

Marketing Research and Analysis -II (Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture - 49
Logistic Regression Analysis

Hello friends, welcome to the class of marketing research and analysis. So we will be continuing from the last lecture where we had discussed about hierarchical regression and then we also discussed about dummy variable regression. So a concept which is very helpful when you have your independent variable as a categorical one right. Today, we will be discussing about a new technique which is also slightly different in nature from the usually discussed regression methods right.

As you know that in regression when you talk about regression the dependent and the independent variable they need to be continuous in nature right. So when we had a case of independent variable not being in a continuous but in a categorical variable, there we talked about the role of dummy variable coding and the dummy variable regression but what if our outcome of the dependent variable is a non-continuous or a categorical you know variable.

So a condition which is violating the general assumption of regression right since regression usually says they both have to be continuous dependent as well as the independent but here we are saying that we do not have a dependent variable which is continuous but it is a categorical variable. So in such a condition what can we do? So there are two techniques which we will be discussing in this lecture and the next lecture.

So one is called the logistic regression right and the other is called the discriminant analysis right. So logistic regression is one technique where the outcome variable or the dependent variable is actually not a continuous variable but a categorical variable and the independent variables are the predictors are all in continuous in nature right. So there also you can have both continuous as well as categorical right.

But the general assumption is the predictor is continuous and outcome is non-continuous. So let us start with it.

(Refer Slide Time: 02:38)

Logistic Regression

Introduction

Logistic regression extends the ideas of linear regression to the situation where the dependent variable, Y , is categorical.

A categorical variable as divides the observations into classes.

If Y denotes a recommendation on holding /selling / buying a stock, then we have a categorical variable with 3 categories.

Each of the stocks in the dataset (the observations) as belonging to one of three classes: the "hold" class, the "sell" class, and the "buy" class.



Logistic regression does not face the strict assumptions of multivariate normality and equal variance among groups.

Decision

H

M

L



2

So what it says logistic regression extends the ideas of linear regression to the situation where the dependent variable Y is categorical, correct. A categorical variable as divides the observation in two classes. Basically, what a categorical variable does? For example, let us see. If Y denotes a recommendation on holding or selling or purchasing a stock, then we have a categorical variable that means with 3 categories okay.

For example, make it more simpler. You want to know about the income category of people, let us say income. So your 3 income categories, high, medium and low, so these are the 3 different classes we are talking about right. So in this case each of the stocks in the data set as belonging to one of the 3 classes right, the hold class, the sell class and the buy class. Logistic regression does not face.

The one good thing about logistic regression is that why it should be used, it does not have a very strict assumptions of multivariate normality and equality of variance among groups. So the good thing about a logistic regression is that when I will talk about discriminant, the difference between logistic and discriminant is that in discriminant you need to be your data should be following all the assumptions of a multivariate of normality right.

But logistic regression does not follow very strictly all these assumptions and it can work better when your data does not follow all these assumptions right of normality, homogeneity of variance and etc.

(Refer Slide Time: 04:11)

Logistic Regression

Logistic regression extends the ideas of linear regression to the situation where the dependent variable, Y , is categorical.

Logistic regression is used in applications such as:

1. Classifying customers as returning or non-returning (classification)
2. Finding factors that differentiate between male and female top executives (profiling)
3. Predicting the approval or disapproval of a loan based on information such as credit scores (classification).

We deal only with a binary dependent variable, having two possible classes.



The results can be extended to the case where Y assumes more than two possible outcomes.

Popular examples of binary response outcomes are

- ✓ success/failure.
- ✓ buy/don't buy.
- ✓ default/don't default, and
- ✓ survive/die.

We code the values of a binary response Y as 0 and 1.

Binary Logistic Reg
Multinomial

3

So as it says logistic regression extends the idea of linear regression right so where the Y is categorical. We have already discussed. So it is used applications such as where, what are the applications? First is to classify the customers as returning or non-returning. For example, a bank classifies whether the person after let us say giving some kind of a service or some kind of a benefit, are they again coming back to the bank or they are not coming back to the bank.

Or let us say a company wants to see whether after giving a sales promotion whether the customers are returning back to the company or they are not coming back to the company right. So in such a condition it helps to classify right. Finding factors that differentiate between male and female top executives so profiling. So sometimes you need to understand what are the differences between the male and female top executives.

So you are profiling the factors right which helps in differentiating them. The third for example another classification method, predicting the approval or disapproval of a loan. So suppose there is a person who has come for a loan, should you approve the loan or should you disapprove the loan. Now on what basis will you do? Based on some information such as credit scores which may be based on your income, your past performance, your let us say age, your stay where you stay right, your ownership of property, etc right.

So we deal only with the binary dependent variable having two possible classes. So generally we will talk about the binary logistic regression right. Binary logistic regression which is the logistic regression we are talking about. Generally, when we talk about logistic regression we

talk about the binary logistics. Although, there is another method called the multinomial logistic regression.

But we are not talking about it, the only difference is that besides the binary and the multinomial that in the multinomial you have more than 2 levels instead of 0 and 1 let us say as I said about you know high, medium, low. So there were 3 classes right, so there it is a multinomial logistic but suppose you have only two classes, approval or rejection of a loan for example entry to a college or you know disqualifying the candidate.

So such kind of cases where there are only two options, we will use a binary logistic. When you have more than two options, we will have the multinomial logistic okay. So for example let us see some cases, success failure, Y can assume two values one is success, failure; buy, do not buy; defaulter, not a defaulter; will the patient survive or will he die. Now there are only two options either he will survive or he will die, there cannot be third option right.

So you may include a third option like for example somebody will say coma but then also that is a survival. So we code the values of Y as 0 and 1, so anything you can take 0 and anything you can take 1. So suppose I take failure as 0 and success as 1, it does not matter or if I take success as 0 or failure as 1 it does not matter right. We may choose to convert sometimes you can also do with some continuous data.

For example, let us say you have a data where the original income of people is known to you right. So for tax for example the government for tax benefits, so what it does, it says divided the people into different income brackets right. So bracket A is somebody maybe I am just giving example above let us say 10 lakhs per month, bracket B is somebody from 5 lakhs to 10 lakhs or bracket C is somebody in between 0 to 5 lakhs.

So this is like 3 different classes being made right so 1, 2, 3 so right so you can may convert continuous data's also into such categories and run a logistic regression right.

(Refer Slide Time: 08:02)

Logistic Regression

We may choose to convert continuous data or data with multiple outcomes into binary data for purposes of simplification, reflecting the fact that decision-making may be binary

approve the loan / don't approve,

make an offer/ don't make an offer)

the independent variables X_1, X_2, \dots, X_k may be categorical or continuous variables or a mixture of these two types.

In multiple regression the aim is to predict the value of the continuous Y for a new observation

In Logistic Regression the goal is to predict which class a new observation will belong to, or simply to classify the observation into one of the classes.

$$y = a + b(x_1) \rightarrow$$
$$0 \leq y \leq 1 = a + b + x_1$$

The independent variables maybe categorical but we generally talk about continuous variables right because that is the general property of a regression model or regression equation right or you can have both right because now by this time you have already learnt about dummy variable regression, you have learnt about simple regression and you are today learning about logistic regression.

So when you take you can understand that it can take any variable that means any kind of value, be it continuous or be it categorical. In multiple regression, the aim is to predict the value of the continuous Y for a new observation that is what we were doing right but in logistic regression the goal is to predict which class a new observation will belong to or simply to classify the observation into one of the classes.

So here there in the normal regression what you were doing, you are measuring the Y on basis of a certain value of X after calculating for a and b. So when X was changing what is the new value of Y, you were trying to find out but here you are not doing that. What in regression equation you are doing since you have only two options 0 and 1, so on basis of this values you will only say whether the person will fall into this category 0 or will fall into this category that is what we are trying to find.

(Refer Slide Time: 09:22)

Logistics Regression model

The "logit" model: log odds ratio, or "logit"



$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

[range = $-\infty$ to $+\infty$]

- p is the probability that the event Y occurs. $p(Y=1)$ is given as

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$
 - [range = 0 to 1] (0, 1)
- $p/(1-p)$ is the "odds ratio" where, $p/1-p = \text{odds} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}$.
 - [range = 0 to ∞]

Note: The estimated probability is: $p = 1/[1 + \exp(-\alpha - \beta X)]$



5

So the logic model or log odds ratio or logit is given as this is the formula how you understand it is given right, so log of $p/1-p$ right. What is this p ? p is the probability says p is the probability that the event will occur, Y occurs right. So log of $p/1-p$ is equal to what? β_0 so that means what, your initial intercept + slope $1 \beta_1 X_1$ sorry this should $\beta_1 X_1 + \beta_2 X_2$ right goes on + $\beta_n X_n$ right.

So this is the formula and this ranges from $-\infty$ to $+\infty$ okay, p is the probability that the event will occur and p is given as you can say this p is given as $p = 1/1 + e^{-\beta_0 - \beta_1 X_1 - \beta_2 X_2}$ up to whatever $\beta_n X_n$ right and obviously as you know the value of a probability lies between 0 and 1. So this is maximum it will lie, $p/1-p$ this part right is called as the odds ratio right where $p/1-p$ is = odds is given as this is the formula.

So just you note down this formula, you remember this formula. So e so here we are using $p = 1/1 + e^{-\beta_0 - \beta_1 X_1 - \beta_2 X_2}$ right, $1/1 + e^{-\beta_0 - \beta_1 X_1 - \beta_2 X_2}$ right. So note the estimated probability is $p = 1/1 + \exp(-\alpha - \beta X)$ so this is what you are saying $-\alpha - \beta X$ right.

(Refer Slide Time: 11:12)

Logistic Regression



Logistic regression thus forms a predictor variable ($\log(p/(1-p))$) which is a linear combination of the explanatory variables.

The values of this predictor variable are then transformed into probabilities by a logistic function.

Such a function has the shape of an S. (See the graph on the next slide) 0 1

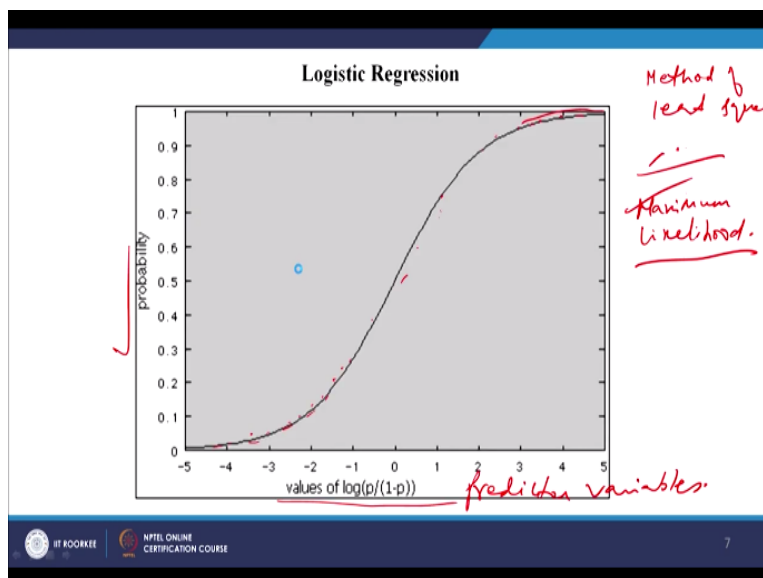
On the horizontal axis we have the values of the predictor variable, and on the vertical axis we have the probabilities.

Logistic regression also produces Odds Ratios (O.R.) associated with each predictor value.



6

So now using this formula we will try to find out right. So logistic regression thus forms a predictor variable of $\log p/1-p$ because the logarithmic value which is the linear combination of the explanatory variables, the predictor variables. The values of these predictor variables are then transformed into probabilities by a logistic function right. We will show this function probabilistic function, it is a probability function, so it lies between 0 and 1. So what happens, I will show you in the next slide maybe, so let us look at this.

(Refer Slide Time: 11:42)



If you see if I take this the probabilities are taken in between 0 and 1 right and the values of the log this value right you see this, on the horizontal axis we have the values of the predictor variable, horizontal axis the predictor variables, so these are all the predictor variables right and on the vertical axis you have the probabilities so this is the probabilities right. So when I draw you know as I plot the values so all the values will lie between 0 and 1.

Why? Because the probability maximum ranges from 0 and 1, so that is why when you plot the different estimates right after iteration so interestingly you should know that logistic regression does not follow a method of least square it does not follow it, why? Because here there is no point of finding out the variance from the regression line right. You do not have any value.

So what you do is basically you follow in the case of a logistic regression you are following a maximum likelihood method. Why it is called a maximum likelihood method? Because it says what is the maximum chance of an event to occur right. So this is termed this value is given through the probability right. So the probability is the chance of occurring right. So what is the maximum probability of occurrence okay?

So when we do this it gives a S-shaped, if you connect the dots right the values these probability values, you will see that it gives you a S-shaped curve called the sigmoid curve right, the sigmoid curve okay. So this is what I was talking about the S curve or the sigmoid curve right okay.

(Refer Slide Time: 13:28)

Odds Ratio in Logistic Regression

The "odds" of an event is defined as the probability of the outcome event occurring divided by the probability of the event not occurring. Odds are usually written as "5 to 1 odds" which is equivalent to 1 out of five or .20 probability or 20% chance, etc.

If there is a $\frac{3}{4}$ chance that it will rain tomorrow, then 3 out of 4 times we say this it will rain. That means for every three times it rains once it will not. The odds of it raining tomorrow are 3 to 1. This can also be understood as $(\frac{3}{4})/(\frac{1}{4})=3/1 = 3$.

If the odds that my pony will win the race is 1 to 3, that means for every 4 races it runs, it will win 1 and lose 3.

IFT KOOKEE NPTEL ONLINE CERTIFICATION COURSE 8

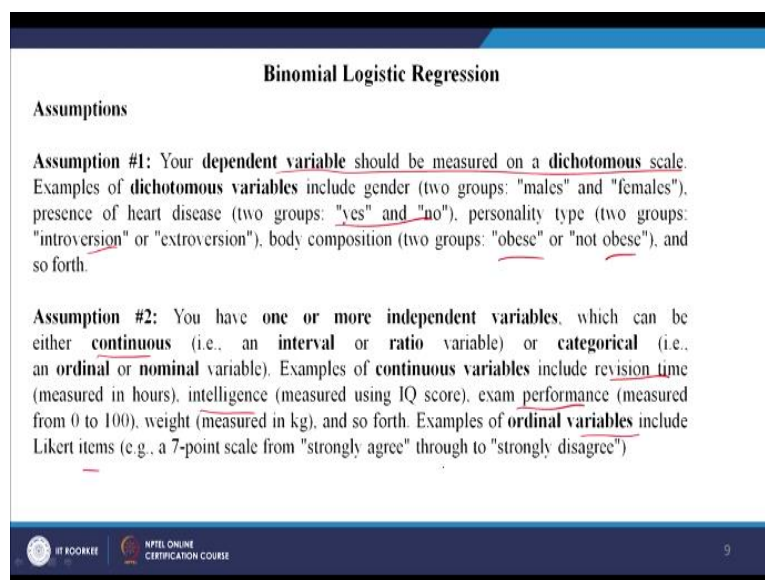
Now what is this odds ratio in the logistic regression? Let us understand it very clearly. The odds of an event is defined as the probability of the outcome event occurring divided by the probability of the event not occurring simple. The probability of the outcome event occurring divided by the probability of the event not occurring, so odds are usually written as for example 5 to 1 odds.

What does it mean? Which is equivalent to 5 times in every 5 times there is only one chance that you might get it or not get it. You see 1 out of 5 or 0.2 or 20% chance that means in every 5 times that I will let us say toss a coin the chance that I will get a head is only one time let us say right. So then I will say what is the probability there is only 20% chance or 20% probability okay.

If there is 75% chance that it will rain tomorrow, then 3 out of 4 times we say that it will rain correct because 75 is 3 times out of 4 right. This means that for every 3 times it rains once it will not right. So the odds of it raining tomorrow are 3 to 1 correct. This can be understood as 3/4/1/4 this is not happening, this is happening is=3/1 is=3. If the odds that my pony my horse will win the race is 1 to 3 that means what?

For every 4 races it runs it will win once and lose 3. So this is what it means, odds means what is the chance of something happening the probability of the event occurring/not occurring right. So some other assumptions of the logistic regression now. What are the assumptions?

(Refer Slide Time: 15:19)



Binomial Logistic Regression

Assumptions

Assumption #1: Your dependent variable should be measured on a dichotomous scale. Examples of dichotomous variables include gender (two groups: "males" and "females"), presence of heart disease (two groups: "yes" and "no"), personality type (two groups: "introversion" or "extroversion"), body composition (two groups: "obese" or "not obese"), and so forth.

Assumption #2: You have one or more independent variables, which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable). Examples of continuous variables include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. Examples of ordinal variables include Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree")

IT KOOKEE NPTEL ONLINE CERTIFICATION COURSE 9

Your dependent variable should be measured on a dichotomous scale; you know about it right. Second assumption now for example yes, no; extrovert, introvert; obese, not obese given some examples. One or more independent variables which can be either continuous or categorical so that also we have explained, explains the example of continuous variables

include revision time, intelligence, performance in exam, etc and ordinal variables include likert items also which can be taken right.

(Refer Slide Time: 15:50)

Binomial Logistic Regression

Assumptions

Assumption #3: You should have independence of observations and the dependent variable should have mutually exclusive and exhaustive categories.

Assumption #4: There needs to be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable.

$\hat{Y} = X_1 X_2$

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 10

Third assumption, independence of observations, again I have repeated it several times where the observation or the respondent is taken once and only once right until unless this is the study for repeated (()) (16:04). There needs to be linear relationship between any continuous independent variable in the logit transformation of the dependent variable. So what it says, there has to be a linear relationship between any continuous independent variables right.

So any predictor variable right X_1 , X_2 and the logit transformation of the dependent variables so the logit transformation that you make there has to be a linear relationship between this and this value right. So but remember the best part is that the best thing about logistic regression is that logistic regression is not you know does not follow strictly you know if your data you know violates the normality assumptions, then in that case you can very well use the logistic regression.



But in such conditions its alternative for example the discriminant analysis which I will be explaining in the next class will not be applicable right okay. Let us take a case.

(Refer Slide Time: 17:04)

Binomial Logistic Regression

Example

A researcher wants to know whether the "credit card default" can be predicted based on "monthly salary", and "gender". To this end, the researcher recruited 159 participants to gather the data. The participants were also evaluated for the present of credit card default. Gender (female=0, male=1) and credit card default (default= 1, no default=0) were coded into the spss data. A binomial logistic regression was then run to determine whether the credit card default could be predicted from their monthly salary and gender.

  11

A researcher wants to know whether the credit card default can be predicted, every company wants to know can we predict a defaulter or not. Based on monthly salary and gender right, so the researcher recruited 159 participants to gather the data right. The participants were evaluated for the present of credit card default, so gender female is=0, male is=1 and default if somebody has defaulted is 1, no default 0 were coded into the data set.

A binomial logistic regression was then run to determine whether the credit card default could be predicted from their monthly salary and gender. So such a situation you must have also faced in life. I want to know whether I would be you know can I get a admission into some university in London or some university in US on basis of my scores and some of my other characteristics such as my age, my past record in my school days and all. So I am interested okay. So how do you do this?

(Refer Slide Time: 18:13)

Binomial Logistic Regression

Test Procedure in SPSS Statistics

Step-2 Transfer the dependent variable, in dependent box and independent variables into covariates

14

And monthly salary which is this one and my gender as my covariates right but you remember that gender is a categorical and monthly salary is continuous okay.

(Refer Slide Time: 18:53)

Binomial Logistic Regression

Test Procedure in SPSS Statistics

Step-3 Click the categorical button. You will be presented with the **Logistic Regression: Define Categorical Variables** dialogue box

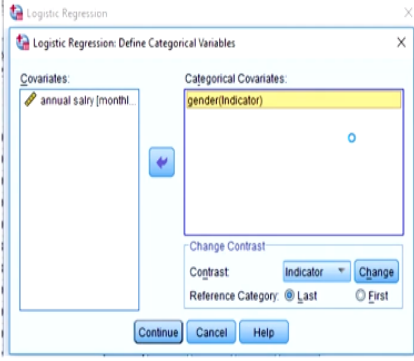
15



(Refer Slide Time: 18:56)

Binomial Logistic Regression

Test Procedure in SPSS Statistics

Step-4 Transfer the independent, categorical variable, gender from covariates to categorical covariates





16

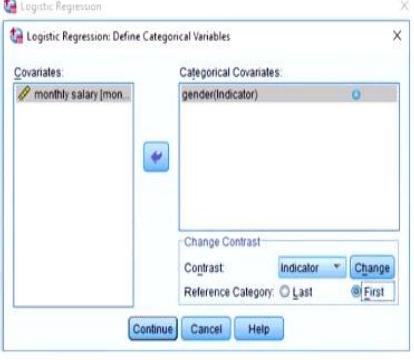
So what I am doing is now I will do one thing, I will take this you know if you okay so I can take this into this categorical variable right so categorical covariates. It has a function; logistic regression has a function to take to explain whether your covariate is categorical or continuous in nature. So whether your covariate is categorical or continuous in nature, it gives you provides you a space for that.



(Refer Slide Time: 19:23)

Binomial Logistic Regression

Test Procedure in SPSS Statistics

Step-5 In the change contrast change the reference category to first (Whether you choose last or first will depend on how you set up your data. In this example, males are to be compared to females, with females acting as the reference category (who were coded "0"). Therefore, first is chosen)





17

(Refer Slide Time: 19:25)

Binomial Logistic Regression

Test Procedure in SPSS Statistics

Step-6 click the continue button
You will be returned to the **Logistic Regression** dialogue box.

Step-7 Click the options button.
You will be presented with the **Logistic Regression: Options** dialogue box

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 18

Now once you have done it right then this is what you get in the options, so there are options right. If you go to this option right, so when you go to options you get this right and here I will show you.

(Refer Slide Time: 19:38)

Binomial Logistic Regression

Test Procedure in SPSS Statistics

Step-8 In the statistics and plots area click the areas shown in the figure

Step-9 then click continue button you will be returned to the **Logistic Regression** dialogue box.

Step-10 Click the **OK** button. This will generate the output.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 19

(Refer Slide Time: 19:39)

Binomial Logistic Regression

Interpreting and Reporting the Output



Variance explained

In order to understand how much variation in the dependent variable can be explained by the model (the equivalent of R^2 in multiple regression), you can consult the table below, "Model Summary":

Model Summary			
Step	-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
1	82.730 ^a	.170	.335

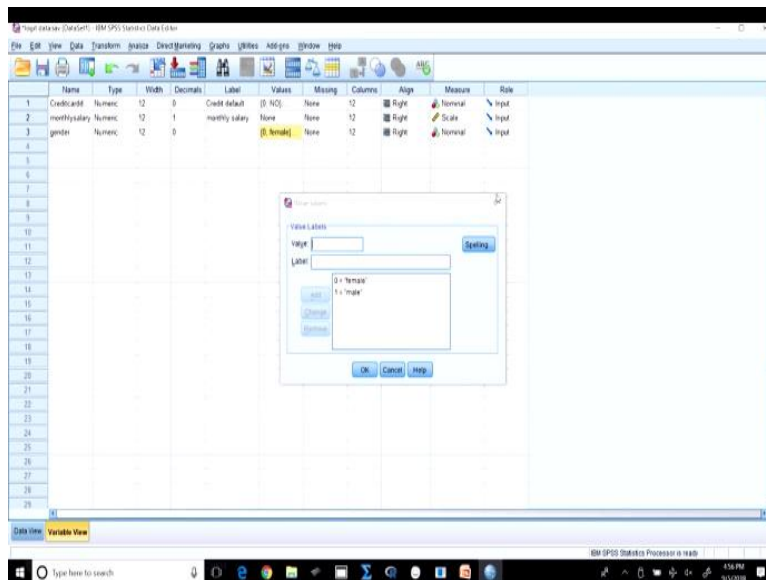
^a Estimation terminated at iteration number 7 because parameter estimates changed by less than .001

table contains the **Cox & Snell R Square** and **Nagelkerke R Square** values, which are both methods of calculating the explained variation. These values are sometimes referred to as *pseudo R²* values (and will have lower values than in multiple regression). However, they are interpreted in the same manner, but with more caution. Therefore, the explained variation in the dependent variable based on our model ranges from 17.0% to 33.5%, depending on whether you reference the Cox & Snell R^2 or Nagelkerke R^2 methods, respectively. Nagelkerke R^2 is a modification of Cox & Snell R^2 , the latter of which cannot achieve a value of 1. For this reason, it is preferable to report the Nagelkerke R^2 value.



20

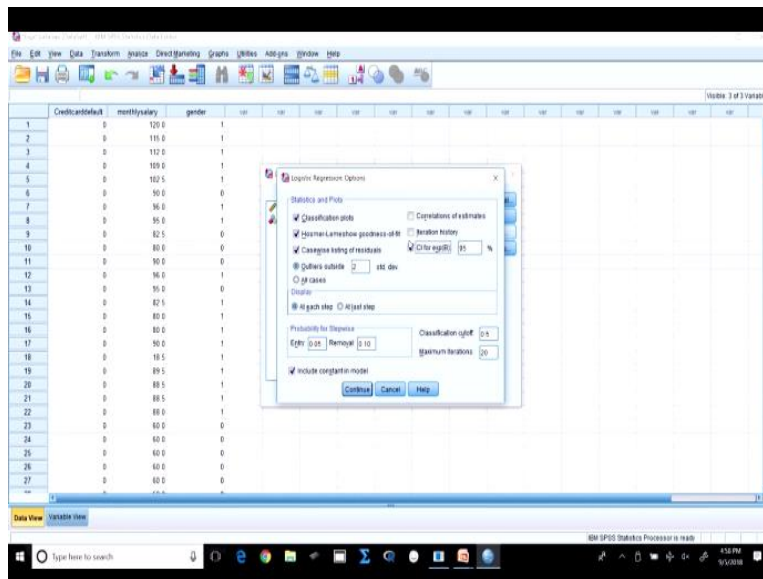
Now let me go to the main you know slide and I will show you the data. Now this is the exact data you are talking about.

(Refer Slide Time: 19:44)



So we know let us go to the variable view so credit card defaulter so let us see what is this, if it is a defaulter then yes, if it is not a defaulter then no. There is no default then no 0 and yes is a defaulter 1 let us say. Similarly, let us go to monthly salary so there is nothing to talk about monthly salary. Gender, what is it? 0 is female, 1 is male okay. So you have got your values. Now let us run a binary logistic regression right.

(Refer Slide Time: 20:18)



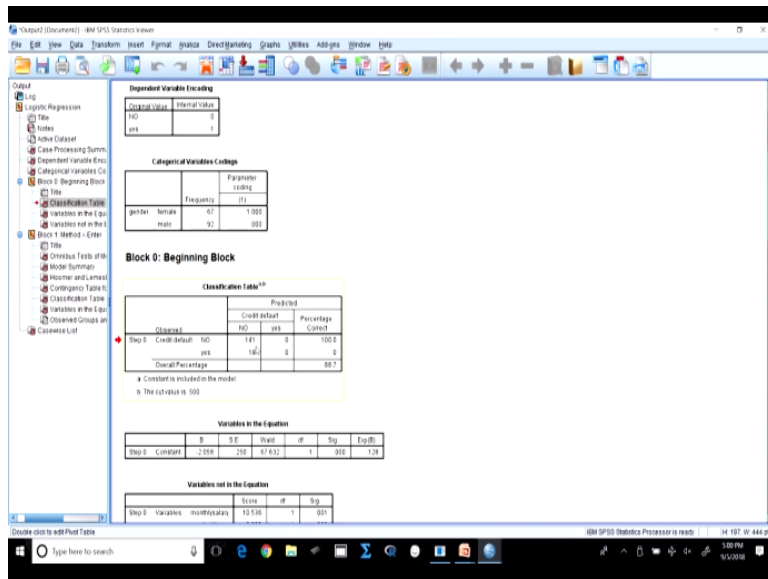
So let us go to regression first and go to binary logistics so what is the dependent variable, I want to know whether my credit card default, whether the person who is coming new whether he will be defaulter or not a defaulter, so I am taking it as a dependent variable right. So what are my covariates? These two are my covariates okay. So now I go to this one categorical. Now you see is there a categorical covariate? Yes, I do have a categorical covariate.

Which one? Gender right because it is categorical 0 and 1 male and female. Now I am saying continuous okay here it is interesting you see. Reference category it tells you something about a reference category. Can you see this? Now this reference category, there are two options, last and first. So if you see now in our case suppose for example you want to compares against what, suppose for example male and female.

Now let us say my reference category is let us say female and female is let us say 0 right that is what we have done. Let us go back okay, let us go back and see variable view so female is 0 right and male is 1 right. So remember this now let us go back. So analyze regression so binary and this is this, this is this, this is here categorical so I am taking my categorical here and so what I am doing is since first I am comparing the female and the male, I am taking the reference category as the female and male is my the other one which has the value of 1.

So I take the first one the first is the male which is 1 right and I am taking as and comparing it right, so now after this I go to options, what I need is this value right the goodness of it and then I need the classification plots, I need the case wise listing of residuals, I need confidence interval for exponential right and this is all I require. Now let us run it.

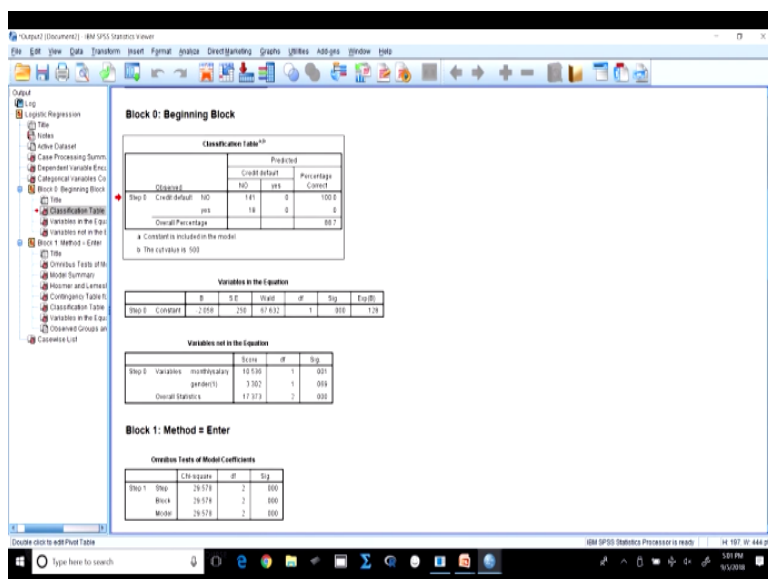
(Refer Slide Time: 22:35)



So as you see as you look at the gender now female 67 and male 92 cases are there. Now if you look at how many people have done credit default right. So beginning the first step right default it was observed that there is no default and what is predictor, this is observed and this is predicted. So how did the model predict? No, no. So there will be no default and there is no default 141 right.

No default and yes so, now overall percentage is saying suppose you see people were predicted that they would make a default but they did not make a default is 18 cases. So the actually the model's accuracy rate is now so because it has been you know because of this 18 now it has come down to 88.7% okay.

(Refer Slide Time: 23:27)



Now let us look at this. So once we have got these values so when you look at the output this output this table the variance the model summary is important for us.

(Refer Slide Time: 23:39)

The screenshot shows the SPSS 'Model Summary' table with the following data:

Step	-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
1	82.730 ^a	.176	.335

Additional tables visible in the screenshot include:

- Variables in the Equation:**

Step	Constant	B	S.E.	Wald	df	Sig.	Exp(B)
1	-.2558	.252	87.632	1	.000	1.291	
- Variables not in the Equation:**

Step	Variables	Wald	df	Sig.
1	monofluoride	11.530	1	.001
1	gender(f)	3.962	1	.048
1	Overall Distance	17.371	2	.000
- Chi-square Tests of Model Coefficients:**

Step	Step	Chi-square	df	Sig.
1	Step	29.578	2	.000
1	Block	29.578	2	.000
1	Model	29.578	2	.000

Now let us go to the output file again. So as you go down this is the model summary I am talking about, I was searching for this. So if you look at this model summary right, so this 2 log likelihood or 2ll it is called as -2ll and the Cox and Snell R square and Nagelkerke. So these are called pseudo R squares right. They are nothing but it is like your normal R square right and it ranges from 0.17 to 0.335 right.

So this value is basically talks about like any normal regression equation like you have a R square. This is similar to that R square value right and then let us go back and look at this. So this table now for example this value right if you look at the same value we are getting here right, so it contains the Cox and Snell R square and Nagelkerke R square value which are both methods of calculating the explained variance like in the regression equation R square right.

These values are sometimes referred to as pseudo R square values right and however they are interpreted in the same manner but with some caution right. So what it says therefore the explained variation in the dependent variable based on our model ranges from how much 17% to 33.5% right. So depending on whether your reference, you can refer anyone but the Nagelkerke one is much better right is more preferred when compared with the Cox and Snell R square right.

So Nagelkerke R square is a modification of the Cox and Snell R square only and the later of which cannot achieve value of 1 right. So what it says, this one the Cox and Snell will never achieve a value of 1 right. For this reason, it is preferable to report the Nagelkerke R square value. Why it does not reach the value of 1, there is a reason behind it also. The scale used R for the Cox and Snell and the Nagelkerke are different actually.

So that is the basic reason why the Cox and Snell value will always be lesser than the Nagelkerke R square value right. So that is why it is a modified Cox and Snell R square only okay.

(Refer Slide Time: 25:50)

Binomial Logistic Regression

Interpreting and Reporting the Output
Category prediction



Binomial logistic regression estimates the probability of an event (in this case, credit default) occurring. If the estimated probability of the event occurring is greater than or equal to 0.5 (better than even chance), SPSS Statistics classifies the event as occurring (credit card default -yes).

Observed		Predicted		
		Credit default		Percentage Correct
		NO	yes	
Step 1	Credit default NO	140	1	99.3
	yes	13	5	27.8
Overall Percentage				91.2

a. The cut value is .500

Firstly, notice that the table has a subscript which states, "The cut value is .500". This means that if the probability of a case being classified into the "yes" category is greater than .500, then that particular case is classified into the "yes" category. Otherwise, the case is classified as in the "no" category (as mentioned previously).

0.5
y
no.



21

Now how do we interpret this? So as we see we did a binomial logistic regression right. The probability of an event in this case the credit default right occurring, we need to check it, so if the estimated probability of the event occurring is greater than or equal to 0.5 right SPSS or any software classifies the event as occurring credit card default yes, suppose it is more than 0.5 then it is yes right.

If it is less than 0.5 it is no right. First, notice that the table has a subscript this one which states the cut off value is 0.5 right. This means that if the probability of a case being classified into the yes category is greater than the 0.5 then that particular case is classified into yes right. Otherwise the case is classified as in the no category. You do not have to understand, just remember what it says is to understand that how it takes 1 or 0.

If it is more than 0.5 then it is if it is even 0.55 then it is a yes category, if it is even 0.49 it is a no category that is what and this is maybe you can say is a drawback because the logic is you know cannot be applied. Sometimes in real life you may have to you know think that 0.49 is as good as 5 but then statistically this is not possible and this is the demerit sometimes of mathematics or statistics you can say right.

(Refer Slide Time: 27:21)

Binomial Logistic Regression

Interpreting and Reporting the Output

Variables in the equation

The "Variables in the Equation" table shows the contribution of each independent variable to the model and its statistical significance. This table is shown below:

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a monthsalary	-.191	.054	12.657	1	.000	.826	.743	.918
gender(1)	-2.146	.746	8.263	1	.004	.117	.027	.505
Constant	8.391	2.740	9.378	1	.002	4406.068		

a. Variable(s) entered on step 1: monthsalary, gender.

The Wald test ("Wald" column) is used to determine statistical significance for each of the independent variables. The statistical significance of the test is found in the "Sig." column. From these results you can see that gender ($p=.004$) and monthly salary ($p=.000$) added significantly to the model/prediction.

The variables in the equation table shows the contribution of each independent variable. Now let us go back to our data set right.

(Refer Slide Time: 27:27)

The screenshot displays the SPSS output for a Binomial Logistic Regression. Key sections include:

- Variables in the Equation:** Shows coefficients for monthsalary and gender(1), with significant p-values.
- Omnibus Tests of Model Coefficients:** Shows a significant Chi-square test result (p < .001).
- Model Summary:** Shows a significant Nagelkerke R-squared value (p < .001).
- Hosmer and Lemeshow Test:** Shows a non-significant result (p = .833), indicating a good fit of the model.
- Contingency Table for Hosmer and Lemeshow Test:** Shows the distribution of observed vs. expected values across deciles.

So let us look at, this is the model summary we saw right. So actually this Hosmer-Lemeshow test is nothing but like similar to chi-square test and it should be significant. That

means the model is significant. It talks about the goodness of it that means what it says that the variables are fitting well to explain the model or the dependent variable okay.

(Refer Slide Time: 27:53)

The screenshot displays the SPSS Statistics viewer interface. The main window shows the following data and tables:

Step	Observed	Expected	Observed	Expected	Total
Step 1	17	16.999	0	0.001	17
2	16	16.937	1	0.63	17
3	15	14.851	0	1.69	15
4	19	18.402	0	5.90	19
5	16	15.876	1	4.01	17
6	22	20.705	1	2.295	23
7	17	15.504	1	2.494	18
8	14	12.530	2	3.470	16
9	5	9.140	12	7.860	17

Observed	Predicted		Percentage Correct
	Credit default: no	Credit default: yes	
Step 1 - Credit default: no	142	1	99.3
yes	13	5	27.8
Overall Percentage	17	6	91.2

Step	Variable(s)	B	S.E.	Model	df	Sig.	Exp(B)	95% C.I. for Exp(B)
Step 1 ^a	monthlysalary	.191	.054	12.457	1	.000	1.21	.743 1.918
	gender(f)	-.214	.046	8.263	1	.004	.817	.677 1.005
	Constant	8.391	2.740	6.378	1	.002	4408.018	

Now let us go to this one right. Now classification table two tables which are important, one is the classification table and here you see the cut off value is 0.5 right. So in the case where the observed variable was no that means there was no default credit default and the predicted value was also no, it is the case of 140 people which is no problem correct but the observed value was that it will the person will not be a defaulter.

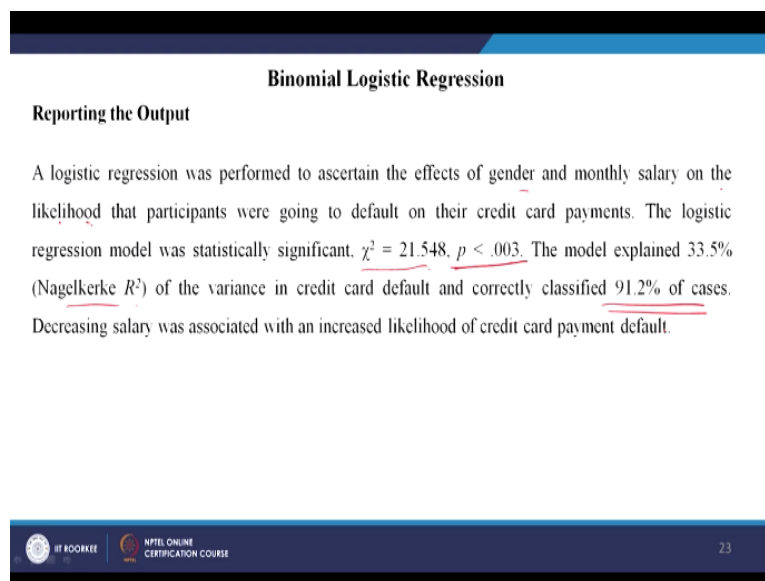
But in real life it became sorry in real life it is not a defaulter but the equation predicted that it will be a defaulter is only one case right. So the observed is he is actually a defaulter but what happened during prediction you have found that it was assumed that they will not be a defaulter so there are 13 cases. So 13 cases were assumed that they will not be defaulters but they actually became defaulters.

So this is the mistake of you know weakness in our technique and there are 5 cases which was observed that they would default and they actually defaulted 5. So taking this overall the percentage when you measured it said that the model is giving a 91.2% right. So 91.2 which is the overall model's strength. Now look at this 2 independent variables, monthly salary and gender. Now if you see this helps you from here you can find out from the significance that monthly salary is significant at higher level and gender is also significant.

That means what from here we can say that monthly salary does predict whether a person will be a defaulter or not a defaulter. Similarly, gender also predicts whether a person will be defaulter or not defaulter okay. Now let us go back to our slide and you see the Wald test the Wald column is used to determine the statistical significance we just saw for each of the independent variables.

And from the significance column you can see that gender had a p is=0.004 and monthly salary 0.000 added significantly to the model to predict the model.

(Refer Slide Time: 30:06)



Binomial Logistic Regression

Reporting the Output

A logistic regression was performed to ascertain the effects of gender and monthly salary on the likelihood that participants were going to default on their credit card payments. The logistic regression model was statistically significant, $\chi^2 = 21.548$, $p < .003$. The model explained 33.5% (Nagelkerke R^2) of the variance in credit card default and correctly classified 91.2% of cases. Decreasing salary was associated with an increased likelihood of credit card payment default.

IIIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 23

Now how am I reporting this? A logistic regression was performed to ascertain the effects of gender and monthly salary on the likelihood right. So remember again this word maximum likelihood it is because a probabilistic model, it is not a method of least square right which is generally you follow in the multiple regression. The participants were going to default on their credit card payments.

The logistic regression model was statistically significant the chi square value which you saw which decide about this right was significant at 0.003. The model explained 33.5% the Nagelkerke R square I am assuming here taking here of the variance in the credit card default and correctly classified 91.2% of the cases. Decreasing salary was associated with an increased likelihood of credit card payment default and it was observed that men tended to default more than women right.

So this is how you write the output right. So I hope you are clear that this logistic regression is a very interesting, very important way of understanding to do a regression when your dependent variable or outcome variable is in a dichotomous mode or in a binary mode right. So in such conditions, you can use logistic regression to predict whether the outcome will happen or not happen right.

So similarly in the next lecture, I will talk about another technique called the discriminant analysis which is similar to this but there are slight differences which I will explain in the next class. Thank you so much.