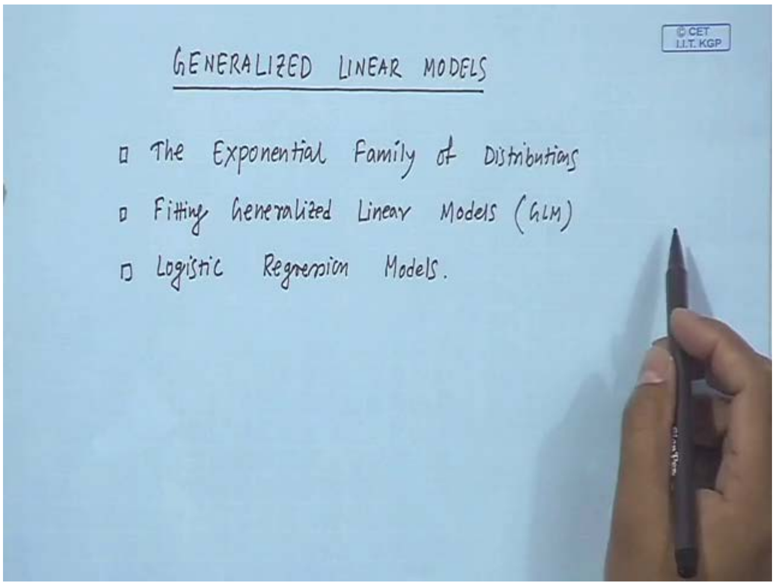
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institution of Technology, Kharagpur

Lecture - 30
Generalized Linear Models

Hi, so today, I will start a new module called generalized linear models.

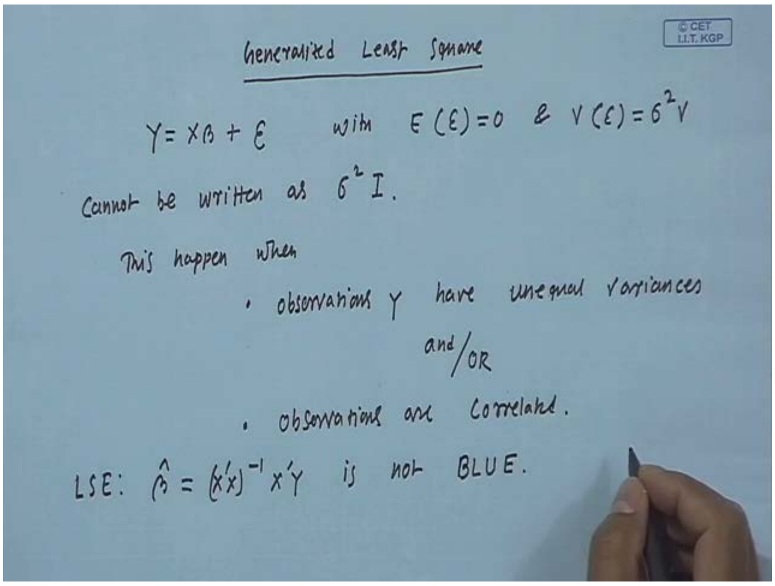
(Refer Slide Time: 00:30)



And, here is the content of this module. 1st will talk about exponential family of distributions and then, fitting generalized linear models and logistic regression models ok. So, before I start this module I want to recall. We talked about a module called transformation and weighting to correct a model inadequacy. And there, we have started something called generalized least square of which weighted least square is a particular case. And, the generalized least square is concerned about the application of ordinary least square technique in situation where y equal to x beta plus epsilon. This is the model and exception of epsilon is equal to 0 but the variance of epsilon is equal to sigma square into v.

So, this v is the variance covariance matrix of the added term which cannot be written in the form of sigma square into i. So, let me write this thing in detail.

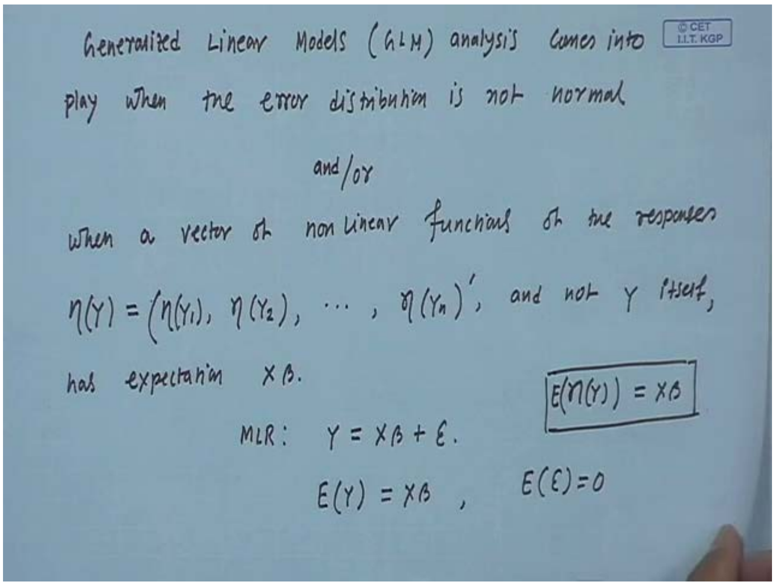
(Refer Slide Time: 02:14)



So, the generalized least square so, this one is concerned about the application of ordinary least square technique in the situation where Y equal to X beat plus epsilon with exception of epsilon is equal to 0. And, variance of this epsilon is the variance, co variance matrix is equal to sigma square into V and this cannot be written as sigma square I ok. So, we have started that, you know this happen, when the observations Y have unequal variances and or observations are co related. That is why you know in this variance co variance matrix V the up diagonal elements are not equal to 0. So, in either case you know the conditions of gross mass of theorems are highlighted.

So, the least square estimate that is beta hat equal to X prime X inverse, X prime Y is not the best linear unbiased estimated. So, here you know we have started in the generalized least square, we have stared transformation on this module to get the best linear unbiased estimated. And also, in that module like, that is transformations and weighting to correct model inadequacy we talked about variance stabilization transformation, which deals with the situation when the response variable are having, you know inequality in variance ok.

(Refer Slide Time: 05:58)

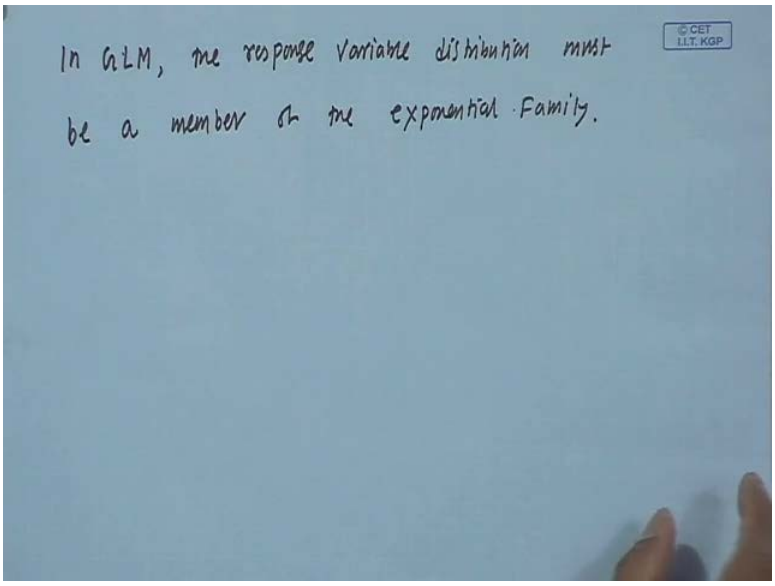


So, the generalized linear models, that is a G L M analysis comes into play when the error distribution is not normal. So, the distribution of error is not normal or which is equivalent to say that, the distribution of response variable is not normal. In that case, we need to use generalized linear model. So, you should understand the different between the generalized linear model and the generalized least square because generalized least square is used to deal with non constant variance in response variable. Or of course, when the observations are correlated whereas know this generalized linear model is used when the error distribution is not normal ok.

So, either the error distribution is not normal and or when a vector of non-linear functions of the response is, that is eta Y which is equal to: eta Y 1, eta y 2 and eta Y n. This vector and, not Y itself has exception X beta. So, I am not in position to explain this thing at this moment but what I can say here is that, in case of multiple linear regression model, we consider the model Y equal to X beta plus epsilon which is same as, we consider the model like expectation of Y is equal Y beta because of the fact that expectation of epsilon is equal to 0. So, in the usual linear model expectation of Y equal to X beta ok. So, that can be written in the linear combination of the regression coefficients but here you cannot. This is not true.

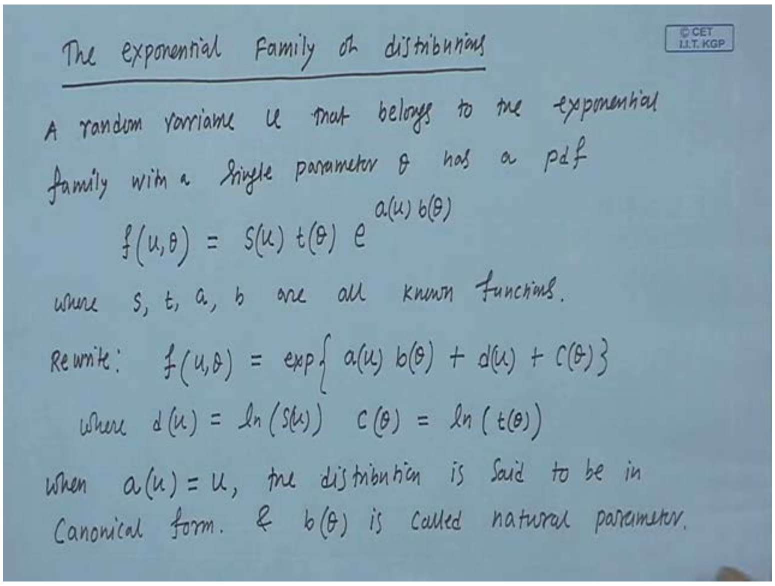
So here, instead of expectation of Y equal to X beta, there exist non-linear function, it is eta may be eta Y, expectation of eta Y is equal to X beta. Anyway, we will come to this point at the end of this topic or module. So, let me mention one more thing. When we use, you know this generalized linear model.

(Refer Slide Time: 10:55)



In generalized linear model G L M, the response variable distribution is not normal. That is what we know I mean, if it is not normal then we go for generalized linear model. But, the response variable distribution must be a member of the exponential family. So, next what will do is that we will learn, what you mean by this exponential family.

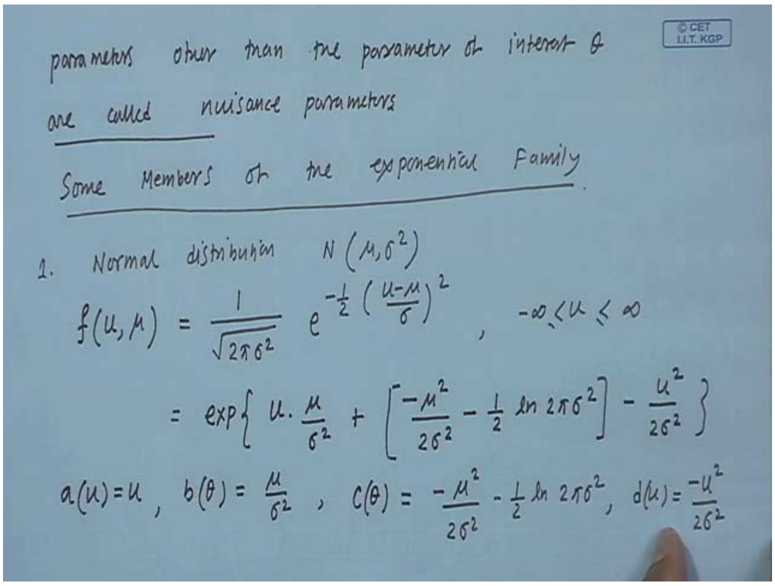
(Refer Slide Time: 12:04)



So, the exponential family of distribution ok so, a random variable u. Here, we denote you know random variable by u, usually use the rotation x or y, that belongs to the exponential family with a single parameter theta, has a probability density function f u theta, which is of the form s u t theta into e to the power of a u b theta ok. So, if the p d f of the random variable u is of this form then, we say that distribution in the exponential family where s, t, a, b are all known functions. So, let me rewrite this p d f. So, I can write this f u theta as exponential a u b theta and then plus d u plus c theta ok, where d u is equal to l n s u.

So, this is a log base e and c theta is equal to l n t theta, right ok. So, the p d f of random variable u which is exponential family can be written in this form and when a u is equal to u, that is a u is a identity function, the distribution if the distribution is said to be in canonical form and this is important. This b theta is called the natural parameter ok. And also, there could be you know several parameters in distribution like, if you consider as a binominal distribution it has 2 parameter: one is a n which is the total number of trials and p, (Refer Time: 17:55) p is the probability of success in one trial. So, here we need to decide which one is correct parameter interest. If p is the parameter of interest then the other parameter n is called nuisance parameter. So, let me write it formally.

(Refer Slide Time: 18:21)

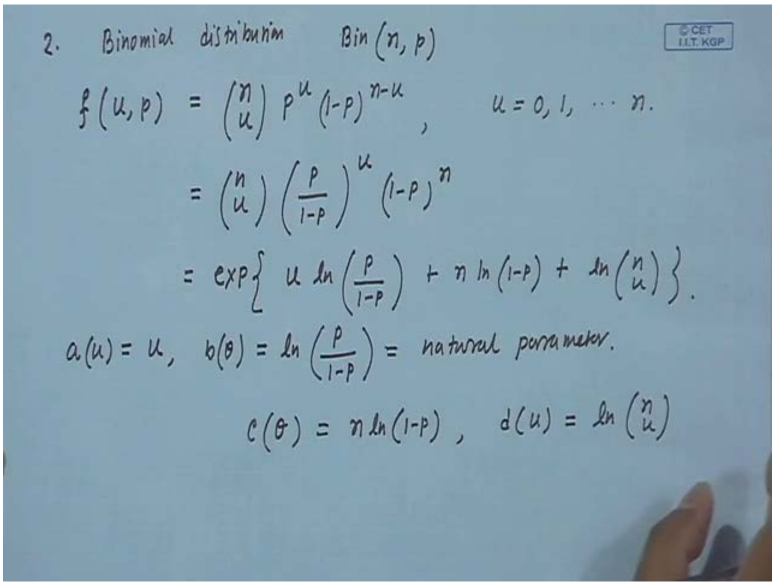


So, parameters other than the parameter of interest theta are called nuisance parameters ok. So, let me talk about some members of the exponential family ok. So, the 1st one is say normal distribution with a parameter nu and sigma square right. So, you know that the p d f of, now see you know whether the p d f of normal distribution which is f u mu. So, I am writing if u mu because u is the random variable and mu is the parameter of interest here and sigma square is a nuisance parameter. So, let me see whether the p d f can be written in that form. So, this one is equal to 1 by root over of 2 pi sigma square into e to the power of minus half u minus mu by sigma square and you know that this u is from minus infinity to plus infinity.

Now, whether I can write this in this form, say exponential you can check that this is u into mu by sigma square plus minus mu square by 2 sigma square minus half l n 2 pi sigma square so, I will put this in bracket, minus u square by 2 sigma square ok. So, this term is coming from here and this three are basically this exponent here ok. Now here, I need to indentify all minus a, u, b theta, c theta and d u. So, here I can see that a u is equal to u. This a u and b theta is mu by sigma square and what is c theta? In fact c mu but, you know theta stands for parameter of interest so, here the parameter of interest is mu and this involve mu. So, c theta is equal to minus mu square by 2 sigma square minus half log 2 pi sigma square and d u is equal to minus u square by sigma square.

So since, a u is equal to u so normal distribution is in canonical form and this is the natural parameter, mu is the mu by sigma square is natural. I mean in fact mu is the natural parameter because mu is the parameter of interest ok. So, next will, I mean basically I will try to write down many distributions which are in the exponential family.

(Refer Slide Time: 24:19)

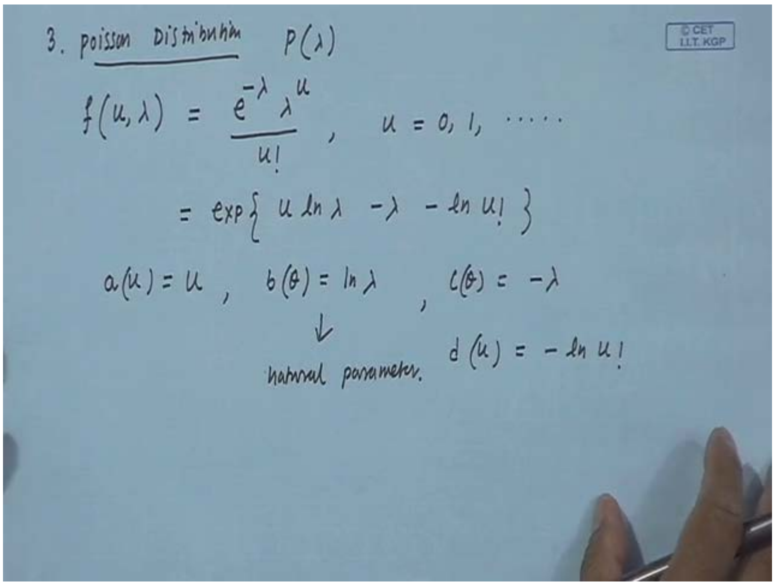


The next one is, see binominal distribution. So, you follows binominal distribution by parameter n and p and p is the parameter of interest and n is the nuisance parameter. So, here you call to mass function but, if u p as you know this is, n c u. So, here the u stands for the number of successes in n trials when, the purity of success in is p ok. So, the purity mass function is n c u p to the power of u and into 1 minus p to the power of n minus u and the u, the range of u is the number of successes would be: 0, 1 and it could up to n, out of n trials ok. Now, see whether this can be you know written in the exponential form, that the form you know I talk about. So, this is equal to n c u. I can write this as p by 1 minus p to the power of u 1 minus p to the power of n.

Now, let me write this as exponential u log p by 1 minus p plus n log 1 minus p plus log n c u ok. So, let me identify now. The functions a u here is equal to u b theta, which is the natural parameter, b theta is l n. I repeat again this natural parameter is important thing, we need to know what is the natural parameter for particular distribution. So, b theta is natural parameter which is log p by 1 minus p, is the natural parameter. And of course, then c theta is n l n 1 minus p and d u is equal to l n n c u. So, this does not involve any parameter of interest ok. So, c theta means it should involve parameter of interest in this function so the parameter of interest is p.

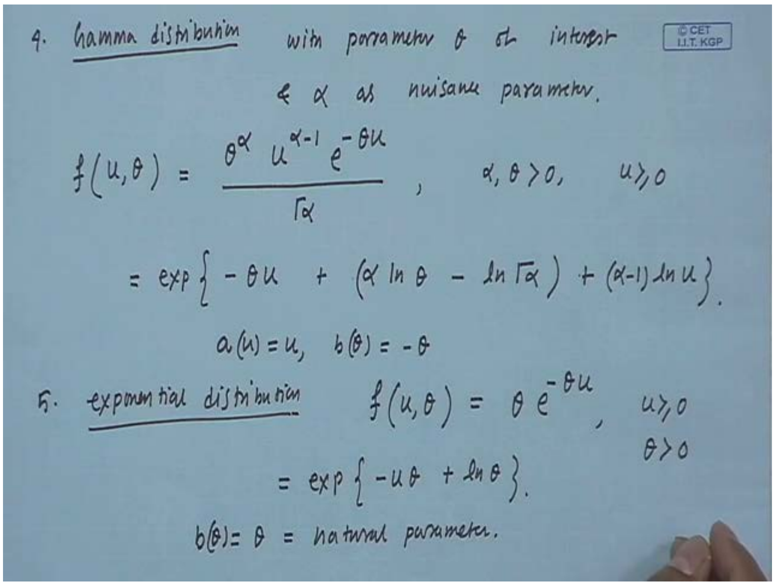
So, c theta is this quantity ok and also this binominal distribution is in the canonical form. Next, will talk about Poisson distribution.

(Refer Slide Time: 28:10)



Poisson distribution, with parameter lambda and you know that the provity density function for Poisson is f u lambda is equal to e to the power of minus lambda to the power of u by u factorial and here, u is equal to: 0, 1 up to infinity. And, this can be written as exponential u l n lambda minus lambda minus l n log u factorial ok. So, it clearly, here my a u is equal to u, so it is canonical form, my b theta is log lambda. So, this is the parameter, this is natural parameter and my c theta is equal to minus lambda and d u is equal to minus l n u factorial and here is the natural parameter. So, this is regarding the Poisson distribution. So, it is in the exponential family.

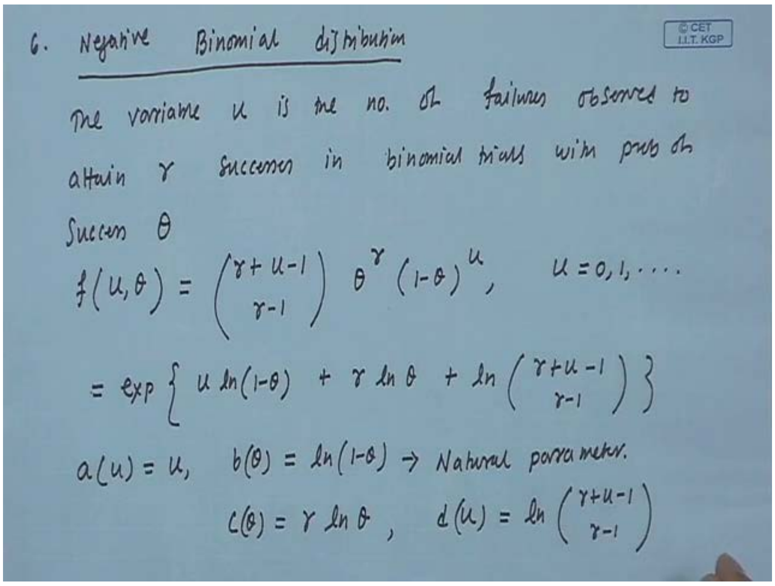
(Refer Slide Time: 30:10)



Next, let me talk about number 4, gamma distribution ok with parameter theta of interest and alpha as nuisance parameter. And, here is the p d f so, f u theta is equal to theta to the power of alpha u to the power of alpha minus 1 e to the power of minus theta u by gamma alpha. So here, all this alpha m theta, they are greater than 0 and u is greater than equal to 0. Now, you write it in the exponential form. So, exponential minus theta u plus alpha log theta minus log gamma alpha, you put them in one bracket because this basically (Refer Time: 32:00) b theta plus alpha minus 1 l n u, is coming from here right.

So, now you can identify the you know this is a u is equal to u b theta is equal to minus theta and this equal to c theta and d u ok. The next one is called exponential distribution. So here, the p d f of this exponential distribution f u theta which is obtained by just puttying alpha equal to 1 here so, this is theta into e to the power of minus theta u. So, u is greater than equal to 0 and theta is greater than 0. So, this can be written as exponential minus u theta plus l n theta ok. So, you understood that you know the a u equal to u and b theta equal to theta and c theta equal to log n theta right. So, the natural parameter here, so here b theta is equal to theta which is natural parameter ok. So, you are almost done. Next, we will talk about one more distribution which is called negative binominal distribution.

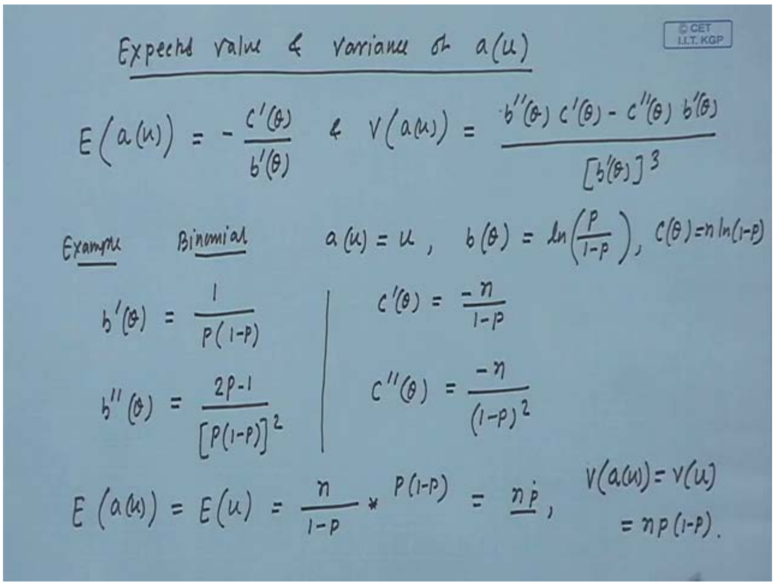
(Refer Slide Time: 34:12)



So, negative binominal distribution so, I hope that you know you understand the experiment here. The variable u is the number of failures observed to attain r successes in binominal trial with probability of success theta. And then, the provity mass function or this one can be written as f u theta is equal to r plus u minus 1 c r minus 1. I hope you understand why this is, this is the provity mass function, theta to the power of r 1 minus theta to the power of u. So u, the number of failure before to obtained r successes it can be it can start from 0, 1 anything ok. So, this can be, my concerned is check whether this negative binominal distribution this, is in the exponential family or not. So, this is exponential u log 1 minus theta plus r log theta plus l n r plus u minus 1 c r minus 1.

So, you must understood that here my a u is equal to u, my b theta is equal to l n 1 minus theta so, the natural parameter this is the natural parameter. I am talking about every time because it is this is important to for the generalized linear model and my c theta is of course, r l n theta and c d u is equal to l n r plus u minus 1 r minus 1. So, this shows that the negative binominal distribution is in the exponential family ok. So, next will talk about the expected value and variance of this a u.

(Refer Slide Time: 38:14)



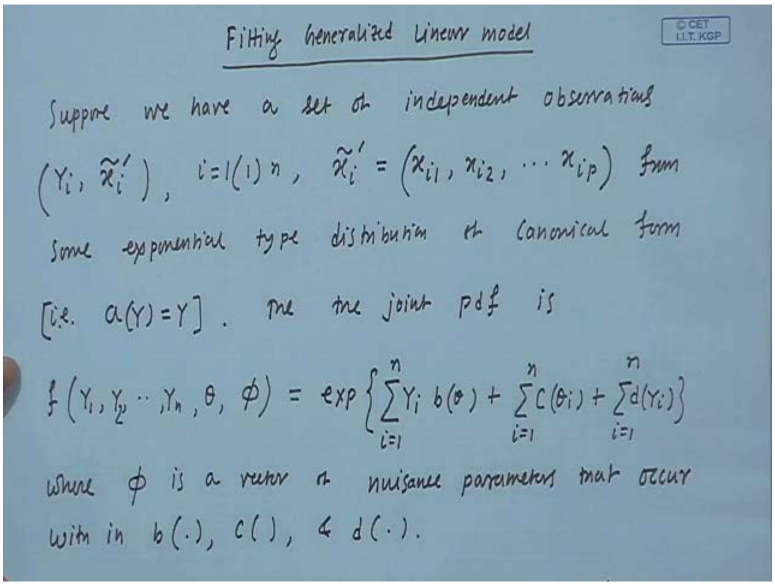
So, expected value and variance of a u so, I am not going to derive it in terms of c theta and b theta. So, you can check that this is e of a u, the expected value of function is minus c prime theta by b prime theta and the variance of a u is equal to b double prime theta c prime theta minus c double prime theta. So, this transfer the double durability of c theta with respect to theta so, b prime theta by b prime theta to the power of 3, just you know you believe me this is correct. I am just giving one example of say, binominal. In case of binominal you just check that we had this a u equal to u and we had b theta, natural parameter that was log p by 1 minus p and c theta was n log n 1 minus p ok. Now, if you compute say, so you know b theta so you can compute b prime theta that is, equal to 1 by p into 1 minus p.

So, you can compute b double prime theta, double derivative that is equal to twice p minus 1 by p into 1 minus p whole square and similarly, you do for c theta. So, c prime theta is equal to minus n by 1 minus p, c double prime theta is equal to minus n by 1 minus p whole square. So, now we can check that, you know that for binominal distribution excepted of u n p so, you can check that expectation of a u which is nothing but binominal it is expectation of u which is equal to minus c prime theta. So, that is n by 1 minus p by p theta so, that is p into 1 minus p so, that is n p. You know that for binominal distribution excepted value is n p and you can check that variance of a u is equal to variance of u here which is equal to n p into 1 minus p. Just put log all this here ok.

So, this is sort of you know preparation for the generalized least square because, we told that you know generalized least square is used when distribution of response variable is not normal but it is distribution is from the exponential family. So, now we have you know idea about distributions are may be this is not interested list but, we know some distribution which are in the exponential family like: binominal, Poisson, normal, gamma, exponential, negative binominal. These are the example we just prove they are in the exponential family. Now, the thing is that suppose you have a set of observation like: x i, y i and the response variable y i is not in formal and it is say from it is from the binominal distribution, why I follows binominal with parameter n i, p i.

So, p i is the parameter of interest and n i is the nuisance parameter. Then, how to deal with such situation because till now, we know that we talked about only if, y follow is normal using the Gauss Markov theorem normal also in the independent identical distribution. Then, by Gauss Markov theorem we know that the least estimate provide the best linear unbiased estimated for the regression co efficient. So, now will talk about how to fit a generalized in the situation when the response variable is not normal but, it is the distribution of the responsible variable is from exponential family ok.

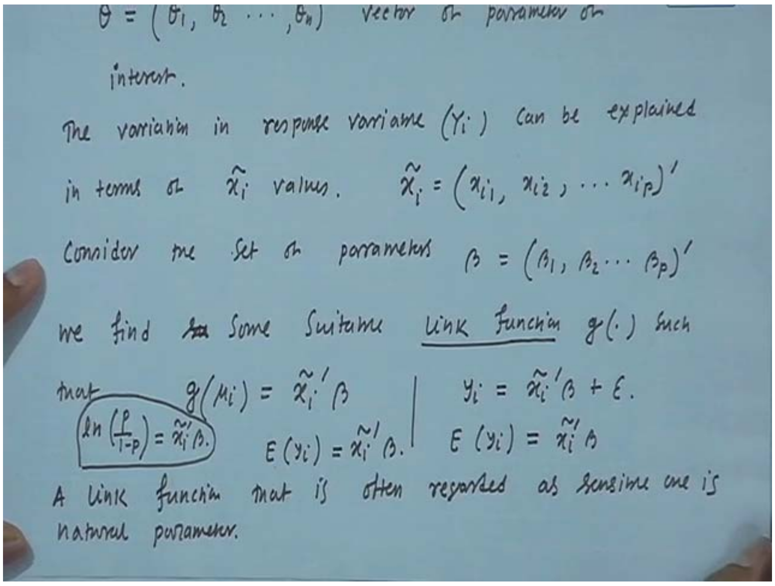
(Refer Slide Time: 44:42)



So, here is a fitting of fitting generalized linear model. As I told suppose, we have a set of independent observations. Suppose, my observations are: y i, y i, so y I, y i. So, if you consider one regression then it is a simple regression but I can generalized, I can make it of vector. So, this is my i-th observation and then, this vector is say it has it consist of r regresses p regresses so, this is my observation for i equal 1 to n. So, I have n observation response variable y i and there are several variable. Let me write this what is this x i prime is that it is x i 1, x i 2, x i p. That means it is a p component vector right and we have a set of independent observation from some exponential type distribution of canonical form. That means, that is a y is equal to y then the joint provity density function is so, I have n observation I just I have independent.

So, the joint provity density function is f y 1, y 2, y n and theta and fee. So, this one is joint free density function is nothing but the product of marginal's. So, this one is equal to you know that observation is exponential of the distribution. So, I can write as exponential y i b theta, I can write this because a y is equal to y that is why plus c theta i plus d y i. But, this is the p d f only for the i-th observation and once you multiple the marginal's here just have to put summation here, for i equal to1 to n, i equal to 1 to n, i equal to 1 to n. So, where this fee is a vector of nuisance parameter that occur within b c and d, ok.

(Refer Slide Time: 50:16)



And, my theta is: theta 1, theta 2, theta n vectors of parameter of interest ok. Now, what we want is that, so we talked about the p d f and the variance in y in response variable y i can be explained in terms of x i values ok. Let me give some time. I will try to explain, what is the different between this generalized liner model and then linear model. So, here my x is regresses variable. So, this is basically x i 1, x i 2, x i p. So, I want to explain the variability in y using the regresses variable that is, what you find the relation between response variable and regression variable, this is the whole purpose of this course also. Consider the parameters that regression co efficient, consider the set of parameters beta which is equal to: beta 1, beta 2, beat p, prime ok.

Now, what we do is that we find some suitable link function. This is important link function, say g such that g of mu i is equal to x i prime beta ok. Let me just explain now, see in usual case what happen is that we consider that the model. So, what is that mu i, mu i is expectation of y i. So, in usual case what happen is that, we consider the model y i equal to say x i prime beta plus epsilon and then of course, expectation of y I, the ordinary case is equal to x i prime beta, that is all. But here, if the response variable is not in normal, if it is from some exponential family then x prime b, the regression variable explain the variability in g mu i. So, this is nothing but you know g of expectation of y i. So here, instead of writing expectation of y i is equal to x prime beta, we write g of expectation of y equal to x prime beta ok.

And, this link function, a link function that is often regarded as a sensible one is natural parameter. So, I will talk about this again in next class. But, let me say if this response variable is from is following binominal distribution then here, this link function will be there are the link function was what that will be the natural parameters. So, that is l n p by 1 minus p which will equal to x i prime beta and in case of normal distribution, we know the natural parameter is mu. So, in case of normal distribution this g mu is nothing but mu.

So, when we assume that y as followed that normal distribution, we can just write expectation of y i is equal to, you can go with this model. But, for the other distribution we need to choose this g function, this link function which is nothing but natural parameter ok. So, we will be talking about this again in the next class. So, today we have to stop now.

Thank you.