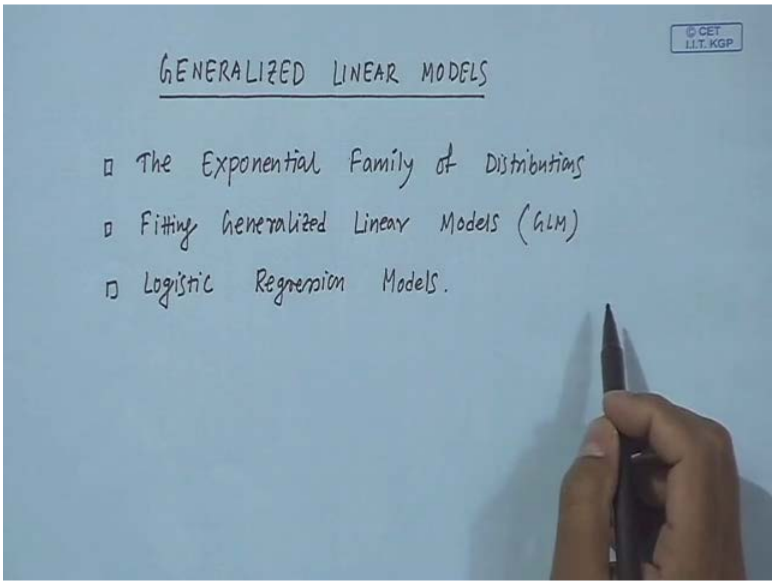
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 31
Generalized Linear Models (Contd.)

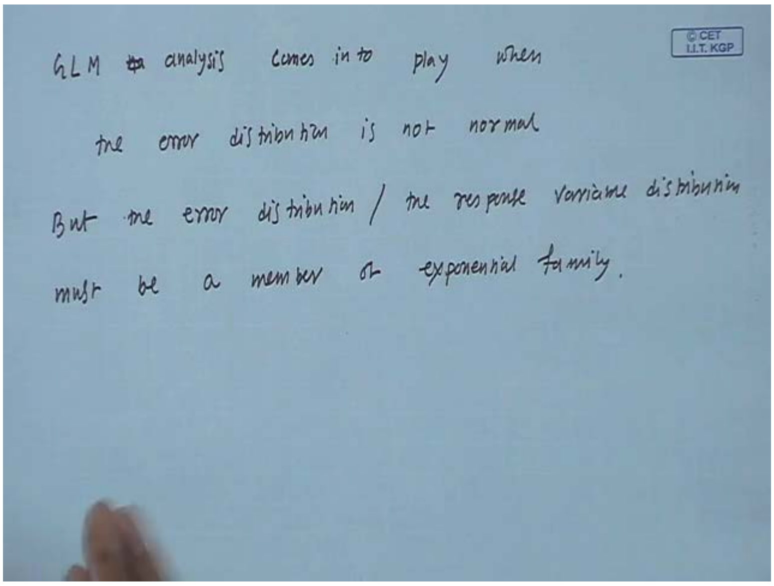
(Refer Slide Time: 00:30)



So, this is my second lecture on a Generalized Linear Models, and here is the content of this module. The exponential family of distributions, fitting generalized linear models and a logistic regression models, well. So, in a simple linear regression or in multiple linear regression model, we make a several assumptions on erratum like, the erratum has a mean 0, variance sigma square. And they are uncorrelated and also we assume that epsilon of the follow, a normal distribution with mean 0 and variance sigma square.

Now, in topic called a transformation and waiting to correct model inadequacy, we have to learnt how to deal with the situation when the assumption on constant variance. And also that uncorrelated assumption is violated and what we learnt in this topic is that, how to deal with the situation when the normality assumption is violated. That means, the response variable or the erratum they follow some order distribution not the normal distribution.

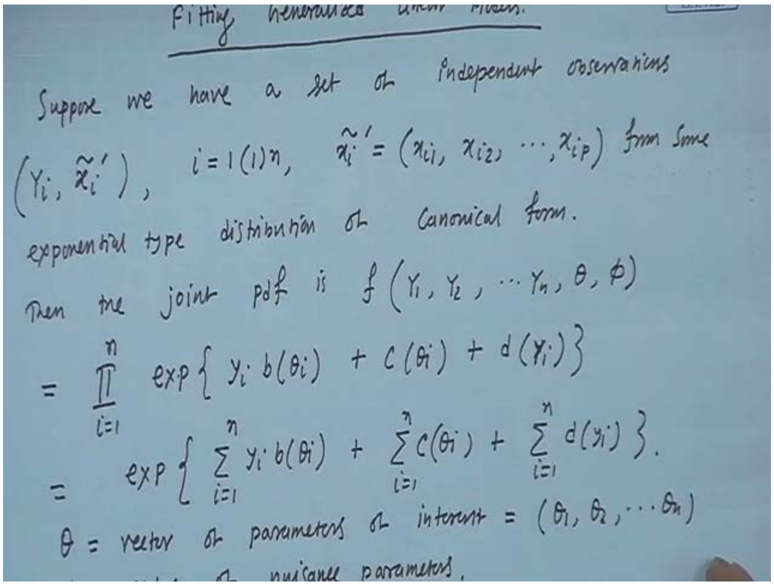
(Refer Slide Time: 02:23)



So, as I told in the previous class, this generalized linear model analysis comes into play, when the error distribution is not normal. So, the error distribution is not normal means, distribution of response variable is also not normal, but let me clear the fact that, the error distribution is not normal, but the error distribution or which is same as the response variable distribution must be a member of exponential family. So, this should be clear, like this generalized linear model is applied, when the error distribution is not normal.

But, the error distribution must follow a distribution from the exponential family, like we learnt in the previous class that, like normal distribution of course, it falls in exponential family. And in binomial, poser and then gamma, exponential, negative binomial, they are in all in exponential family, so what I will do; I sort of repeat this fitting a generalized linear models again.

(Refer Slide Time: 04:58)

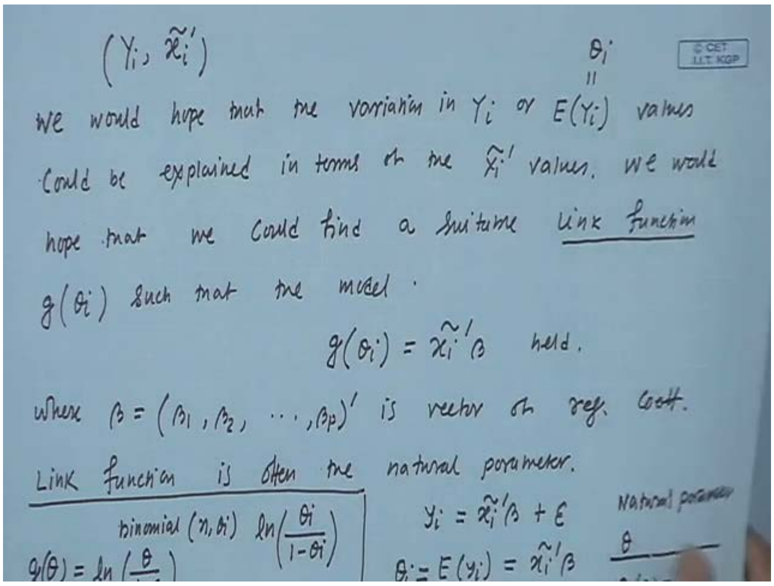


Because, this is very important fitting generalized linear models, suppose you have a set of independent observations, the observations are y i this is the response variable, and x i prime. So, these is a vector and suppose it has p components associated with p regressor variables, so these are the observation we have and we have n observations, i is from 1 to n. And as I told this x i prime, this is a x i 1, x i 2, up to x i p, and here this response variable is not form normal, so this is from some exponential type distribution of canonical form, so we know of when distribution is a exponential type.

And then the joint p d f, probability density function is, so f (y 1, y 2, y n, theta and phi), which is basically product of this marginal p d f. So, the marginal p d f is it is a exponential type, so the p d f is of this form exponential y i b theta i plus c theta i plus t y i. And since this distribution is a canonical form that is why, a y is equal to y and this is a product of marginal; so i is from 1 to n because the observations are independent that is why you can find the joint p d f just by multiplying the marginal p d f.

And this can be written as I wrote in the previous class, this is exponential sum i is from 1 to n y i b theta i plus c theta i; i is from 1 to n plus sum over i equal to 1 to n d y i. And here this is a theta, this theta is a vector of parameters of interest, here the parameter of interest is theta and this vector is a theta 1, theta 2, theta n. So, the i th observation is coming from an exponential type distribution with parameter theta i, and this phi is vector of nuisance parameter.

(Refer Slide Time: 10:30)



So, next what we want is that we are given say, y i and x i like the previous cases, so we are given the response variable and we are given the value for the p regressors, and what we want is that we want to explain the variability in y in terms of x i’s. But, the only problem here is that, if the y i is from the normal distribution, then we know how to fit a model between y i and x i, but here the only problem is that y i, the response the distribution of the response variable does not follow normal distribution here.

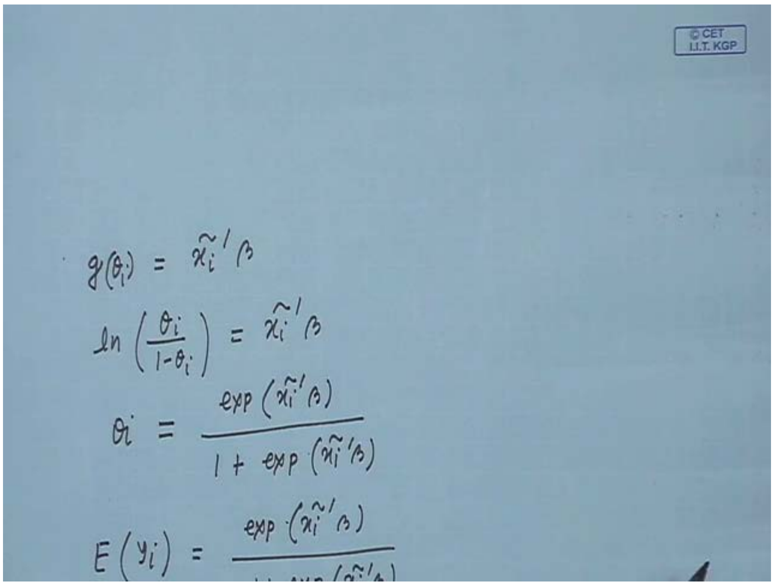
And then how to fit a appropriate model between the response variable and the regressor variable, so that is the main objective here. So, what we expect is that, we would hope that the variations in y i or say expectation of y i that is nothing but theta i and this all this theta i, theta 1, theta 2, theta n, they can all be different. So, the variance in y i or theta i values could be explained, in terms of the x i’s values and we would hope that we could find a suitable link function say g of theta i such that, the model is, a model g theta i is equal to x i prime beta this held.

Let me just complete my writing here and then I will try to explain this part little bit, so where beta is the regression coefficient, so beta is, beta 1, beta 2, beta p is vector of regression coefficient. And this link function is often the natural parameter, I am sure that you may face problem here understanding this part, but let me try to explain this one. So, in usual case, when y i is from normal distribution, what the model we fit is that, we fit the model y i is equal to x i prime beta plus epsilon, so this is a simple or multiple linear regression model.

And then this can be, I can write this model as say expectation of y i equal to x i prime beta, because of the fact that expectation of epsilon is equal to g to 0, so my model I can also write this model as theta i is equal to x i prime beta. So, this is a model in case of, so the model finally, the model is theta i is equal to x i prime beta, when the response variable is from the normal distribution. Now, I hope you can recall that the natural parameter for normal distribution is also in the exponential family, and the natural parameter for normal distribution is the natural parameter is theta.

If theta, suppose theta is mean, if I mean the link function is associated with this natural parameter, so the link function here, the function I am talking about this is g theta, so g theta is equal to the natural parameter theta. So, when it is normal my g theta is equal to theta, so that is why I fit the model theta equal to x i prime beta, now in the other case this suppose y does not follow normal distribution, it follows some other distribution from the exponential family, say binomial. Then my natural parameter for binomial is l n theta i by 1 minus theta i, so this theta is the probability of success in i th trial, so this is a natural parameter for the binomial case with parameter n theta i. So, this is a natural parameter in that case, in case of binomial, so my g theta is l n theta by 1 minus theta.

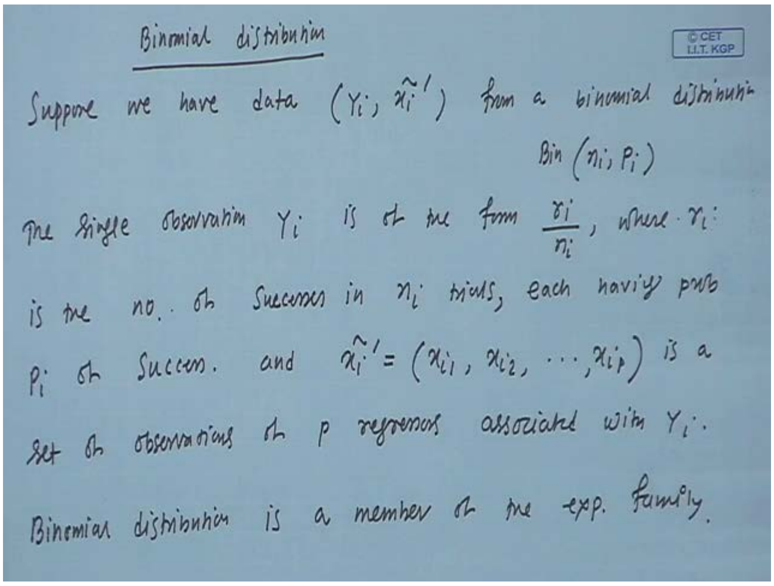
(Refer Slide Time: 18:13)



So, in case of binomial that we will go for the model of g theta i is equal to x i prime beta, and we know that in case of binomial this g theta is equal to l n theta by 1 minus theta, i equal to x i prime beta. So, here you can write this in the compact form also may be finally, the model is theta i you can write it as theta i is equal to exponential x i prime beta 1 plus exponential x i prime beta.

So, this is a model in case of binomial and this is nothing but the final model is expectation of y i is equal to exponential x i prime beta by 1 plus exponential beta. So, this is the model we have to show that means, the y equal to this plus epsilon is the model, in case of the response variable of follows binomial distribution. So, I will talk about this case right now of in detail, so let me consider this fitting generalized linear model in case of binomial distribution.

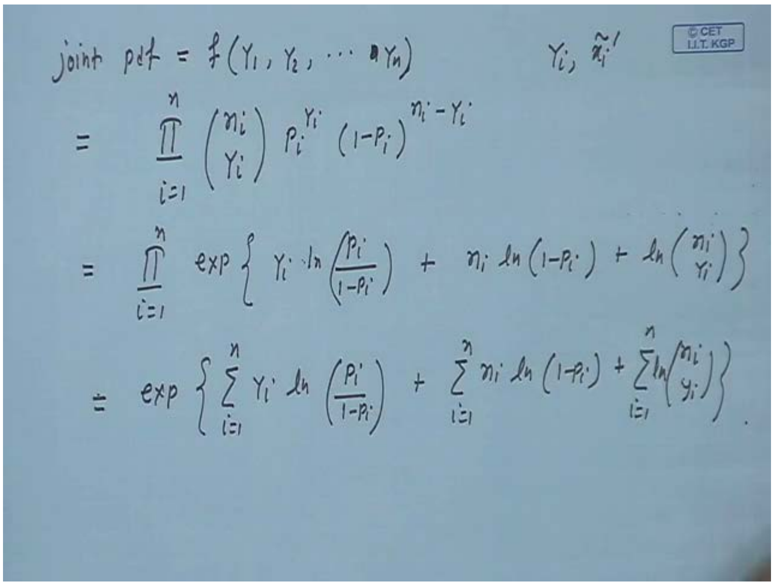
(Refer Slide Time: 20:18)



So, let me write it clearly, suppose we have data say (y i, x i prime) from a binomial distribution with parameter say binomial n i, let me write it p i, I am instead of theta I am writing p i, so p i is the parameter of interest and n i is are nuisance parameter. So, I have a set of observation from binomial distribution, then how to how to fit a model I already talked about this one, but I will write it very clearly here. Now, this y i the single observation y i is of the form r i by n i, where r i is the number of successes in n i trials.

So, binomial distribution and here y i is not really number of successes in n i trials, it is a proportion of success basically, each having probability P i of success. And this x i prime this one is basically x i 1, x i 2, x i p is a set of observations of P regressors, associated with y i. And we know that this binomial, I will also give an example to illustrate this part, this binomial distribution is a member of the exponential family. So, what I will do first is that, I will write down the joint p d f, because that one we get the natural parameter and also the link function.

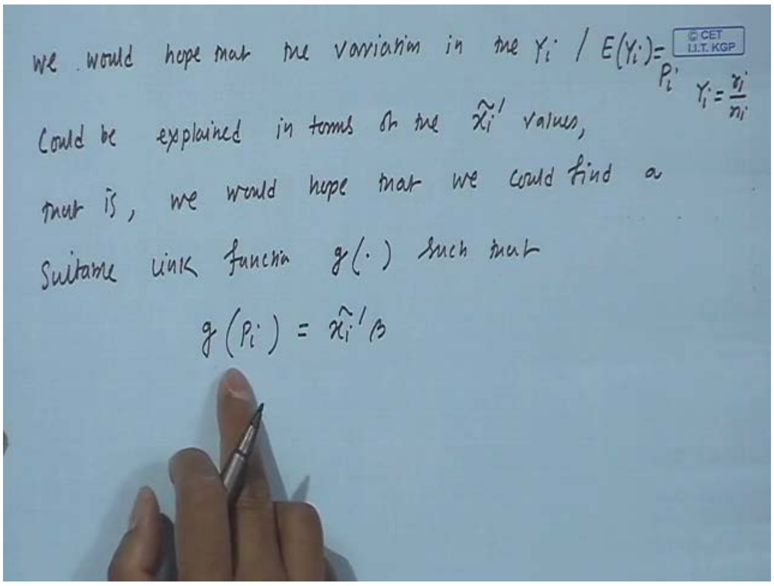
(Refer Slide Time: 24:33)



So, the joint p d f of y 1, y 2, y n, joint p d f, let me write it as a f of y 1, y 2, y n is equal to product of n i, y i, n i c y i, P i to the power of y i into 1 minus P i to the power of n i minus y i, from i equal to 1 to n. And this can be written of as i equal to 1 to n exponential say y i l n P i 1 minus P i it is not difficult to verify this plus n i l n 1 minus P i plus l n n i c y i.

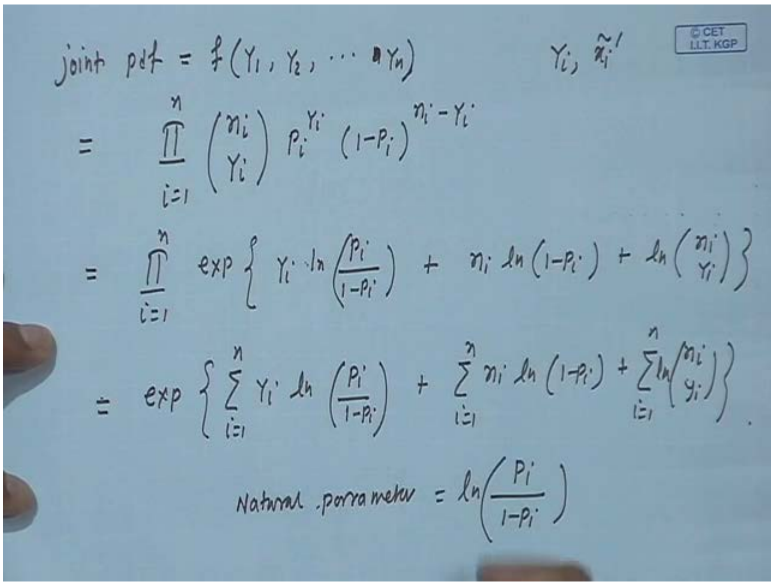
And then you can finally, write it as exponential just a sum over i equal to 1 n y i l n P i 1 minus P i plus sum n i l n 1 minus P y, i equal to 1 to n plus i equal to 1 to n n i c y i l n, this is a joint p d f of the observations, we have which are from the binomial distribution. Now, the same thing what we want we are given y i and x i prime, and we would try to explained the variability in y in terms of x i.

(Refer Slide Time: 27:27)



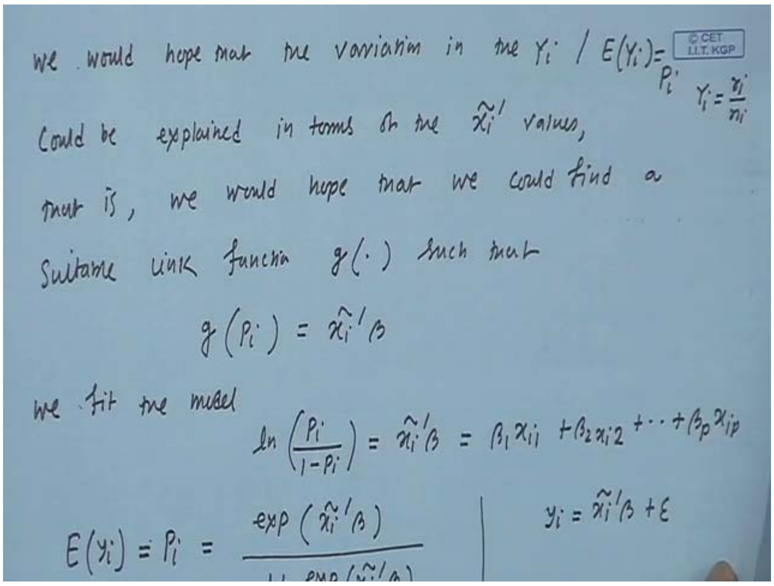
So, same thing, so we would hope that, the variation in the response variable y i or in expectation of y i, so here this generally it is n p, but here we are assuming again this y i the observation is number of successes by n i, so this one is basically P i. So, the variation in y i or P i could be could be explained in terms of the x i's values that is we would hope that, we could find a suitable link function. So, here function g such that, this g of P i is equal to x i prime beta, and this link function is obtained from the natural parameter.

(Refer Slide Time: 29:30)



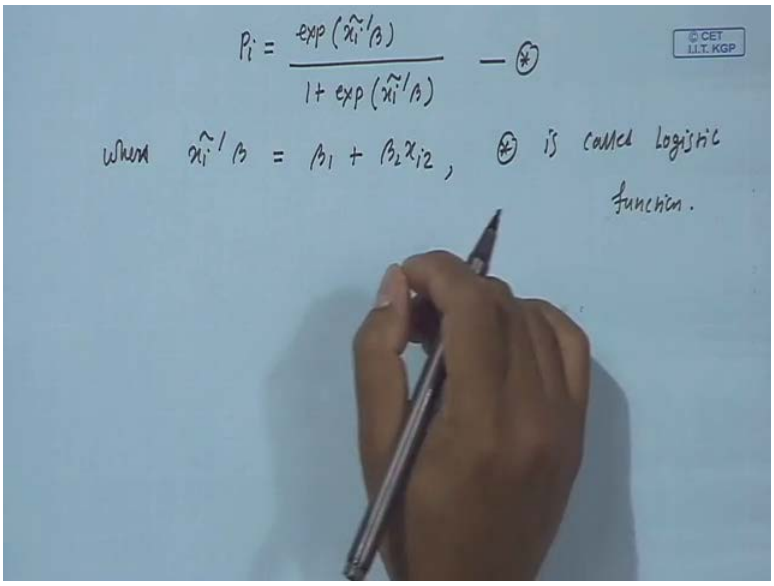
And for binomial distribution the natural parameter is, so the natural parameter here, meter is l n P i 1 minus P i, so this is basically g of P i.

(Refer Slide Time: 29:50)



So, that is why we fit the model, we fit the model l n p i by 1 minus P i equal to x i prime beta, so this one is nothing but beta 1 x i 1 plus beta 2 x i 2 plus beta p x i p, and this one can be written as finally, that p i is equal to exponential, as I wrote before also exponential x i prime beta by 1 exponential x i prime beta. So, this is same as writing expectation of y i is equal to this, and which is equivalent to say that y i is equal to this plus epsilon. So, this is a model instead of fitting y i equal to x i prime beta, which is the case for normal distribution, we fitting the model like y i is equal to this expression plus epsilon and the expectation of y i is this one.

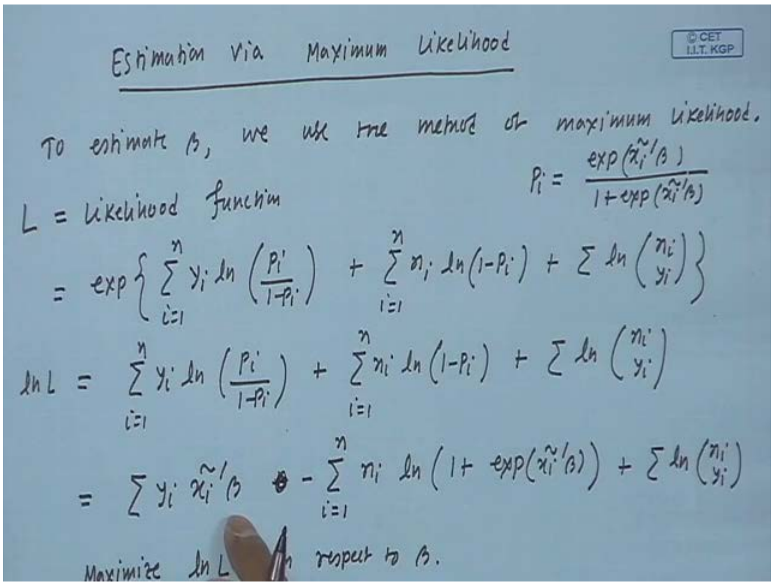
(Refer Slide Time: 31:50)



So, the model we got is finally is P i is equal to exponential x i prime beta by 1 plus exponential x i prime beta, now when x i prime beta is equal to beta 1 plus beta 2 x i. That means, only one regressor and the other one is of course, dummy variable you can x i 1, which is a dummy variable, this is one for all observations. When this is true let me call star, when this is a situation, then star is called logistic function, so we have the model with us, now this is this is the model we have to fit when the response variable is a binomial.

So, the model we got is that expectation of y i or which is equal to P i is equal to exponential x i prime beta by 1 plus x i prime beta, exponential. So, we have the model and then how do we fit the model, fitting the model means here the regression coefficients are beta 1, beta 2, beta P, so you have to estimate things. So, here we will use maximum likelihood method to fit them.

(Refer Slide Time: 34:06)



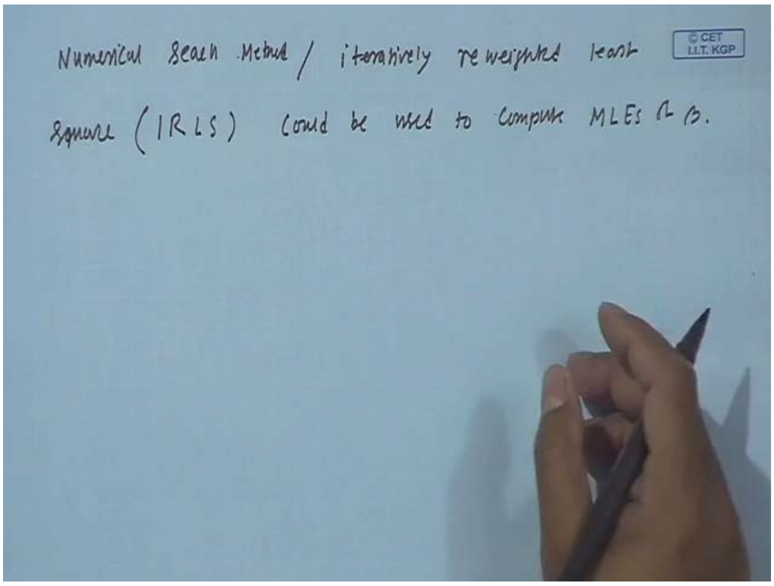
So, estimation via maximum likelihood, so to estimate beta we use the method of maximum likelihood. So, first what we do is that, we construct the likelihood function or compute the likelihood function l, which is the likelihood function is nothing but joint troll d of y 1, y 2, y n and we know that this is exponential just now we computed, this is y i l n P i 1 minus P i plus sum over i equal to 1 to n, i equal to 1 to n. n i l n 1 minus P i plus l n n i c y i, this is a likelihood function and then it is convenient to work with log likelihood, log likelihood is nothing but log of the likelihood function l n.

So, this simply it becomes sum over y i l n P i 1 minus P i i is equal to 1 to 1, so this is the log likelihood plus n i l n 1 minus P i, i is from 1 to n plus l n c i. So, this is a log likelihood and now what I want to do is which see, ultimately we have to estimate the parameter beta and we have to fit the model P i is equal to exponential x i prime beta by 1 plus exponential x i prime beta. And then what we will do is that, we will just a write this log likelihood in terms of beta, so you can check that this one is x prime beta, so this one is summation y i x i prime beta, because this l n P i by 1 minus P i equal to x prime beta from there only we get this one.

Plus i is from 1 to n you can check that this can be replaced by, you have to put minus here (Refer Time: 37:59) by n i l n 1 plus exponential x i prime beta it is not difficult to check this one. So, from here you can check that this l n 1 minus P i is equal to this plus say l n n i c y i. So, we have the likelihood function or log likelihood function in terms of beta now, so how do we estimate beta you maximize log likelihood l n with respect to beta.

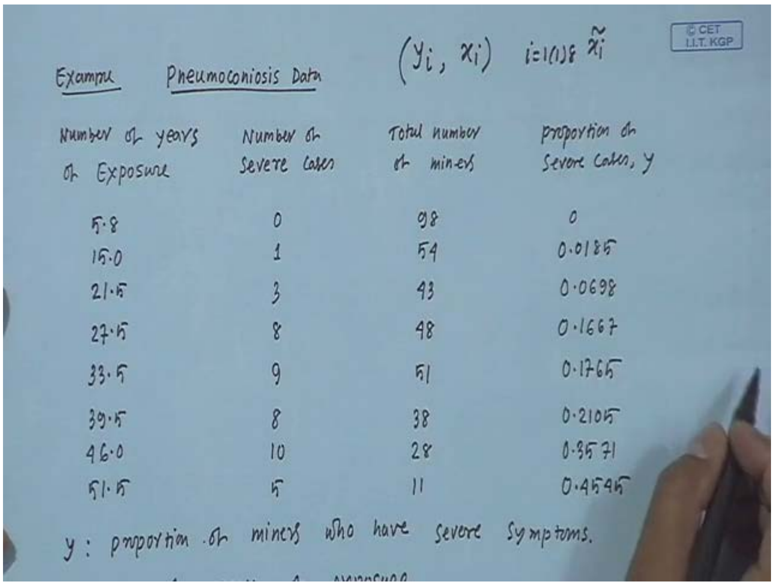
That means, so here this beta is a vector and it has P components, beta 1, beta 2, beta P, so you have log likelihood involving beta, now you differentiate this log likelihood with respect to beta 1, beta 2 and beta P. And so then you will get P equations, and you have P unknown and then you can solve for beta 1, beta 2 and beta P, so this is how you have to find the estimates of a regression coefficient beta and it is not so easy to do this for a given problem.

(Refer Slide Time: 40:01)



So, may be this numerical search method or something called a iteratively reweighted least square, this is IRLS could be used to compute maximum likelihood estimates of beta. So, now again to explain this example of binomial distribution consider, now I will give a numerical example to illustrate the binomial case.

(Refer Slide Time: 41:05)



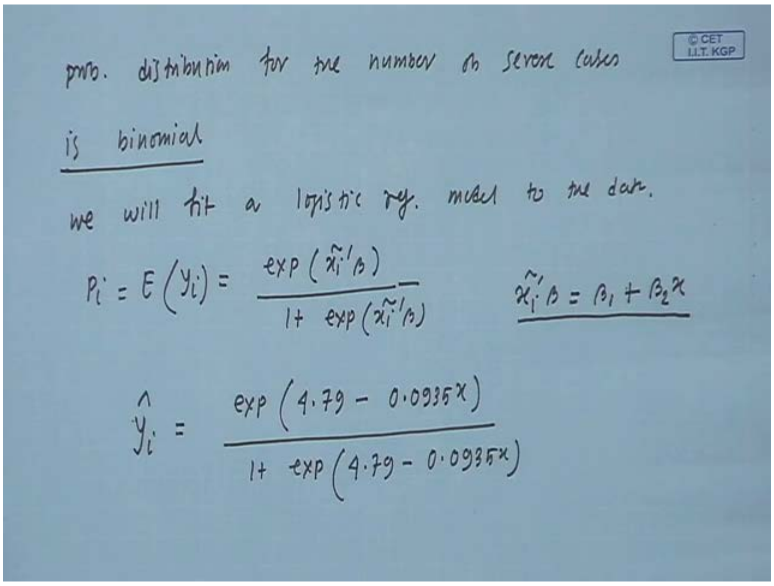
So, here we have data called a pneumoconiosis data and this is a long digits, this pneumoconiosis is a long digits that results from breathing in dust in coal mines. And here you have the data like number of years of exposure, and the data can be written in this way. So, number of years of expo exposure is say 5.8 years and total number of minor is 98, so this many workers number of severe cases is 0, so if I mean then the proportion of severe cases.

So, y is the proportion of severe cases and that is 0, so 0 by 98 is 0, so the number of years of exposure, if it is 5.8 or say 6 year, then the probability that somebody will be severely affected by this pneumoconiosis is a like 0. Similarly, you see that, if the number of year of exposure is more, then there are chances of severely affected by this digit. And here you can see if somebody is exposed for say almost like 50 years, then it is a almost the probability is a half that a person will be affected by this digits.

So, I am sure that understood the problem here, so I have the data now let me write in terms of my requirement like, so I have a response variable y i, so y i is the proportion of minors who have severe symptoms. So, this proportion, these are the proportion and I want to see whether the variation in this proportion can be explained in terms of the number of years of exposure and that is my x i. So, here I talked about this x i vector, so this vector is consisting of only one component, it is a simple linear simple regression model type of things.

So, I have this data for i equal to 1 2 3 4 5 6 7 8 for i equal to 1 to 8, so my y 8 is equal to 0.45 and my x 8 is 51.5 years, so what I want is that, I want to see whether the variation in y i or in the proportion of severe cases can be explained in terms of the number of exposures. But, the problem here is that this y i is not from the normal distribution, so it is sort of binomial if you I mean this number is of course, binomial number of severe cases is binomial.

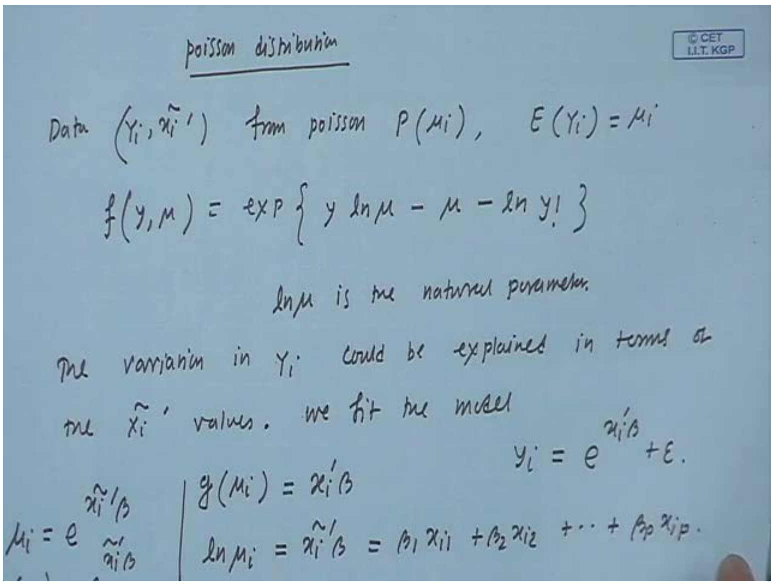
(Refer Slide Time: 45:14)



So, here the probability distribution for the number of severe cases is binomial, so we will fit a logistic model and there only one regressor, so we will fit a logistic regression model to the data. And my model is like a y i is equal to exponential x i prime beta by 1 plus exponential x i prime beta, and I should write this expectation of y i which is equal to p i basically. And here you must have observed this x i prime beta is equal to beta 1 plus of beta 2 x, because there is only one regressor that is the number of years of exposure.

And then you go for, so you have the model and then you know how to fit this model using maximum likelihood estimated and finally, you can check that the fitted model is y i hat, which is equal to exponential 4.79 minus 0.0935 x. That means, I am writing this beta 1 is this, and beta 2 is this, 1 plus exponential 4.79 minus 0.0935 x, so this is the binomial, I given an example to illustrate the binomial case, let me go for the Poisson distribution now.

(Refer Slide Time: 47:54)

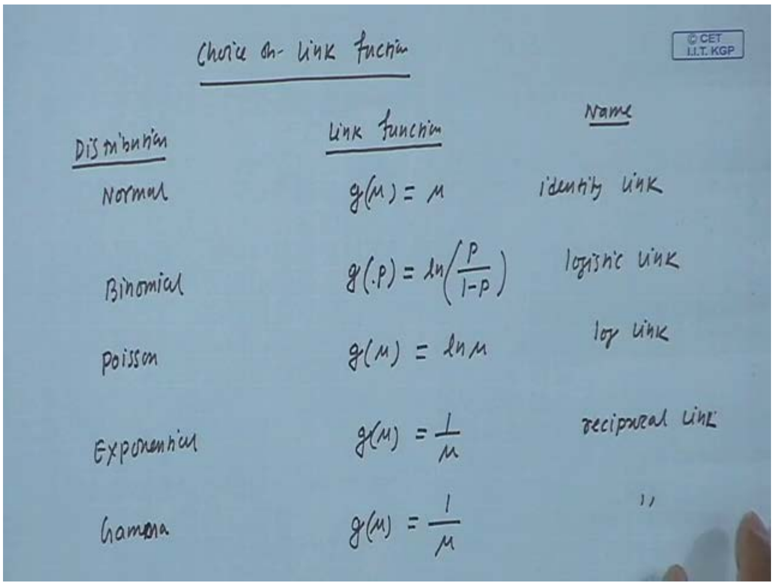


So, suppose we have data (y i, x i) prime form Poisson P with parameter say mu i, that means, expectation of y i is equal to mu i generally we write lambda i. So, that this distribution also is in the exponential family, and the probability mass function f y mu can be written as exponential y l n mu minus mu minus l n y factorial. And here the natural parameter is l n mu is the natural parameter, I am specific about this, because this from this natural parameter, we will get the link function.

So, the g mu, g mu is l n mu, so again, suppose we are given y i and x i and we want to explain the variability in the response variable y i, in terms of x I, so the variation in y i could be explained in terms of the x i's values. And the model we fit is that, we fit the model a like g mu i is equal to x i prime beta and we know that, this link function is equal to l n mu, so the model we fit is, so l n mu i is equal to x i prime beta, which is equal to beta 1 x i 1 plus beta 2 x i 2 plus beta P x i p.

And finally, you can write this as a mu i is equal to x exponential x i prime beta, this is a final model you have to fit and this is nothing but expectation of y i is equal to e to the power of x i prime beta, so this is same as writing that y i. So, you have to fit the model y i equal to e to the power of x i prime beta plus epsilon, so whereas, for the normal case it is y i equal to x i prime plus beta, because for the normal case your g mu is equal to mu that is why. Now, what I will do is that, I will talk about some reasonable choices of link function.

(Refer Slide Time: 52:20)



Suppose choice of this link function, because the model depends on this choice of link function, how the distribution was is normal distribution and the link function. So, in case of normal, you see I am just bringing one slide from my previous lecture, (Refer Time: 53:03) here you can see the a natural parameter is mu, you can forget this sigma square, because this is nuisance parameter. So, you can write it simply mu, so then the link function will be g mu equal to mu and this is called the identity link, in case of binomial we know we have just establish the model.

So, g p let me write p, p is the probability of success in one trial, this is l n P by 1 minus P, this is called logistic link these are the name, and the Poisson my g mu equal to l n mu, this is called log link. And for exponential my g mu is equal to 1 by mu, and this is called reciprocal link and of course, for the gamma distribution is same because exponential is a particular case of, gamma distribution it is g mu equal to 1 by mu, it is also called a reciprocal link. So, in this module, so we have learnt if the error distribution or the distribution of the response variable is not normal, but it is from some exponential family. The distribution is from exponential family, then how to deal with the situation, how to fit a model, so we have to stop now that is all.

Thank you.