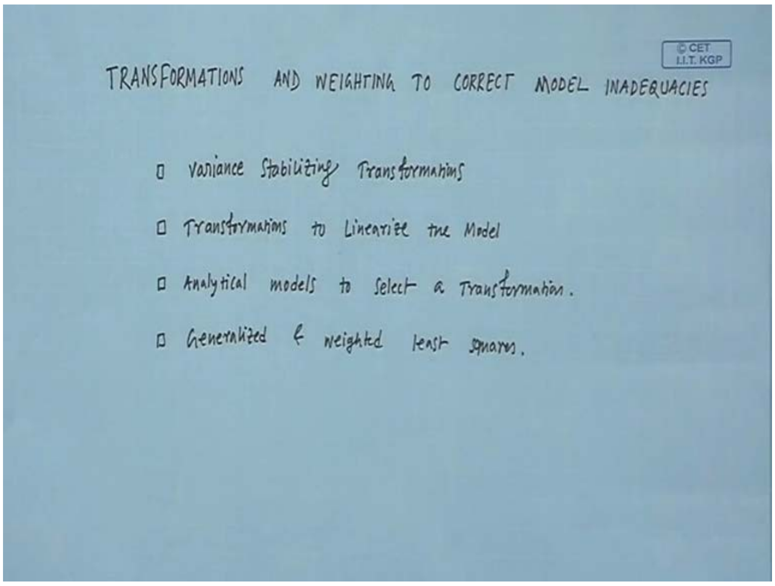
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture No. - 21
Transformations and Weighting to Correct Model Inadequacies

(Refer Slide Time: 00:19)



TRANSFORMATIONS AND WEIGHTING TO CORRECT MODEL INADEQUACIES

© CET I.I.T. KGP

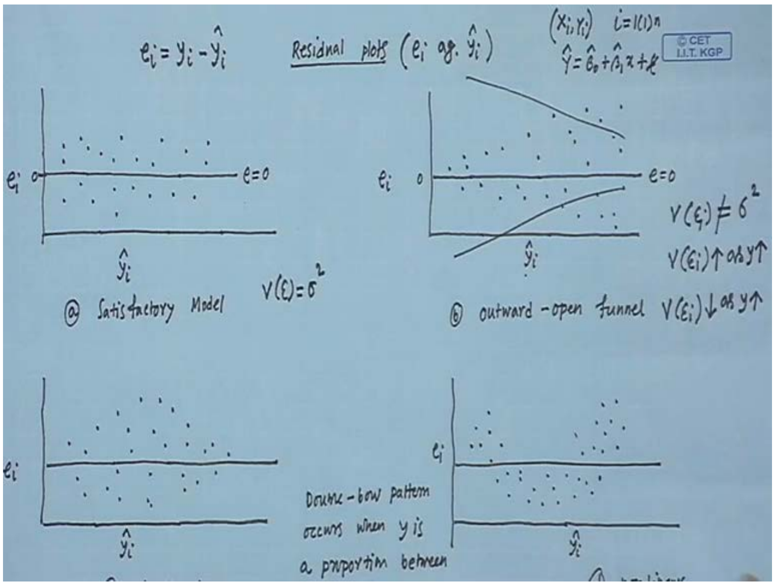
- □ Variance Stabilizing Transformations
- □ Transformations to Linearize the Model
- □ Analytical models to select a Transformation.
- □ Generalized & weighted least squares.

Hi, so, today we will start a new module called transformations and waiting to correct module inadequacy. Here is the content of this module. It consists of: variance stabilizing transformations, transformations to linearize the model, analytic models to select a transformations and finally, generalized and weighted least square. So, before I start a talking about this module I want to talk about the objective of this module. So, for this you know you first recall the simple linear regression model y equal to beta naught plus beta 1 x plus epsilon where the epsilon is the error term.

And similarly, in the multiple linear regression model we have Y equal to x beta plus epsilon and while fitting the simple linear regression model or multiple linear regression model we make some assumptions. The 1st one is the error term say epsilon i as expected value is equal to 0 and variance is equal to sigma square and they are uncorrelated and also we assume that this error terms epsilon i follow the normal distribution. So, normal with mean 0 and variance sigma square and epsilon i r independent and identically distributed random variable. So, this normal assumption is particularly required to test several hypotheses on regression coefficients and also to find the confidence interval for the regression coefficients.

Now, in the previous model called model adequacy checking we have studied different techniques to check whether the basic assumptions are satisfied or not. And, the purpose of this module is that if the basic assumptions are not satisfied if the some of assumptions are violated. Then how we can handle the situation? So, here first we will recall particularly that residual plot was very important to check the basic assumptions. And, then we will study in this module how to handle the situation if some of the basic assumptions or not satisfied?

(Refer Slide Time: 04:28)


$e_i = y_i - \hat{y}_i$
Residual plots (e_i ag. $\hat{y}_i$)
(X_i, Y_i) $i = 1(1)n$
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$
© CET I.I.T. KGP
e_i
0
$e=0$
$\hat{y}_i$
ⓐ Satisfactory Model
$V(\varepsilon) = \sigma^2$
e_i
0
$e=0$
$V(\varepsilon_i) \neq \sigma^2$
$\hat{y}_i$
$V(\varepsilon_i) \uparrow$ as $y \uparrow$
ⓑ outward-open funnel
$V(\varepsilon_i) \downarrow$ as $y \uparrow$
e_i
$\hat{y}_i$
Double-bow pattern
occurs when y is
a proportion between
e_i
$\hat{y}_i$

So, first let me recall the residual plot which is a very important tool to check whether the assumptions are correct or not. So, suppose you have given a set of observation: Y i X i. So, Y i is the response variable and X i is the regressor variable and you are given n observations. And then, you know how to fit simple linear regression model like this. So, Y hat is equal to beta naught hat plus beta 1 hat into x and once you have this fit it model you can find the residual, the i-th residual is y i minus y i hat. So, this y i is the absorbed response and this is estimated response and e i is the difference. This is called the residual and this is also you know most specifically this is called regular residual.

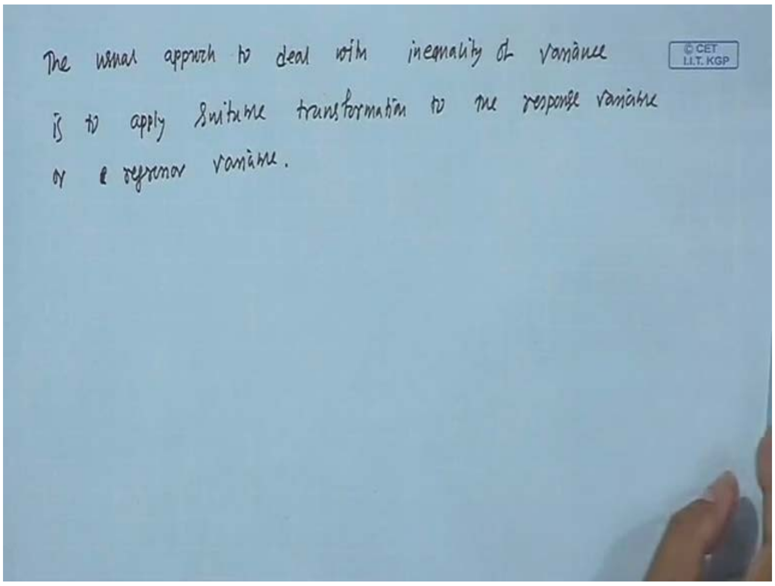
Now, what is the residual plot? The residual plot is the plot of residual e i against the fitted response y i hat. So, here is the scatter plot of the residuals against the fitted response. And, this is the line e equal to 0 and if you see the residuals are sort of centered about the line e equal to 0 and then the model is a satisfactory model. And here, this sort of plot you know suggest that the assumption variance of epsilon is equal to the sigma square is satisfied.

Now, look at this scatter plot here. Forget about these two lines. So, if you see this is called the outward open funnel if you see the residuals here you can see the residual value increases as y i hat increases and this is called outward open funnel and in this situation if this occurs. Then we sort of conclude that the constant variance assumption is not correct. So, we cannot assume that variance of epsilon i equal to sigma square for all i. So, this is not true and what actually happen here is that the variance of epsilon i increase as y increases.

Now, instead of this outward open funnel it could be like inward open funnel that means e i, the residual decreases as y i hat increases. So, in that case also the constant variance assumption is not satisfied and in case of inward open funnel, variance of epsilon i decreases as y increases. And, the other situation could be, this is called double bow. So, here you see the scatter plot of the residual and this sort of scatter plot occurs when y i is proportion and the response variable y is in between 0 and 1. So, this sort of scatter plot also indicates that the constant variance assumption is validated ok. And, here is the final I mean the 4th scatter plot. This sort of scatted plot is called nonlinear and this sort of nonlinear scatter plot indicates that the relationship between the response variable y and the regressor variable x is not linear.

So now, if you see that using residual plot or some other technique you learn in model adequacy checking, if you see that model assumptions are violated. Then specifically if the constant variance assumption is violated then we consider some transformation either on the response variable or in the regression variable to make the variance constant.

(Refer Slide Time: 10:16)

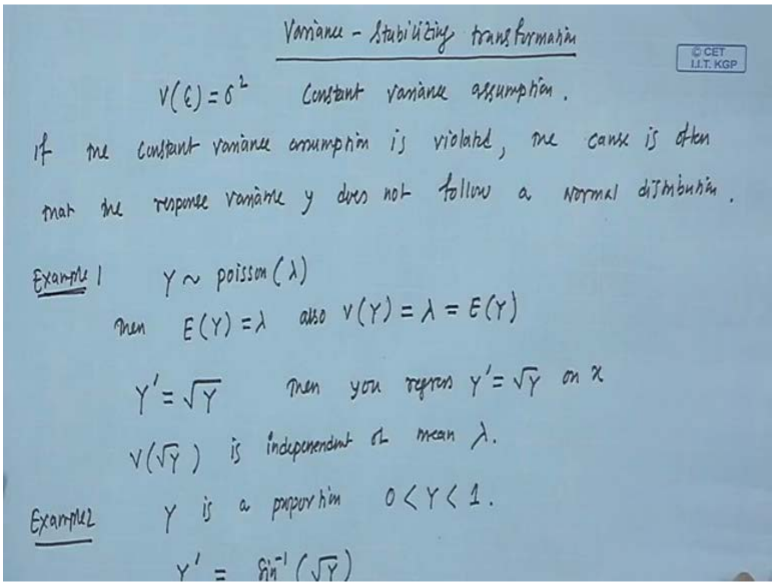


The usual approach to deal with inequality of variance is to apply suitable transformation to the response variable or a regressor variable.

© CET I.I.T. KGP

So, the usual approach to deal with inequality of variance is to apply suitable transformation to the response variable or regressor variable. So, first we will talk about the variance stabilizing transformations.

(Refer Slide Time: 11:25)



Variance - Stabilizing transformation

© CET I.I.T. KGP

$V(\epsilon) = \sigma^2$ Constant variance assumption.

If the constant variance assumption is violated, the cause is often that the response variable y does not follow a normal distribution.

Example 1 $Y \sim$ poisson (λ)

then $E(Y) = \lambda$ also $V(Y) = \lambda = E(Y)$

$Y' = \sqrt{Y}$ then you regress $Y' = \sqrt{Y}$ on x

$V(\sqrt{Y})$ is independent of mean λ.

Example 2 Y is a proportion $0 < Y < 1$.

$Y' = \sin^{-1}(\sqrt{Y})$

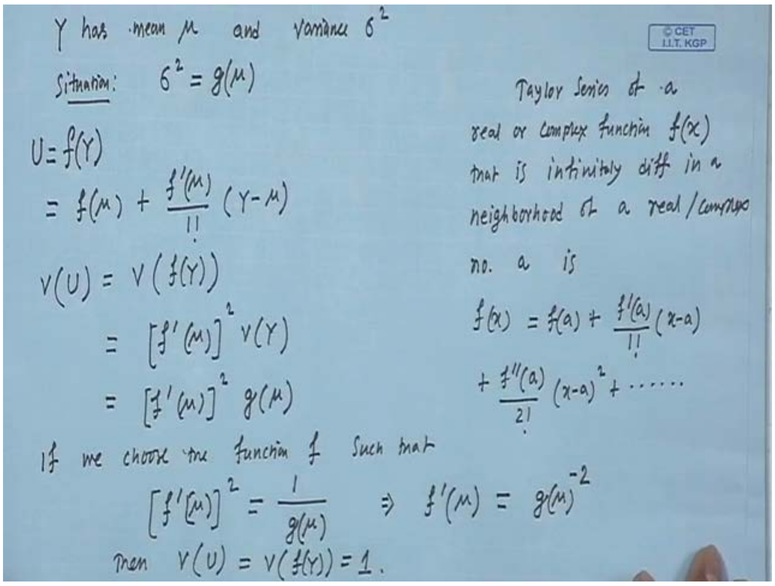
So, what we assume is that variance of epsilon is equal to sigma square. So, this is what is called the consent variance assumption. Now, if the constant variance assumption is violated the cause is often that the response variable Y does not follow a normal distribution. So, let me take some example, say example 1. Well, suppose the response variable Y follows Poisson distribution with parameter lambda. Then we know that expected value of Y is equal to lambda and also variance of Y is lambda. So, here you see that the variance of the response variable is the function of expectation. So, this is nothing but lambda is nothing but expectation of Y. So, in this case what we have to do is that we take some transformation on the response variable to make this variance constant.

Now, if you take the transformation say Y prime which is equal to root Y and then you regress Y prime which is equal to root Y on x and you can check that, it is not difficult to check that variance of root Y is independent of mean lambda. So in 2nd example, suppose, the response variable Y is a proportion in between 0 to 1 and when Y is the proportion between 0 and 1 we have seen in the residual plot that the residual plot sort of follow double bow pattern.

And here, we take the transformation Y prime which is equal to sign inverse square root of Y to make the variance of the response variable constant. Well, so if you see that the constant variant variance assumption is violated then in that case you know most like the response variable is not from the normal distribution. It follows some other distribution where variance is the function of mean and we have absorbed in case of Poisson distribution if the response variable y follows Poisson distribution then the transformation we took is that Y prime is equal to square root of Y to make the variance constant.

It is not difficult to check that variance of Y prime which is equal to variance of square root of Y is constant. And similarly, in case if Y proportions between 0 and 1 then we take the transformation Y prime is equal to sin inverse square root of Y. Question is: how do you decide about which transformation to take? So, in this variance stabilizing transformation we learn this we will talk about how to decide on which transformation to take to make the variance constant.

(Refer Slide Time: 18:32)



Y has mean μ and variance σ^2

Situation: $\sigma^2 = g(\mu)$

$U = f(Y)$

$= f(\mu) + \frac{f'(\mu)}{1!}(Y-\mu)$

$V(U) = V(f(Y))$

$= [f'(\mu)]^2 V(Y)$

$= [f'(\mu)]^2 g(\mu)$

if we choose the function f such that

$[f'(\mu)]^2 = \frac{1}{g(\mu)} \Rightarrow f'(\mu) = g(\mu)^{-2}$

Then $V(U) = V(f(Y)) = 1$.

Taylor Series of a real or complex function $f(x)$ that is infinitely diff in a neighborhood of a real/complex no. a is

$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots\cdots$

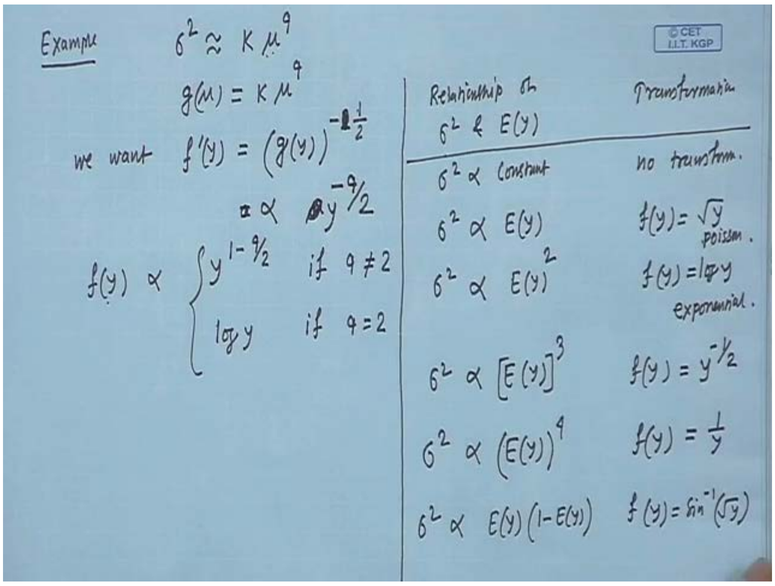
So, Y is response variable and Y has mean mu and variance sigma square and the situation is that this variance sigma square is a function of g mu. So, that means the variance of response variable Y is not constant. It depends on mean and in case of Poisson distribution the variance was equal to mean and in case of Poisson distribution, g is identity function. So, depending on this g we will try to find the transformation on Y, say call it f Y such that the variance of f Y is constant. It does not depend on mu. So, how to find this transformation if one Y so that the variance of f Y is constant?

Well, so you call it U, U equal to f Y. Now, let me talk about Taylor series. The Taylor series of a real or complex function say f x. That is, infinitely differentiable in a neighborhood of a real or complex number say a is f x is equal to f a plus f prime a by 1 factorial x minus a plus f double prime. This is the double derivative at a by 2 factorial into x minus a whole square like this. So, this Taylor series is the polynomial approximation of the function f at a neighborhood of point a. So, here we are looking for a transformation f on Y. You do not know what this function is. So, I mean we do not have the idea about what is this function at this moment.

Let me just write using the Taylor series expression f x is equal to f mu plus a prime mu by 1 factorial into Y minus mu. So, I am considering Taylor series of f Y after the 1st term and this is the neighborhood of mu and I am ignoring the higher order terms. Now, we want to make the variance of U constant. So, the variance of U which is equal to the variance of f Y, this is equal to f prime mu whole square into variance of Y. I hope you understand that variance of this is equal to this quantity because variance of Y minus mu is nothing but variance of Y and this variance of Y is a function of mu. So, this one is equal to f prime mu whole square into g mu. I am replacing the variance of Y which is equal to g mu. Now, if we choose this function f such that f prime mu square is equal to 1 by g mu. Then the variance is equal to 1.

So, if you can choose the function f such that this is true then you are done. So, this is equivalent to f prime mu is equal to g mu to the power of minus 2, right. Then, if you can find the function f such that f prime mu is equal to this, because g mu is given. Then, variance of the transform random variables, we are looking for transformations of Y such that f Y has constant variance then variance of mu which is nothing but variance of f Y is equal to 1.

(Refer Slide Time: 27:17)



Example

$\sigma^2 \approx K\mu^q$

$g(\mu) = K\mu^q$

we want $f'(y) = (g(y))^{-\frac{1}{2}}$

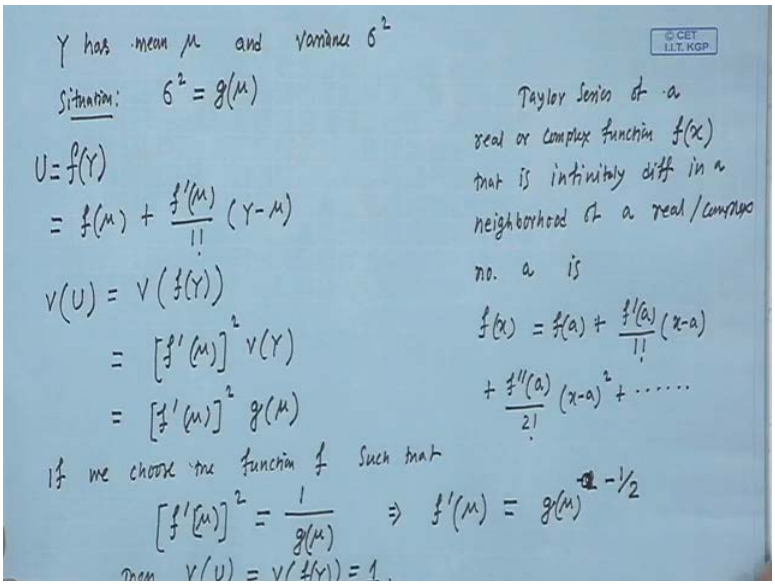
$\propto y^{-q/2}$

$$f(y) \propto \begin{cases} y^{1-q/2} & \text{if } q \neq 2 \\ \log y & \text{if } q = 2 \end{cases}$$

Relationship bn σ^2 & $E(y)$	Transformation
$\sigma^2 \propto$ constant	no transform.
$\sigma^2 \propto E(y)$	$f(y) = \sqrt{y}$ poisson.
$\sigma^2 \propto E(y)^2$	$f(y) = \log y$ exponential.
$\sigma^2 \propto [E(y)]^3$	$f(y) = y^{-1/2}$
$\sigma^2 \propto (E(y))^4$	$f(y) = \frac{1}{y}$
$\sigma^2 \propto E(y)(1-E(y))$	$f(y) = \sin^{-1}(\sqrt{y})$

So, let me give one example to illustrate this idea. Suppose, the variance of y sigma square is approximately or proportional to K mu time q so, what I am trying to say is that g mu, the variance of sigma square. Variance of y which is sigma square is function of mu and that function is equal to K mu to the power of q. And what you want is that, f prime y equal to g y to the power of minus 2. So, you are looking for a function f or transformation on y that is f such that this is true. So, then f prime y is proportional to mu to the power of minus q by 2.

(Refer Slide Time: 29:45)



Y has mean μ and variance σ^2

Situation: $\sigma^2 = g(\mu)$

$$U = f(Y)$$
$$= f(\mu) + \frac{f'(\mu)}{1!}(Y-\mu)$$

$$V(U) = V(f(Y))$$
$$= [f'(\mu)]^2 V(Y)$$
$$= [f'(\mu)]^2 g(\mu)$$

If we choose the function f such that

$$[f'(\mu)]^2 = \frac{1}{g(\mu)} \Rightarrow f'(\mu) = g(\mu)^{\cancel{2}\,-1/2}$$

Then $V(U) = V(f(Y)) = 1$.

Taylor Series of a real or complex function $f(x)$ that is infinitely diff in a neighborhood of a real/complex no. a is

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots\cdots$$

© CET I.I.T. KGP

Let me check, maybe I have made some mistake here. Yeah so, I made a mistake here. So, f prime mu is to the power of minus half. So, here it is to the power of minus half and this quantity. So, we want a transformation f such that this is true and from here oh, sorry this is y. So now, you can check that f y is proportional to y to the power of 1 minus q by 2, if q is not equal to 2. Because, if you take the derivative of this one you get back this one and this is log of y if Q equal to 2. So, here is the transformation.

So, what you see here is that if you see the response variable has variance sigma square which is a function of mu to the power of q, which is the function of mu and the function is mu to the power of q. Then the transformation you have to consider is this one ok. Now, I will talk about several commonly used transformations. Suppose the relationship of sigma square and E y and the transformation here. Suppose the relationship between variance and mean is this: sigma square is proportional to constant. Then you do not need to take any transformation because constant variance assumption is satisfied here.

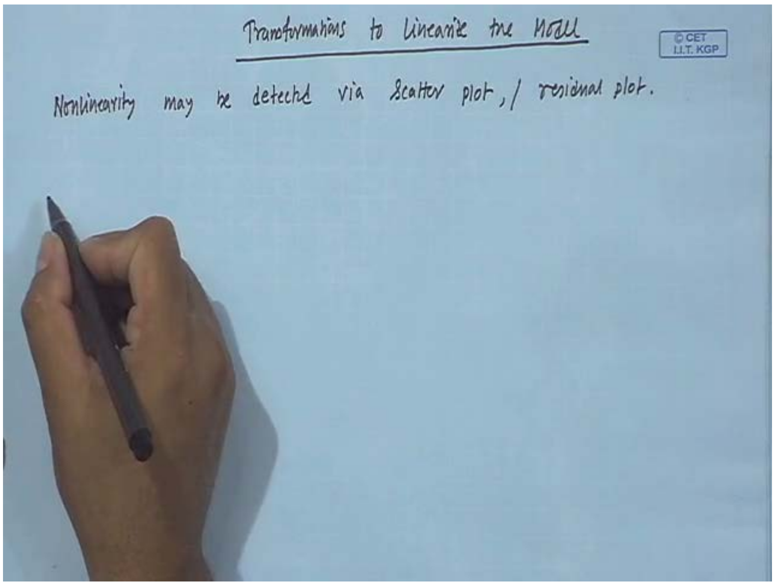
Suppose, sigma square is proportional to expectation of y that is case were y follows Poisson distribution and here you can check that the variance is proportional to mu that means q is equal to 1. So, if you put q equal to 1 here, then the transformation you have to take is f y, f y is equal to y to the power of half. So, that is square root of y. Now, if sigma square is proportional to expectation of y square then q is equal to 2 here.

So, you put q equal to 2 here and then f y is equal to log y and this is the case when y follows exponential distribution. This is the case of y follows Poisson distribution. Now, if sigma square is proportional to expectation of y to the power of 3, then q is equal to 3 and the transformation f y is equal to y to the power of minus half.

Similarly, if sigma square is proportional to expectation of y to the power of 4, then f y, the transformation y to make it constant variance is 1 by y. If the function g mu is of this form you can find the transformation very easily, but suppose if sigma square is proportional to expectation of y into 1 minus expectation of Y. This is the case when y is the proportion between 0 and 1. So, in this case f Y is equal to sin inverse root Y. So, this is all about variance stabilizing transformation. So, what is the basic message from this technique is that if the constant variance assumption is violated then most probably the response variable follows some other distribution not normal distribution like Poisson distribution or it might be the proportion.

And, here if the constant various assumptions are not correct or if the assumption is violated then we learned about the technique of how to transform the response variable to get constant variables.

(Refer Slide Time: 37:09)



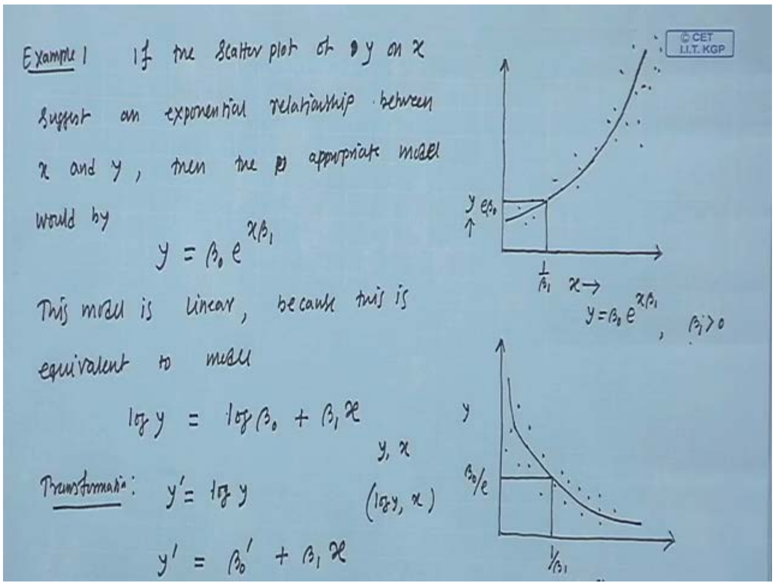
Transformations to Linearize the Model

Nonlinearity may be detected via Scatter plot, / residual plot.

So, next we will talk about transformations to linearize the module. So here, given a set of date, one response variable and one regressor variable or several variables we assume a linear relationship between the response variable and the regressor variable and the assumption of this linear relationship is just a starting point. You know occasionally this assumption might not be correct. So, if the relationship between response variable and the regression variable is not linear, how to detect that?

So, the best technique to detect or you know to get some idea about the relationship between the response variables and the regressor variable is the scatter plot of response variable and regressor variable or one can also go for the residual plot. So, we will talk about several nonlinear relationships here and there are some nonlinear relationship between the regressor variable and response variable which can be linearized easily by using some suitable transformation. So here, if the relationship between the variable is between response variable and regression variable is nonlinear. So, the nonlinearity may be detected via a scatter plot or residual plot.

(Refer Slide Time: 40:14)



Example 1 If the scatter plot of y on x suggest an exponential relationship between x and y, then the appropriate model would by

$$y = \beta_0 e^{x\beta_1}$$

This model is linear, because this is equivalent to model

$$\log y = \log \beta_0 + \beta_1 x$$

y, x

$(\log y, x)$

Transformation: $y' = \log y$

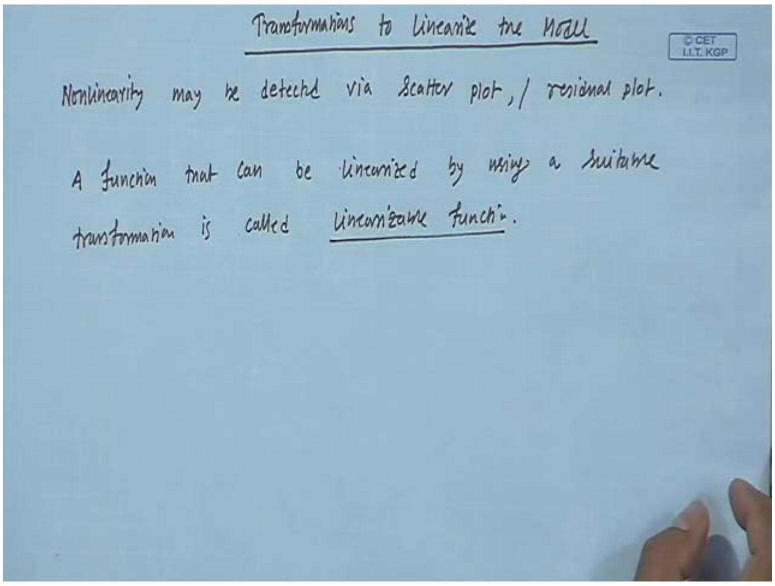
$$y' = \beta_0' + \beta_1 x$$

Let me give one example. If the scatter plot of y on x suggest an exponential relationship between x and y then the appropriate model would be y equal to beta naught e to the power of x beta 1. So, what it says is that a given set of say regressor variable and the response variable you first find the scatter plot and if you see the scatter plot indicates that the relationship between y and x is exponential. Then the appropriate model for this one is y equal to beta naught e to the power of x beta 1. And, this is the scatter plot. This is the exponential relationship between y and x when beta 1 is greater than 0 and it could be like this also.

This also suggests exponential relationship between response variable and the regression variable, but here beta 1 is a negative. And, if the relationship between y and x is exponential then this model is in fact, linear. This model is linear because this is equivalent to the model log y equal to log beta naught plus beta 1 x. So, here the transformation you are taking is y prime equal to log y. That is all. So, the model, the final model is y prime equal to beta naught prime plus beta 1 x.

So, even if you see that the relationship is exponential between the response variable regression variable then the appropriate model is this one. And then you can transform this model to a linear model. So, here given y x you transform y to log y and x and then you fit a linear model using this transform data.

(Refer Slide Time: 44:59)



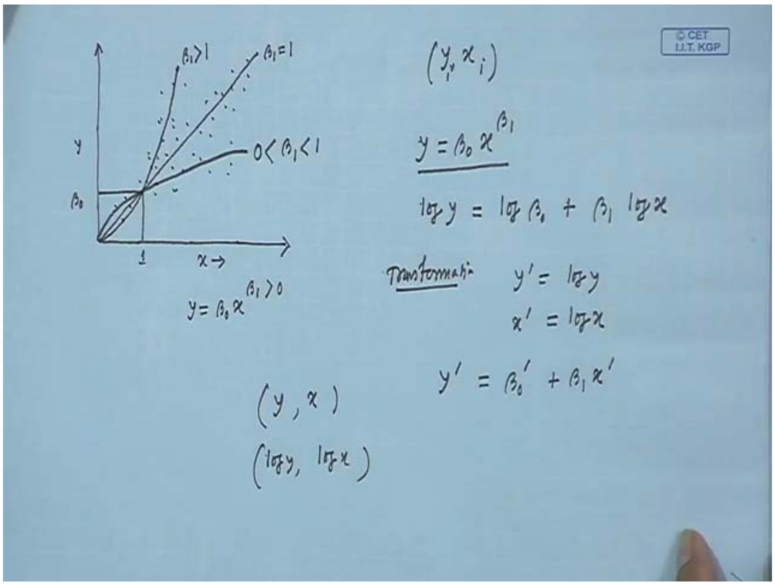
Transformations to Linearize the Model

Nonlinearity may be detected via Scatter plot, / residual plot.

A function that can be linearized by using a suitable transformation is called linearizable function.

So, a function that can be linearized by using a suitable transformation is called linearizable transformation function, is called linearizable function. So, there are some functions which can be a linearized by using the suitable transformation very easily those are called a linearizable functions.

(Refer Slide Time: 46:06)



(y_i, x_i)

$y = \beta_0 x^{\beta_1}$

$\log y = \log \beta_0 + \beta_1 \log x$

Transformation $y' = \log y$

$x' = \log x$

$y' = \beta_0' + \beta_1 x'$

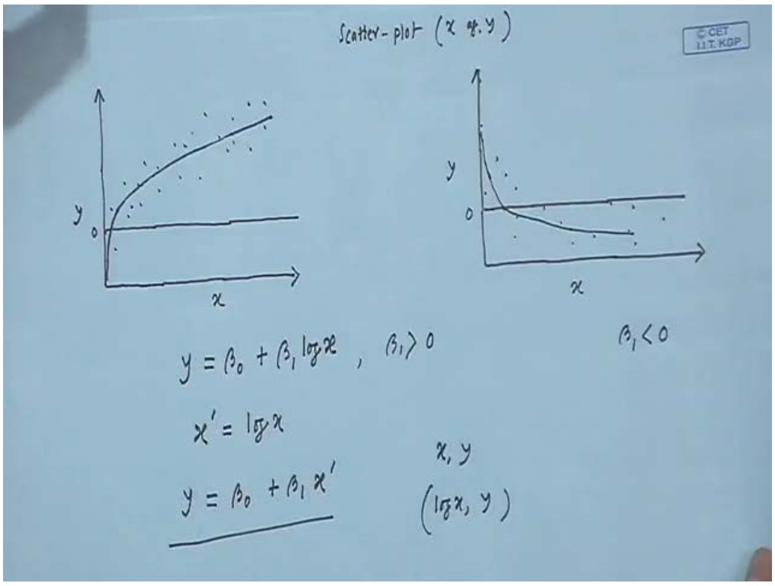
(y, x)

$(\log y, \log x)$

Next, let me talk about some more example. So, you are given response variable and the regressor variable you fit the scatter plot. So, if you see the scatter plot it is centered about the line. So, this is the line y equal to x and then you go for linear fitting. But, if you see the scatter plot is centered about this curve or may be the scatter plot is centered about this curve and then the relationship between response variable and regressor variable is some sort of polynomial, it is not a linear relationship. The relationship is y equal to beta naught x to the power of beta 1. So, this one is the case when beta 1 is greater than 1, this is the case when beta 1 is equal to 1 and this is the case when beta 1 is less than 1 but greater than 0.

So, as I told before there are some nonlinear relationships which can be easily transformed into linear form. Those are called linearisable function. So, this one is also a linearisable transform function because you can easily make it linear by taking log function, log Y is equal to log beta naught plus beta 1 log x. So, the transformation here you are choosing is: the transformation are y prime is equal to log y and x prime is equal to log x and the final model is y prime is equal to beta naught prime plus beta 1 x prime. So, you transform the given data y, x to log y log x and if you plot the scatter plot for this transform data perhaps, you will get the scatter plot a centered about the straight line and you can go for a linear fit.

(Refer Slide Time: 49:38)



$y = \beta_0 + \beta_1 \log x$, $\beta_1 > 0$ $\qquad \beta_1 < 0$

$x' = \log x$

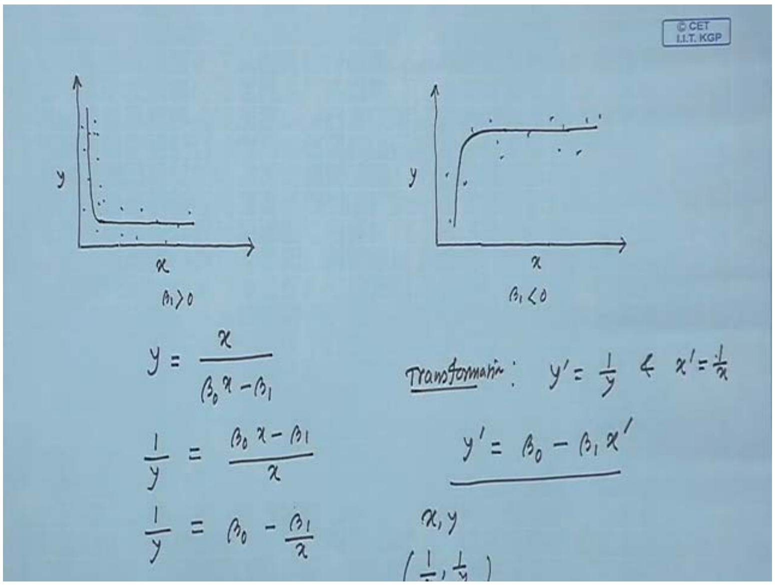
$y = \beta_0 + \beta_1 x'$

x, y

$(\log x, y)$

So, one more example is here. So, this is the scatter plot of y against x and if you see the scatter plot similar to this, which is centered about this curve. Then the relationship between y and x here is y equal to beta naught plus beta 1 log x, for beta 1 greater than 0 and this one is for beta 1 less than 0. And it is very easy to realize that this is a linearizable function because here you just take the transformation x prime equal to log x and then the model become y equal to beta naught plus beta 1 x prime. So, given the data x, y you transform that to log x, y and fit this linear model. I give one more example and then stop here.

(Refer Slide Time: 51:15)



$\beta_1 > 0$ $\qquad \beta_1 < 0$

$$y = \frac{x}{\beta_0 x - \beta_1}$$

Transformation: $y' = \frac{1}{y}$ & $x' = \frac{1}{x}$

$$\frac{1}{y} = \frac{\beta_0 x - \beta_1}{x}$$

$$y' = \beta_0 - \beta_1 x'$$

$$\frac{1}{y} = \beta_0 - \frac{\beta_1}{x}$$

x, y

$\left(\frac{1}{x}, \frac{1}{y}\right)$

Suppose, you scatter plot is centered about this curve or this curve ok. Then the relationship between y and x is again a linearizable function. Here, y is equal to x by beta naught x minus beta 1 and this one is for beta 1 greater than 0 and this one is for beta 1 less than 0. And you can transform this to linear function. What is 1 by y? 1 by y is equal to beta naught x minus beta 1 by x and so here, 1 by y is equal to beta naught minus beta 1 by x. So, what transformation you are taking here is that, transformation is y prime equal to 1 by y and x prime is equal to 1 by x and the final model is y prime equal to beta naught minus beta 1 x prime.

So, what you have to do is that given the data x, y, if you see the scatter plot is similar to this one. Then you take the transform data 1 by x, 1 by y and fit a straight model. So, this is what we want to mean by a linearizable function. So, if you see the relationship between the response variables and the regressor variable is not linear and if it is similar of one of this thing you can take some easy transformation on the variable. And, then the problem is equivalent to fitting a linear model between the response variable and the regressor variable.

So, thank you for your attention.