 39 | 1 | 16 |
| 40 | 0 | 7 |
| 41 | 1 | 17 |
| 42 |  |  |

Regression

Input

Input Y Range: $A$2:$A$41

Input X Range: $B$2:$B$41

□ Labels □ Constant is Zero

□ Confidence Level 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

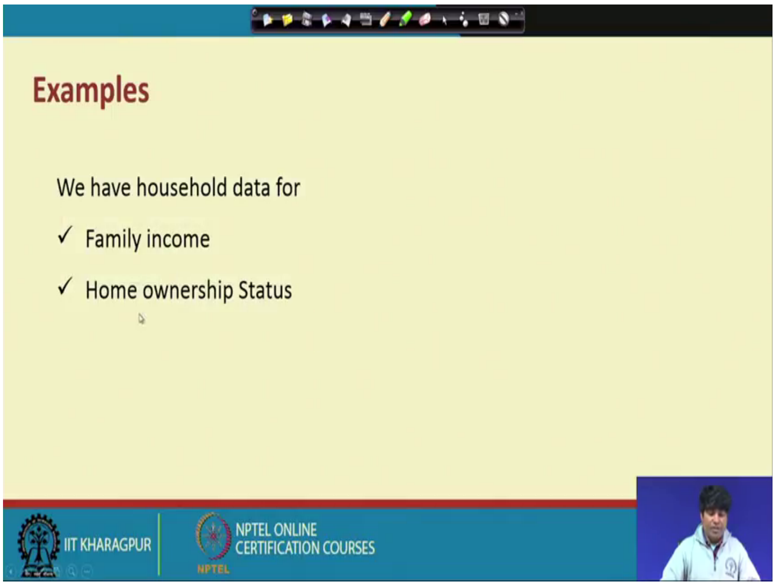
Normal Probability

□ Normal Probability Plots

OK Cancel Help

So, we start with a first linear probability model and in this in this particular you know models let me take this particular you know example and the example is here.

(Refer Slide Time: 18:42)



## Examples

We have household data for

- ✓ Family income
- ✓ Home ownership Status

.

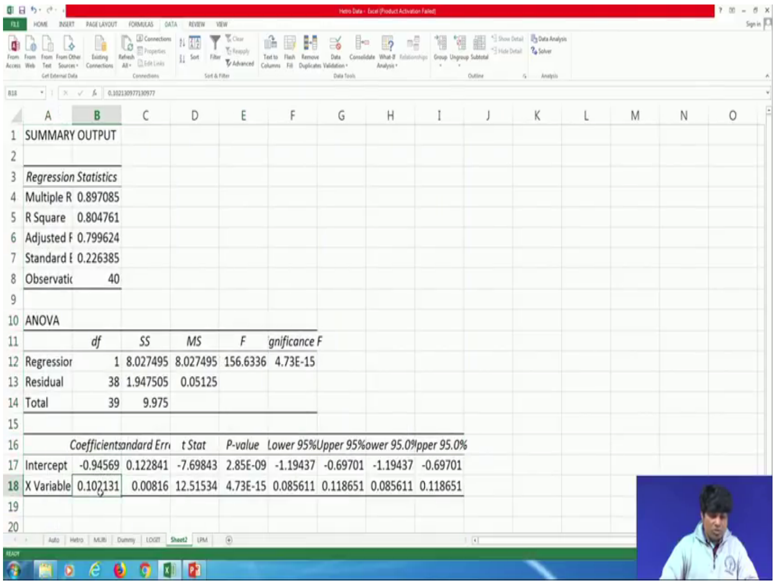
So, family income and home ownership status so; that means, we are trying to solve a problem where homeownership status is your dependent variable and it is actually predicted with respect to family income. So, that means, the literarily understanding is that. So, if your income is very high then you may have you know ownership you can you know home ownership and if your income is a law then you may not have actually a home ownership. So, that is so what will you do will do the you know predictions or the we will address this particular problem, with respect to some you know and that offends.

Let us say. So, start with you know linear probability model and y is represented here actually people have been their house, ownership status and their income you know income. So, then this is a first individual. So, he has no ownership status and his income is 8 lakhs for a month. And against one indicates that you know he has a family ownership status and his income is actually 16 lakhs. So, again so the third one is having ownership status and income is having 18 lakh.

So, likewise we have a plenty of you know data and what we like to you know check here that you know, whether income is a factor to justify that you know people having there you know ownership status or not then again you go to the data analysis package and choose the regression and finally. So, here actually you remove or clean the kind of you know entry early entry and then you first give the dependent variable structure in input and after doing this is actually independent variable structure.

So, let us we start with you know dependent variable cost and this will be give you the indication about the dependent variable. So, allow up to forty one and then we go to the dependent variables and dependent variable is again b to 2 to b 41. So, same sample size and then you put. So, this this gives the idea about the kind of you know linear probability estimation see here.

(Refer Slide Time: 20:55)

.

SUMMARY OUTPUT

| Regression Statistics | |
|---|---|
| Multiple R | 0.897085 |
| R Square | 0.804761 |
| Adjusted F | 0.799624 |
| Standard E | 0.226385 |
| Observatic | 40 |

ANOVA

| | df | SS | MS | F | gnificance F |
|---|---|---|---|---|---|
| Regressior | 1 | 8.027495 | 8.027495 | 156.6336 | 4.73E-15 |
| Residual | 38 | 1.947505 | 0.05125 | | |
| Total | 39 | 9.975 | | | |

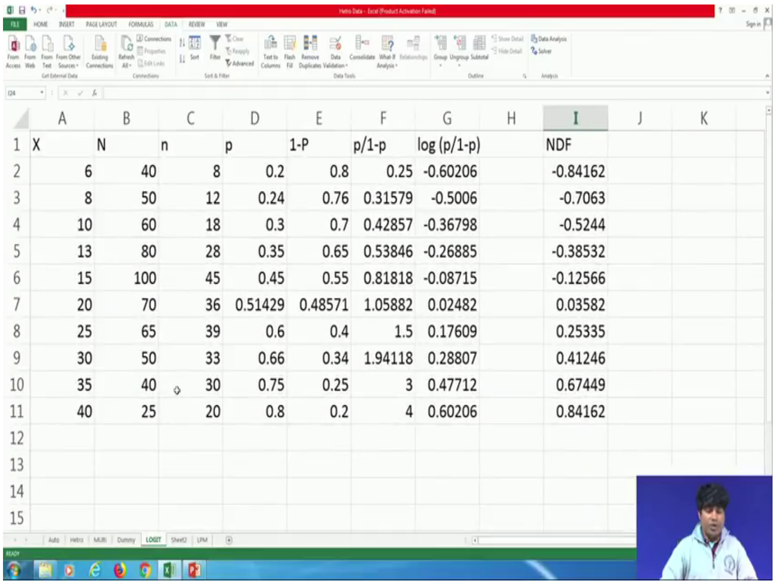
| | Coefficients | andard Erri | t Stat | P-value | Lower 95% | Upper 95% | ower 95.0% | pper 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | -0.94569 | 0.122841 | -7.69843 | 2.85E-09 | -1.19437 | -0.69701 | -1.19437 | -0.69701 |
| X Variable | 0.102131 | 0.00816 | 12.51534 | 4.73E-15 | 0.085611 | 0.118651 | 0.085611 | 0.118651 |

And as usual it is a standard procedure. So, intercept is coming negative and this is coming actually positive. So that means; family you know family income is a positive factor through, which you know income you know people having the ownership status housing ownership status. So, that means, the way which we are actually expecting earlier. So, these data is actually validating the same thing to linear quality model.

So, it is showing actually the relationship positively a linked and it is statistically also highly significant and R square is also having actually, a huge kind of you know support that is the fitness measures and it is followed by f statistic that is the anova.

So, that means, technically. So, corresponding to the theoretical background and the kind of you know choice of the model the model output is actually very useful for predicting this kind of you know business problem. And that too come with you know some fantastic management decision. And this is similar kind of you know our problem which we can actually address through other issues that is the kind of you know that is the kind of you know logit model and probit model.

(Refer Slide Time: 22:11)



| X | N | n | p | 1-P | p/1-p | log (p/1-p) | | NDF |
|---|---|---|---|---|---|---|---|---|
| 6 | 40 | 8 | 0.2 | 0.8 | 0.25 | -0.60206 | | -0.84162 |
| 8 | 50 | 12 | 0.24 | 0.76 | 0.31579 | -0.5006 | | -0.7063 |
| 10 | 60 | 18 | 0.3 | 0.7 | 0.42857 | -0.36798 | | -0.5244 |
| 13 | 80 | 28 | 0.35 | 0.65 | 0.53846 | -0.26885 | | -0.38532 |
| 15 | 100 | 45 | 0.45 | 0.55 | 0.81818 | -0.08715 | | -0.12566 |
| 20 | 70 | 36 | 0.51429 | 0.48571 | 1.05882 | 0.02482 | | 0.03582 |
| 25 | 65 | 39 | 0.6 | 0.4 | 1.5 | 0.17609 | | 0.25335 |
| 30 | 50 | 33 | 0.66 | 0.34 | 1.94118 | 0.28807 | | 0.41246 |
| 35 | 40 | 30 | 0.75 | 0.25 | 3 | 0.47712 | | 0.67449 |
| 40 | 25 | 20 | 0.8 | 0.2 | 4 | 0.60206 | | 0.84162 |

So, in the case of you know logit model and probit model. So, our mod modelling structure is a like this. So, compared to previous ones let us see here.

So, in the case of you know linear probability model. So, this is your you know modelling structures. So, which you have already tested, but now in the case of you know logit. So, our right hand transformation will be like this. So, log of you know odd ratio that is P by 1 minus P and agains in the case of you know probit model.

So, we need actually z transformation that is in the form of this particular you know structure. So, that means, compared to linear probability model. So, logistic model and probit model need further kind of you know additional transformations, before you start the estimation process.

The some in fact, if you use actually standard software like R stata SPSS, then automatically the model is there you can connect the software automatically help you to give the estimation, but if you use the spreadsheet then by default what you will do you have to do yourself manually, but a what is happening here’s it compare to the previous case.

So, it is the game between 0 and 1. So, go through the linear probability model and now this is actually 0 1 0 1.

So, now this is this is actually a probability p means 1 means then, 1 minus p equal to 0. So, that means, the odd ratio will be very complicated so for instance if you if you go to the logistic model. So, this one so when you put actually P equal to 1 and then by default 1 minus P equal to 0. So, by default this will give you infinite value.

So, you are not in a position to estimate, but when you put actually P equal to 0 by default this is actually 0 by 0. So, log 0. So, this is again actually having you know kind of you know problem.

So, as a result so, what is happening? So, in this kind of you know data set you cannot actually directly apply logistic model and you know probit model. So, for that so we need actually the concept called as a you know group sampling so; that means, we will rearrange the particular, you know structure were x is having actually income and that starts with you know 6 lakhs, 8 lakhs, 10 lakhs, 13 lakhs, 15 lakhs and up to 40 lakhs.

And then what is the group sampling concept that you know we like to you know pick up you know 40 different family who is having actually 6 lakhs annual income and out of which then we may ask everybody, whether they have the ownership you know housing ownership or not.

Then you know in this particular sample out of forty samples. So, 8 are having actually yes yes of sons. So, again with respect to 8 lakhs category so we surveyed 50 50 samples and then ownership status is having 12; likewise you know 10 lakhs then 60 60 family out of which 18 18 18 families are having ownership status so as a result.

So, now, the p value will be the ratio between N upon n. So, small n to capital N. So, as a result in this case so what we have done here? So, this is the actual actual data and then from modelling will start from here. So, we first calculate the p value and which is a ratio between small n to capital N the number of happenings with respect to total number of cases and as a results you define the probability values here.

Now, compared to linear probability models so, logit model and probit model case. So, it will be not a 0 and 1 structure, but it will be in between 0 and 1 structure you see here in the case of you know P value. So, all these all these you know probability values are in between 0 to 1 non 0 and no one's.

If only 0 and 1 then the right choice to predict the kind of you know problem is linear through linear probability model, otherwise you can choose here by logistic structure or you know the same problems, if you like to connect then you know you have to go for you know data adjustment like what we have done here?

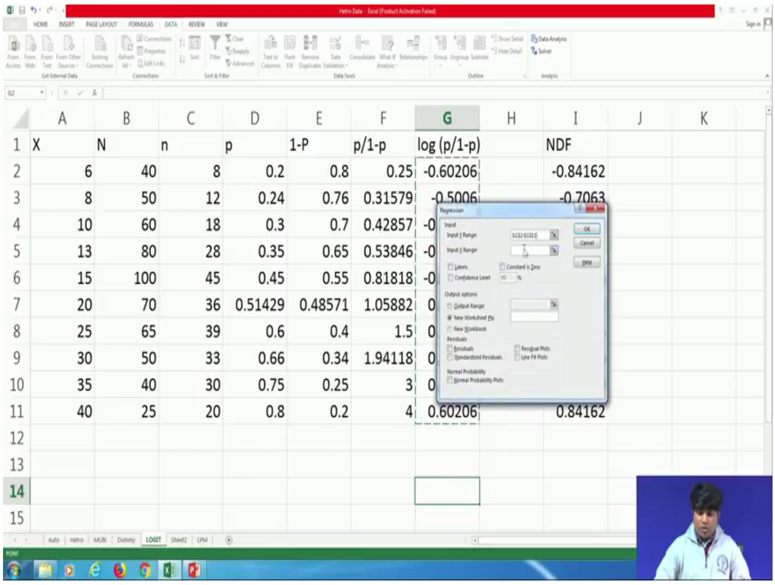
Now, what is happening here? So, after calculating P so we need actually 1 minus P, that is the probability of failure case. So, 1 minus P is the case of you know. So, whatever p will be find you have to here. So, 1 minus P is nothing, but 1 minus you know 1 minus P will give you the kind of you know failure case.

So, P and then 1 minus P that is nothing, but 1 minus probability of success. So, this is probability of success and this is probably of failure so; that means, it is the yes case and it is the no case and that is the ratio will give you the odd ratio. So, what will you do p by 1 minus p will give you the odd ratio.

Then finally, we calculate actually log p by 1 minus p, log p by 1 minus p. So, that means, after getting the p a p by 1 minus p. So, you calculate actually log p by p by 1 minus p.

So, now finally so, far as estimation is concerned go to the data analysis then again you choose regressions and now what is happening? So, you clean the kind of you know entry early entry and then here the dependent variable is this ones that is log p by 1 minus p. So, you put these options. So, this is this is here only.

(Refer Slide Time: 27:53)


.

| X | N | n | p | 1-P | p/1-p | log (p/1-p) | | NDF |
|---|---|---|---|---|---|---|---|---|
| 6 | 40 | 8 | 0.2 | 0.8 | 0.25 | -0.60206 | | -0.84162 |
| 8 | 50 | 12 | 0.24 | 0.76 | 0.31579 | -0.5006 | | -0.7063 |
| 10 | 60 | 18 | 0.3 | 0.7 | 0.42857 | -0 | | |
| 13 | 80 | 28 | 0.35 | 0.65 | 0.53846 | -0 | | |
| 15 | 100 | 45 | 0.45 | 0.55 | 0.81818 | -0 | | |
| 20 | 70 | 36 | 0.51429 | 0.48571 | 1.05882 | 0 | | |
| 25 | 65 | 39 | 0.6 | 0.4 | 1.5 | 0 | | |
| 30 | 50 | 33 | 0.66 | 0.34 | 1.94118 | 0 | | |
| 35 | 40 | 30 | 0.75 | 0.25 | 3 | 0 | | |
| 40 | 25 | 20 | 0.8 | 0.2 | 4 | 0.60206 | | 0.84162 |

Regression

Input

Input Y Range: $G$2:$G$11

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level 95 %

Output options

○ Output Range

● New Worksheet Ply

○ New Workbook

Residuals

□ Residuals □ Residual Plots

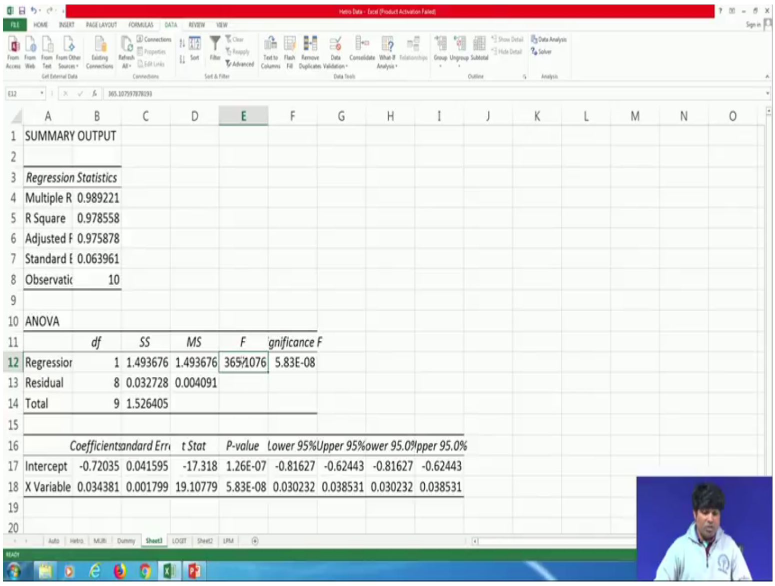
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

So, now you enter these options. Then finally, you give the indication about the independent variable structure and then as usual you can actually click. So, you will get the output here.

(Refer Slide Time: 28:04)


.

SUMMARY OUTPUT

| Regression Statistics | |
|---|---|
| Multiple R | 0.989221 |
| R Square | 0.978558 |
| Adjusted F | 0.975878 |
| Standard E | 0.063961 |
| Observatic | 10 |

ANOVA

| | df | SS | MS | F | gnificance F |
|---|---|---|---|---|---|
| Regressior | 1 | 1.493676 | 1.493676 | 365.1076 | 5.83E-08 |
| Residual | 8 | 0.032728 | 0.004091 | | |
| Total | 9 | 1.526405 | | | |

| | Coefficients | andard Err | t Stat | P-value | Lower 95% | Upper 95% | ower 95.0% | pper 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | -0.72035 | 0.041595 | -17.318 | 1.26E-07 | -0.81627 | -0.62443 | -0.81627 | -0.62443 |
| X Variable | 0.034381 | 0.001799 | 19.10779 | 5.83E-08 | 0.030232 | 0.038531 | 0.030232 | 0.038531 |

So, this is what actually and the model is actually very interesting. Now like the previous linear probability case LPM model case. So, it is also ninety percent impact and. In fact, F is a highly significant and the variable impact income impact on the ownership is also highly statistically significant and that too positively related.

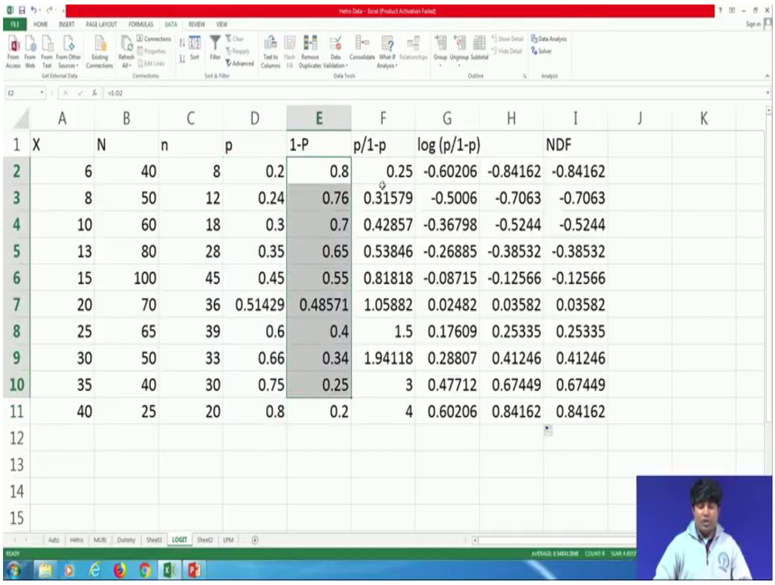
So, that means, compared to the previous 1 logit model is much much you know providing much better results for this particular you know problems. So, far as you know prediction is concerned.

Now again go to this particular you know structure. So, we have actually same problem we can also solve through the kind of you know probit model. So, but in the case of you know probit models our modelling structure is like this. So, we have we have a model actually with model. So, we need actually z transformation so; that means, technically. So, again you go to the excel sheet.

So, now we have here we have here we have here actually p by 1 minus p. So, what will you do? So, you again you go for you know z transformation. So, it is a simply a it will simply find out you know normal distribution sampling. And then you connect with the kind of you know p value here.

And then you will get the kind of you know entry this one. So, this will give you the kind of you know structure. So, in fact, actually normal distribution function this is a so what will it do? So, we whatever we p by 1 minus p we have calculated here. So, log of p by 1 minus p is the entry to the kind of you know logit model, but in the case of you know probit models. We need in normal density function and that to the a the kind of you know the normal kind of you know sample inverse functions. And then we have to we have already calculated here's and again what will you do.

(Refer Slide Time: 30:28)



=1-D2

| X | N | n | p | 1-P | p/1-p | log (p/1-p) | | NDF |
|---|---|---|---|---|---|---|---|---|
| 6 | 40 | 8 | 0.2 | 0.8 | 0.25 | -0.60206 | -0.84162 | -0.84162 |
| 8 | 50 | 12 | 0.24 | 0.76 | 0.31579 | -0.5006 | -0.7063 | -0.7063 |
| 10 | 60 | 18 | 0.3 | 0.7 | 0.42857 | -0.36798 | -0.5244 | -0.5244 |
| 13 | 80 | 28 | 0.35 | 0.65 | 0.53846 | -0.26885 | -0.38532 | -0.38532 |
| 15 | 100 | 45 | 0.45 | 0.55 | 0.81818 | -0.08715 | -0.12566 | -0.12566 |
| 20 | 70 | 36 | 0.51429 | 0.48571 | 1.05882 | 0.02482 | 0.03582 | 0.03582 |
| 25 | 65 | 39 | 0.6 | 0.4 | 1.5 | 0.17609 | 0.25335 | 0.25335 |
| 30 | 50 | 33 | 0.66 | 0.34 | 1.94118 | 0.28807 | 0.41246 | 0.41246 |
| 35 | 40 | 30 | 0.75 | 0.25 | 3 | 0.47712 | 0.67449 | 0.67449 |
| 40 | 25 | 20 | 0.8 | 0.2 | 4 | 0.60206 | 0.84162 | 0.84162 |

.

So, you can just once again, I will show you normal sample distribution. So, this one's and then you connect with the a particular you know sample and then this is.

Yes now it is coming. So, then you will just scroll it. So, this is this will be generated and cause sorry we have made wrong entry. So, corresponding to this one so what will you do? So, our entries actually with respect to P we need not require actually odd ratio that is why it is showing wrong? So, what will you do? So, you will go to the normal distribution again.

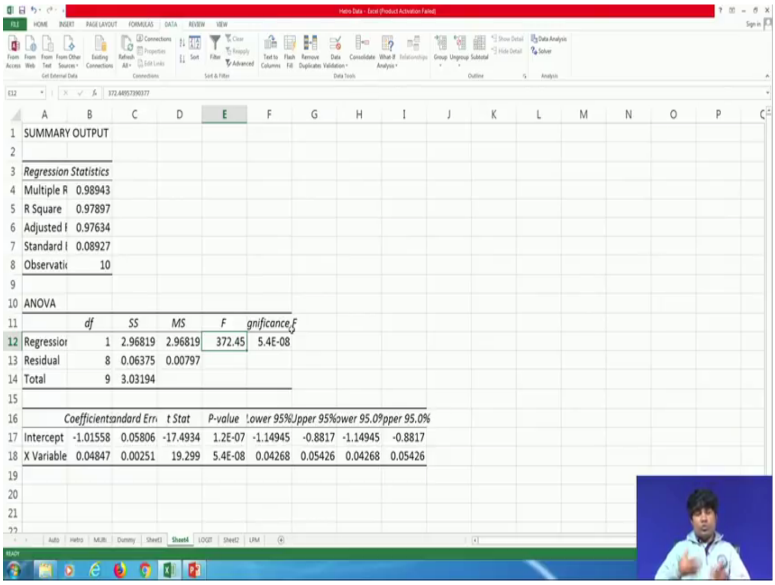
So, then the entry will be with respect to probability of success on D. So, that will give you the entry yes, now it is coming yes this is the; that means, technically if you choose the wrong row. So, it will give you wrong prediction.

So, that means, you see here. So, our main variable is you know independent variable is X and then it will be followed by the dependent variable structure n up N upon n that is the probability that the probability of success. And for logistic logit model. So, we need actually 1 minus p then again p by 1 minus p and again we will find log p upon 1 minus p.

And then finally, the estimation process will follow log p upon 1 minus p with x, but in the case of normal days you know probit model. So, first we have to calculate the probability of you know success. So, that is n upon n then we have to transfer this you know probability into normal density function and for that we have already done this transformation again and after doing this one. So, this is actually generated and then finally, you go to the data analysis and in the data analysis. So, you again connect with this dependent variables. So, first you clean this a previous entry and then the dependent variables.

So, this is again a wrong entry. So, what will go to the you go to the dependent variable structure. So, this way you have to enter. So, this is dependent variable and then you connect with you know independent variable. So, now, it is ok.

(Refer Slide Time: 33:08)


.

SUMMARY OUTPUT

| Regression Statistics | |
|---|---|
| Multiple R | 0.98943 |
| R Square | 0.97897 |
| Adjusted I | 0.97634 |
| Standard I | 0.08927 |
| Observati | 10 |

ANOVA

| | df | SS | MS | F | gnificance F |
|---|---|---|---|---|---|
| Regressior | 1 | 2.96819 | 2.96819 | 372.45 | 5.4E-08 |
| Residual | 8 | 0.06375 | 0.00797 | | |
| Total | 9 | 3.03194 | | | |

| | Coefficients | andard Err | t Stat | P-value | Lower 95% | Upper 95% | ower 95.0% | pper 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | -1.01558 | 0.05806 | -17.4934 | 1.2E-07 | -1.14945 | -0.8817 | -1.14945 | -0.8817 |
| X Variable | 0.04847 | 0.00251 | 19.299 | 5.4E-08 | 0.04268 | 0.05426 | 0.04268 | 0.05426 |

So, this is how the a probit model results right again; again compared to the previous case. So, the model R square is a 90 a 99 percent and again X is a highly significant and that to 19 and then F is highly significant.

So, now what is happening? So, the problem is actually to predict people having the ownership or not with respect to you know their family income. So, we tested this particular you know problem through linear poverty model, logit model and probit model.

In the case of you know linear probability models we need simply the family you know information, whether they have ownership or not. So, it is a kind of you know 0 1 status and then their income levels and the usual estimation process is very simple, because the 0 1 is the requirement of you know linear probability model, but in the case of you know logit model. So, the same sampling cannot work, because we need to calculate the odd ratio or you know z transformation for logic and probit respectively.

So, in that case to get the odd ratio so or the kind of you know probability we have to go for you know group sampling and that too we have to first fix the family income. And then with you know with a particular you know income level.

So, we will check you know number of you know family out of is how many having their in ownership then in that context we can find out the probability. And once you get the probability, we can find out you know the that is what the probability of you know success; that means, they have the ownership. And then we will get you know failure site 1 minus p and finally, you can find out the odd ratio by doing the p by 1 minus p and then we connect with the logit model log of p by 1 minus p with the income you know factors.

Similarly, in the case of you know probit model. So, we have to first calculate the probability of success by the you know group sampling again n by N. And then we transfer this p value with you know normal density function and then we will get the z structure which we have already highlighted here.

So, this is the this is the entry of you know logic logistic, you this is the entry for a logit model that too dependent variable and this is the entry for the kind of you know probit model and that too for dependent variable and every case x is the kind of you know independent variables.

And as a result what is happening compared to 3 models. So, every time there is a kind of you know improvement. So, this is the this is the kind of you know logit model status and then this is the kind of you know probit model status, and we have already discussed the linear probability model status and with respect to these samples.

So, now what is happening? So, we will just compare with you know logit model, with the probit model, because we are using the same data structure for logit and probit model that is the group sampling.

And in this case what is happening happening in the case of logit logit model. So, it is showing actually 99, F is coming 365 and then T is coming 19.1. Now coming to probit model so, it is actually a drastically improved instead of you know F value 365 it is now moving to 372. And instead of actually T value earlier case it was 19.1 1. So, now, it is actually having 19.3 altogether. And again R square is more or less you know same, but still you know F value is coming high and T value is coming very high.

So; that means, what is actually happening. So, the same problems is invest you know you can investigate through linear probability structure logit structure and probit structure, but there is a high degree of you know prediction you know improvement.

So, that means, if you compare all these 3 models finally, probit model output is coming more interesting and much better than compared to linear probability model and the logit model, but it is not always true that you know every time the probit model is more more you know attractive to solve the problem. Sometimes say sometimes the probit model in logit model can give better results sometimes linear probability model can give better result and sometimes can be probit model can give better results.

So, now you know it is again like you know continuous process to solve this problem and go through you know continuous search process till you get the better model and through which you can do the predictions and then you can go for the kind of you know management requirement.

So, that means, actually what we have already discussed here? So, we have discussed dummy independent modeling, dummy independent modelling and various ways you know simple structure, multiple structure, linear model non-linear models and again we have a different options.

So, this itself gives you know enough kind of you know flexibility in the kind of you know modelling framework, that too we are highlighting here the predictive analytic structures where the variables informations are more or less categorical. So, either through dependent and independent, but it gives you know some kind of you know different kind of you know structure altogether and again it is very useful for the today's you know business requirement.

Because some of the problems behavior is always like that only some variables may be categorical with respect to dependent variable and some variables again catalytically with respect to independent variable. Now having different you know situation and different kind of you know atmosphere and different kind of you know environment.

So, what is the best suggestion is you have to pick up a particular you know model and test continuously a you know with respect to a particular you know objective and the kind of you know the requirement. And till you get the model which is actually better fit your best fit and free from all kind of you know error starting with you know multicollinearity autocorrelation heteroscedasticity. And as usual your fitness will be and the variables most of the variables should be statistically significant and at the higher level. And then it will it will give you enough exposure about the predictions and then finally, you can go for you know perfect management decision and with this we will stop here.

Thank you very much have a nice time.