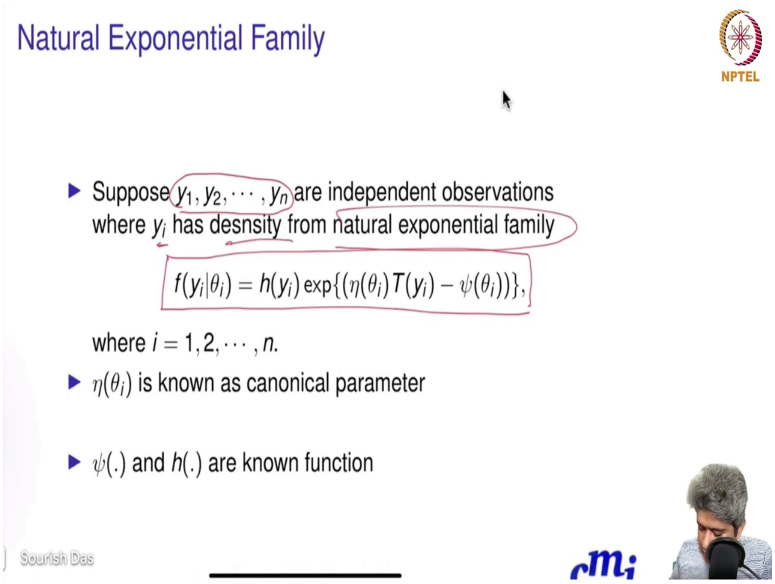
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 47
Generalised Linear Model

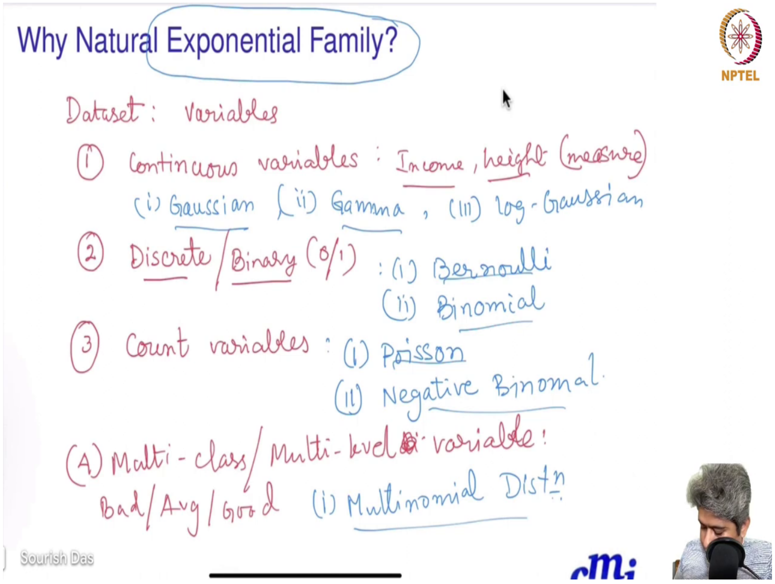
Hello all, welcome back to Predictive Analytics, Regression and Classification course. We are going to talk about Predictive Analytics, Regression and Classification for lecture 14 and in this lecture, we are going to talk about today Natural Exponential Family.

(Refer Slide Time: 00:34)



So, natural exponential family is a class of distributions. Suppose you have bunch of random variables y 1, y 2, y n they are independent observations and y i has a density of the form something called natural exponential family and this family this density of which can be cast into this functional form. I will talk about what is this functional form and how can we cast into this form. Now, question is why should we study natural exponential family?

(Refer Slide Time: 01:18)



Why we want to study natural exponential family? So, if you look into any data set, any data set ok and this data set will have a different kinds of variables ok. So, bunch of different variables and pretty much you can classify these variables into certain categories.

1st category will be a bunch of continuous variables, then 2nd category will be bunch of a discrete variable or discrete binary variable like 0 1 kind of thing. And then and 3rd will be count variable ok count variables ok.

So, continuous variables will be like you know income ok and maybe height some measures some it is going to measure something that is a feature of typically if you if a variable is coming out of a measurement, then it is typically a continuous variable. If it is like some kind of action decision here, I am going north, south I am going to go going or not yes no then it comes a binary or discrete 0 1 kind of variable.

And sometimes it is if it is count variable then like you know if you are if the variable is coming out of counting process number of you know tumour number of goals or something like that you know or number of courses you want to take something like that.

So, now typically how can you model income height this continuous variable the models that the ones which models the continuous variable they are typically like a Gaussian distribution, then there are gamma distributions then there are log Gaussian distributions. So, bunch of distributions comes into display for binary and discrete you have typically Bernoulli distributions or typically we call it binomial family of distribution binomial family of distributions.

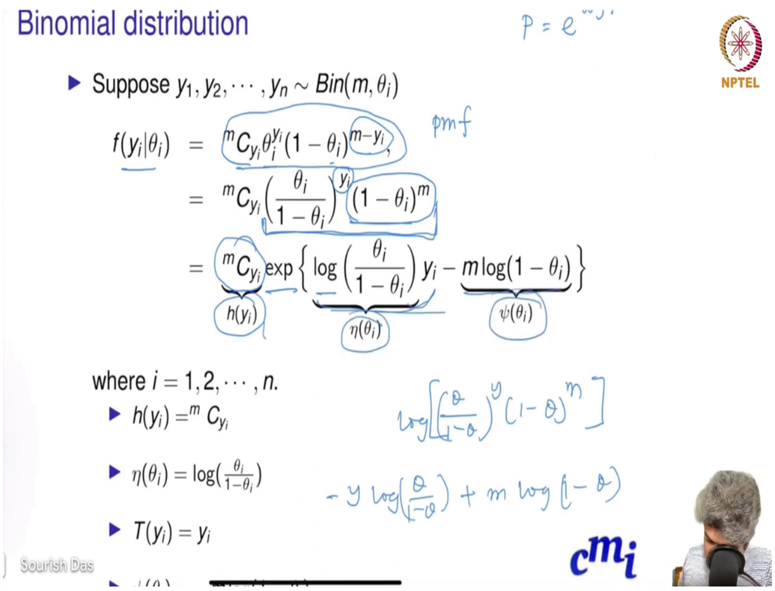
And the count variables typically we model with either Poisson distributions negative binomial distributions. So, these are the typical popular distributions which are we use to model the count variables ok. The 4th category there is a 4th category I would say called multi class or multi level variables multi class or multi level distribution multi level variables ok.

Now, what is the example of multi class or multi level variables let us suppose you want to you are asking your subject is it a bad or average or good service ok. So, some kind of multi level variable or multi class variable or how was the performance of a student is it A grade is it a B grade is it a C grade is it a D grade. So, how what is the performance of a student right. So, grade could be a multi level variable. Ok

So, for that kind of variable we can have a interesting distribution called multinomial distribution ok. So, we call it multinomial distribution ok. Now, these all these distribution Gaussian, gamma, Bernoulli, binomial, Poisson, negative binomial, multinomial distribution these all these distributions falls in the big family of exponential class of family.

So, if we can if we can build a model a predictive model where for exponential class of family then automatically all other models will become a special case for this model and that is where natural exponential family is going to play a very crucial role. Ok.

(Refer Slide Time: 07:08)

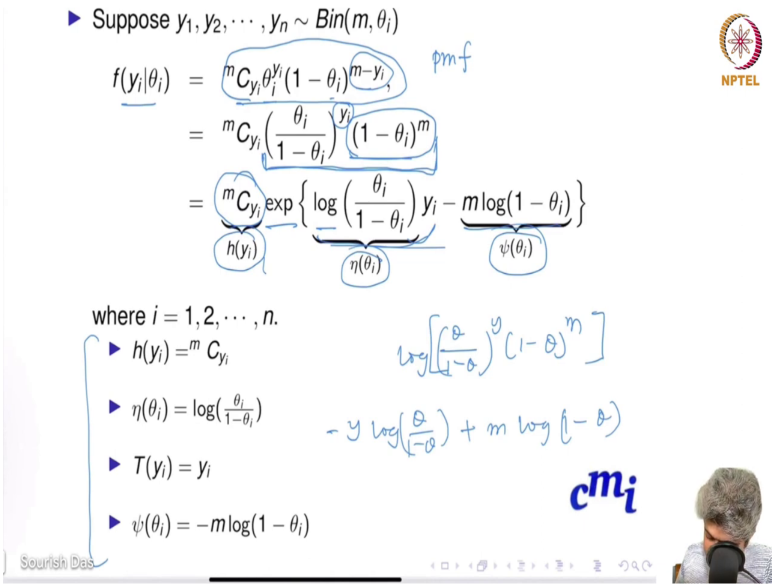


So, the first model I am going to talk about binomial distributions. Suppose y 1 y 2 y n are binomial distribution with some m and theta i ok. Now, that means, my distribution is f of y i theta i m choose m C y i theta i to the power y i 1 minus theta i to the power m minus y i. I can write this this is the simple pmf of binomial distribution. I can write the same pmf in a slightly differently I can write it as theta to the power theta by 1 minus theta to the power y i and 1 minus theta times to the power m.

Now, you see this part does not involve any y ok and then this entire thing I can raise to e to the power log. So, I can write a p as e to the power log p I can write it in this way. So, I can write this guy as e to the power log of the entire thing and then log of theta 1 minus theta to the power y 1 minus theta to the power m.

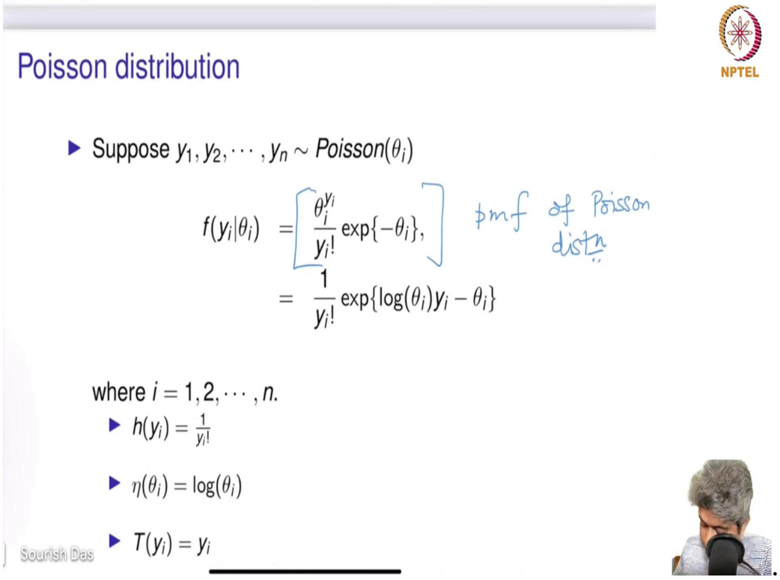
This entire thing I can write as log of y times log of theta 1 minus theta plus m times log of 1 minus theta ok. And then if I have a minus y here e to the power this is like this then you can write it in this way. So, then you can write this part as h of y this part as eta theta i and this part as psi theta i.

(Refer Slide Time: 09:32)



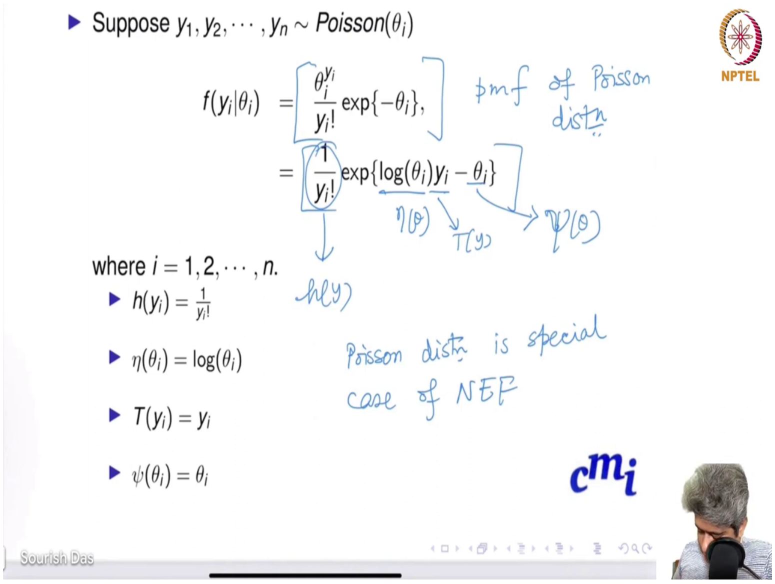
So, you can put this in this format. Now, if you go back this binomial distribution function the PMF of binomial distribution function can be formed in this kind of cast into the natural exponential family with each part has its own functional form ok.

(Refer Slide Time: 09:59)



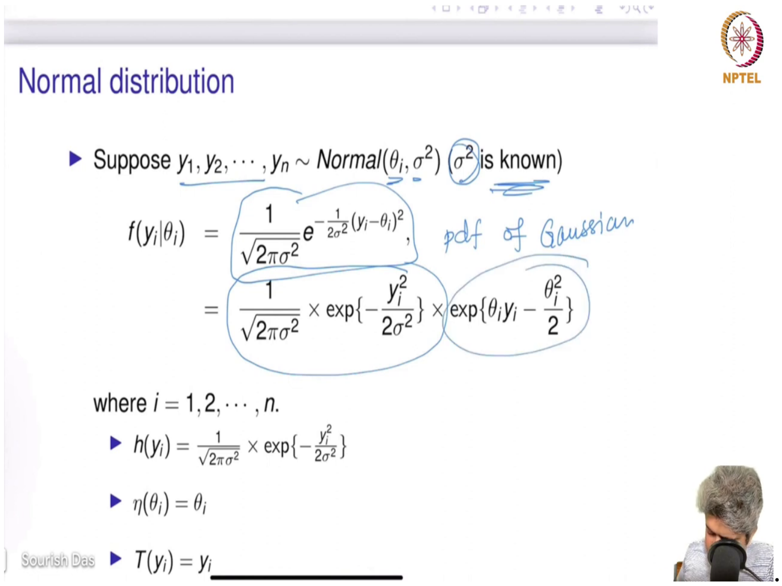
Now, next you see if you Poisson distribution if you see Poisson distribution similarly you can write Poisson this is the pmf of Poisson distribution ok pmf of Poisson distribution.

(Refer Slide Time: 10:23)



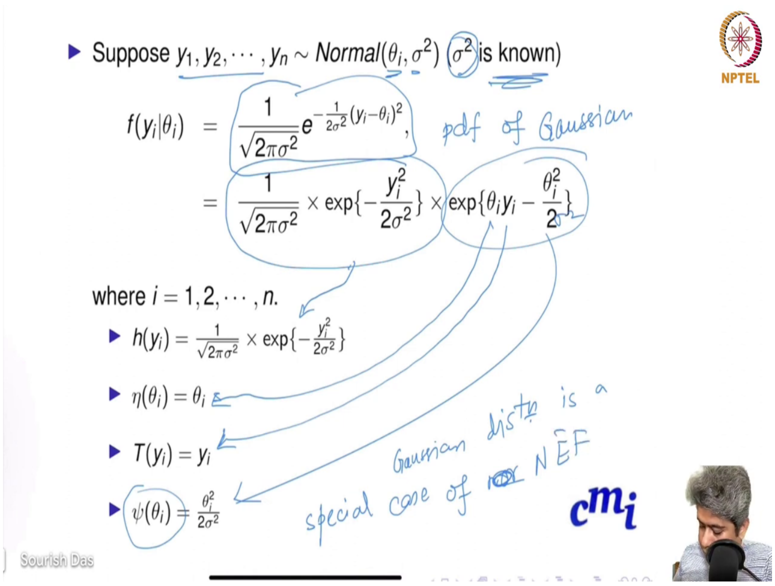
And I can write it in this format and this log theta is my eta theta this y i is my T of y psi theta is this theta is my psi of theta and 1 over y is my h of y. So, now, this I can write it as a I can say that Poisson distribution is a spatial is a spatial case of natural exponential family class of distributions.

(Refer Slide Time: 11:07)



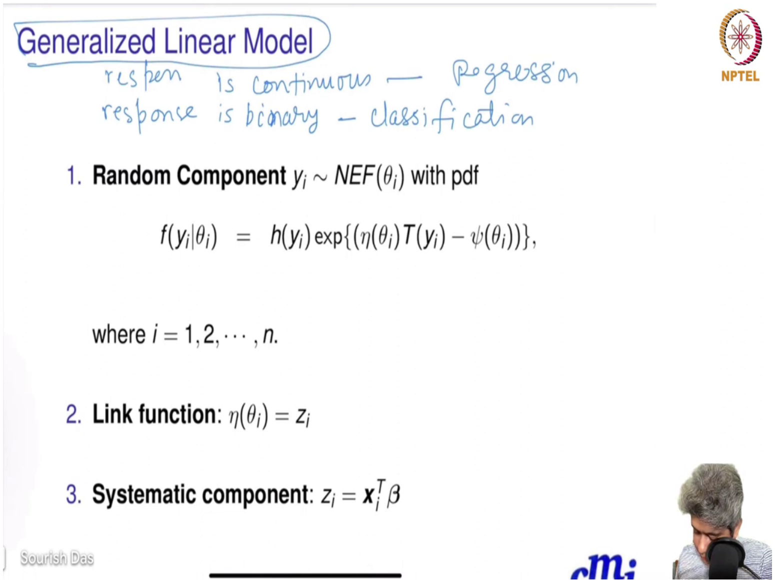
Similarly, you can choose normal distribution suppose normal theta y 1 y 2 y n follow normal theta i sigma square where sigma square is normal. I am just taking it as you know and then you can write this is the pdf of normal distribution pdf of Gaussian or normal distribution you can take the pdf and you can write it in this format ok.

(Refer Slide Time: 11:37)



And then h of y this guy is my h of y theta i is this y i is this and theta i square by 2 sigma square there must be sigma square is my psi theta i. So, I can now say Gaussian distribution is a spatial case of normal a natural exponential family of distributions.

(Refer Slide Time: 12:21)



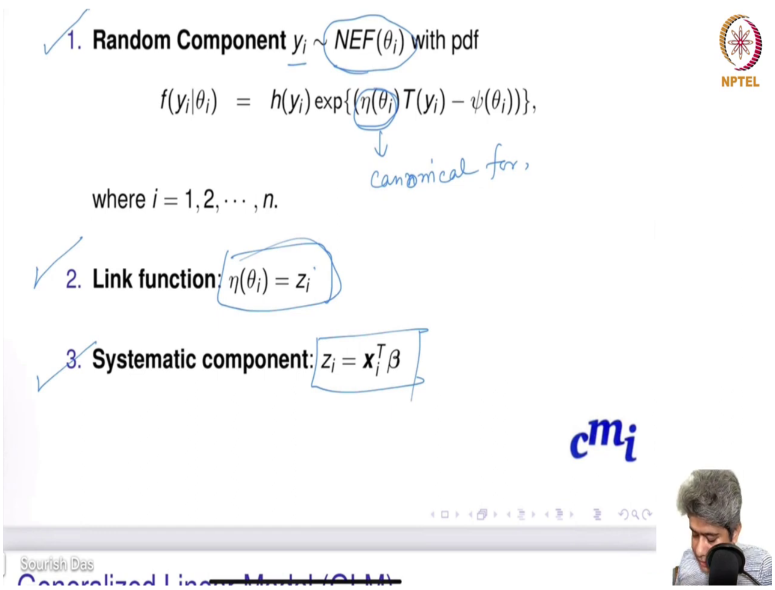
So, now if I build a predictive model based on natural exponential family that will be called general that it will be called generalized linear models. And this will work for any kind of response if your response is binary the if your response is continuous and if your response is count everything will work that is an interesting point.

If your response is binary then it must be a classification problem. So, if your response is binary then it is a classification problem right, then it is a binary classification problem and if your response is continuous then it is a regression problem correct. We knew so, far in machine learning they treats separately, but in statistics they all fall into the same problem class of problem called generalized linear model because binary and binary classification and regression they have all fall in the same class of natural exponential family.

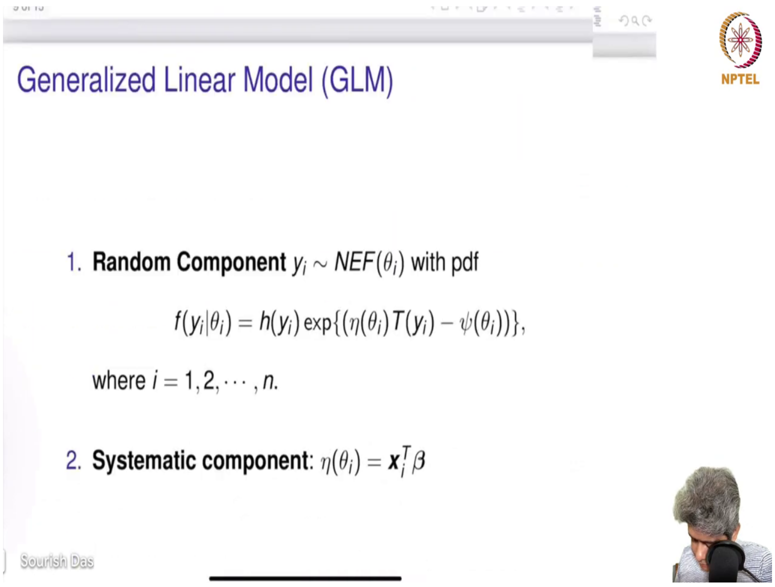
Hence any predictive model based on the natural exponential family will be simply called as generalized linear models ok. So, now what we are doing is very simple we have a random component y i because the general generalized linear model will have three component random component link function and systematic component.

So, y i for a some distribution natural exponential family and link. So, this eta theta i is typically called the canonical form eta theta i equal to z i and z i equal to x i transpose beta that is how we write the generalized linear model.

(Refer Slide Time: 14:29)

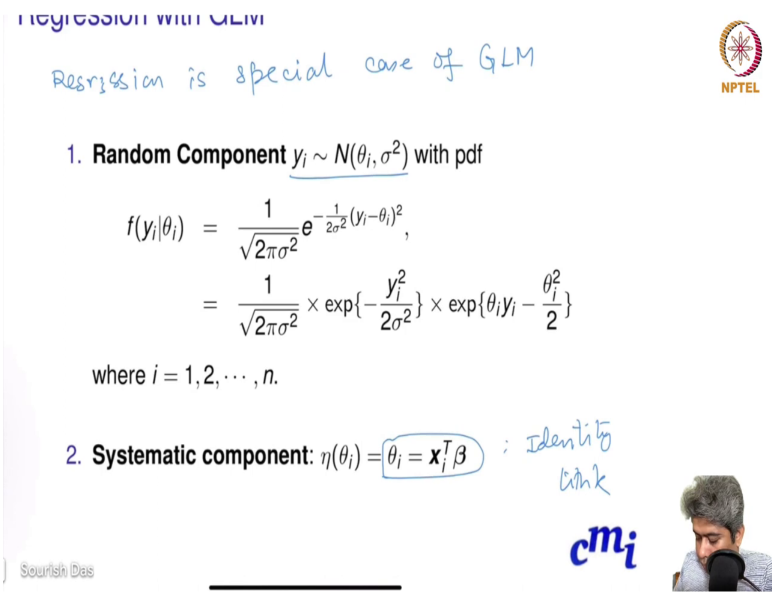


(Refer Slide Time: 14:45)



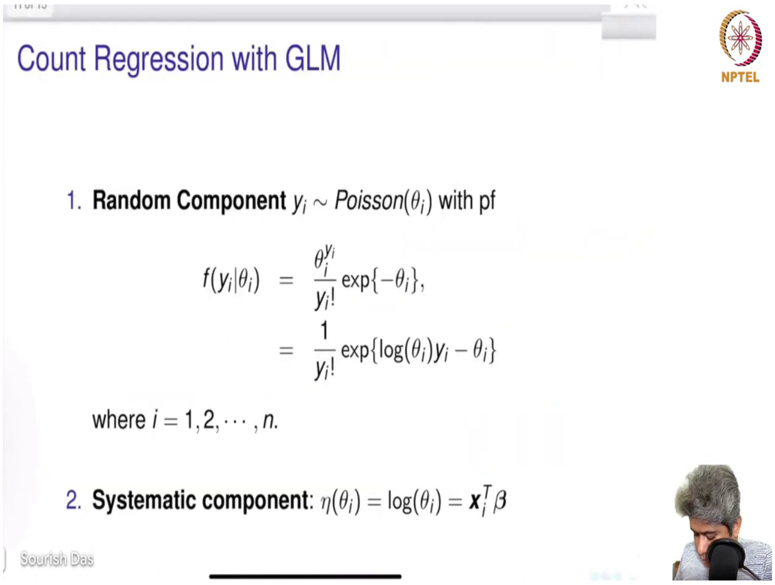
And sometimes we just simply write it in this way eta equal to x i transpose beta ok. Now, regression with g lm is simply you just say y equal to normal theta i sigma square.

(Refer Slide Time: 15:04)



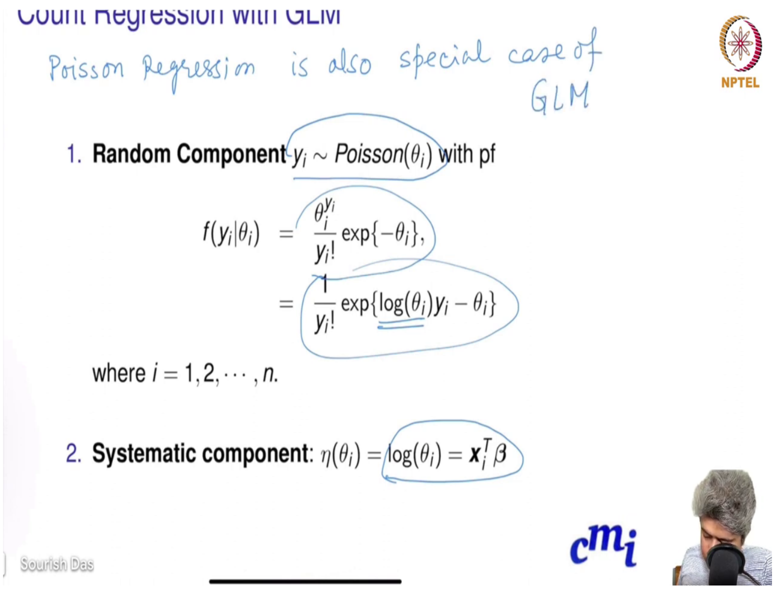
And then theta i equal to x i transpose beta it is called identity link simple very simple that is how a simple regression works in GLM framework. So, regression simple regression is special case of simple regression is special case of GLM ok.

(Refer Slide Time: 15:36)



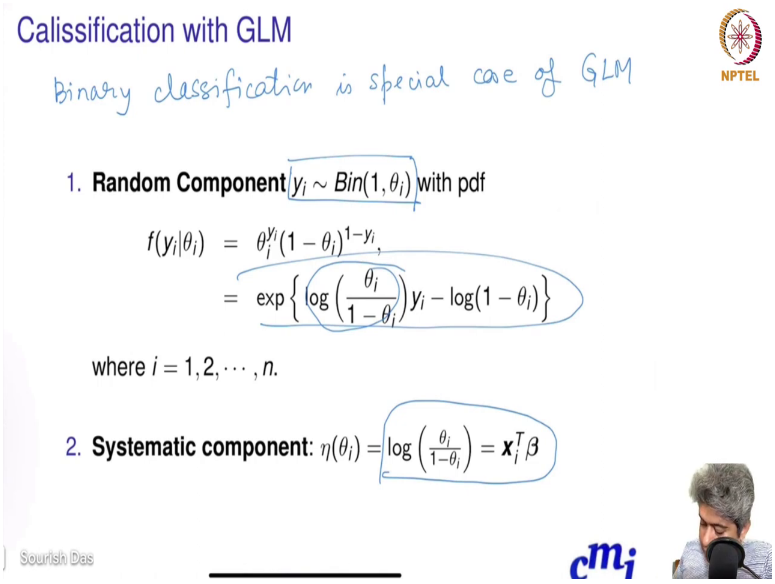
Now, count regression all you have to do that ok y follow Poisson theta i then you write it in this format and that ok then eta theta i is log theta i. So, we just say log theta i equal to x i transpose beta that is it.

(Refer Slide Time: 15:58)



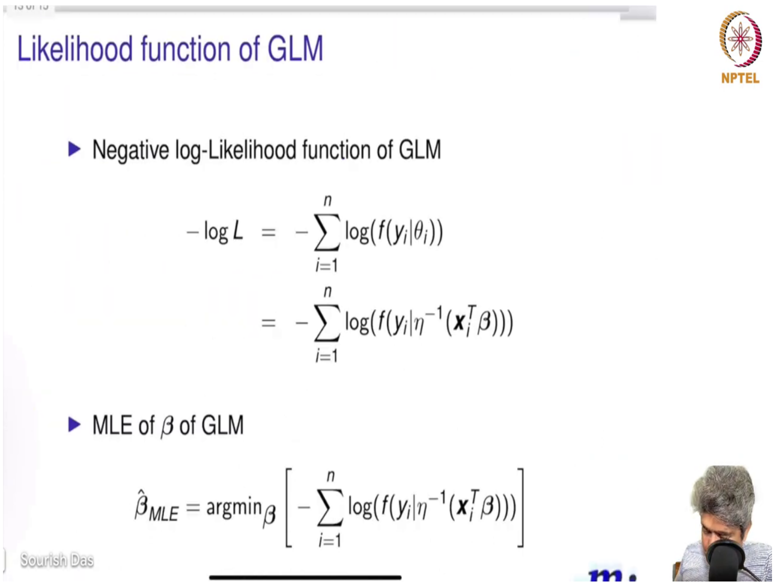
So, count regression or Poisson regression is also special case of GLM or Generalized Linear Model Poisson regression is also special case of GLM.

(Refer Slide Time: 16:19)



Then classification binary classification well if a y i follow binomial 1 theta i or Bernoulli theta then you can write it in this format and this is your canonical form. So, log theta i by 1 by theta i equal to x i transpose beta and damn i simple binary classification is now binary classification is special case of GLM ok.

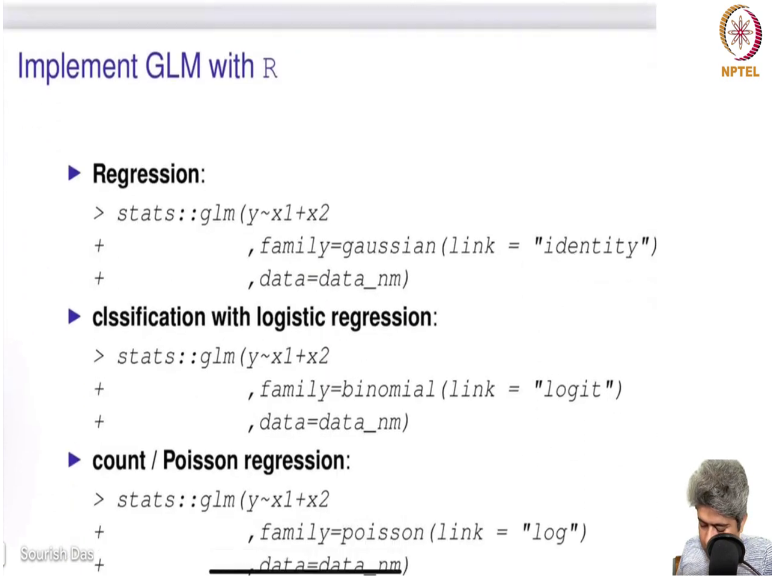
(Refer Slide Time: 17:08)



So, now all you have to do you can define each function of the natural exponential family separate ok like when I you go back to the natural exponential family the way you define it. So, a h y eta theta T, y and psi theta each of them you define as a separate form separate function in your program ok.

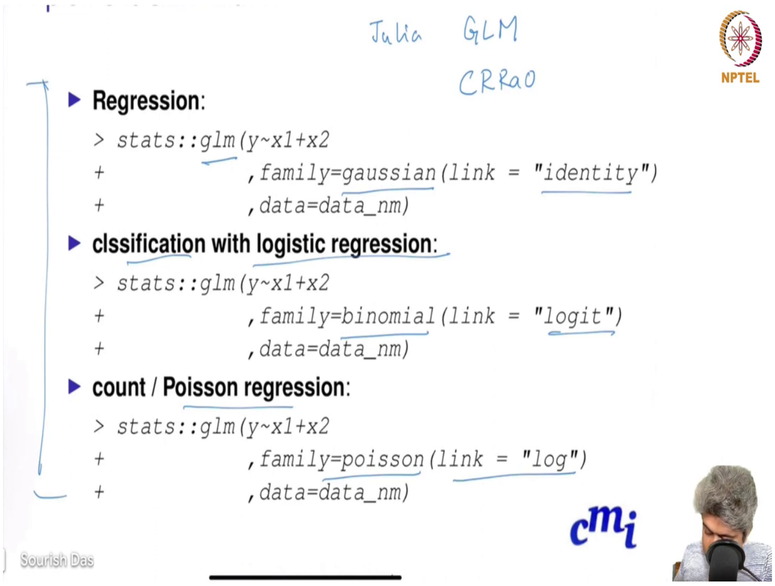
Now, if you just go here come here and define it as a negative log likelihood in that and then where theta i equal to eta inverse x i transpose beta and then just optimize it and you get the beta hat MLE that is it.

(Refer Slide Time: 17:56)



So, the implementation is very simple in GLM in r there is a stats package glm and you call family as Gaussian link equal to identity key. If it is a classification problem with logistic regression then we call family equal to binomial with logit link if it is Poisson regression you call Poisson with log link.

(Refer Slide Time: 18:25)



If it is Julia in Julia there is a GLM package our CRRao package is also can handle all three models. With this we will stop here and we will do some hands on with GLM.

Thank you very much see you soon.