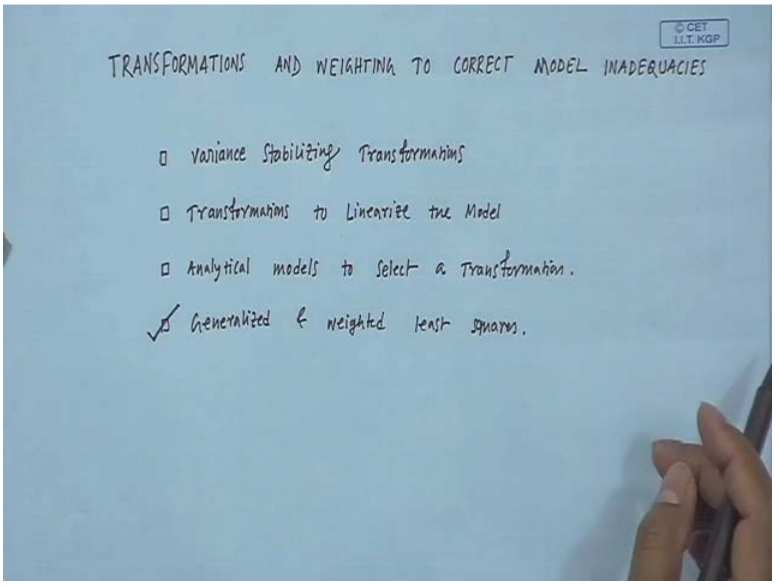
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institution of Technology, Kharagpur

Lecture - 23
Transformations and weighting to correct model inadequacies (Contd.)

Hi, so this is my 3rd lecture, on transformation and weighting to correct model inadequacy.

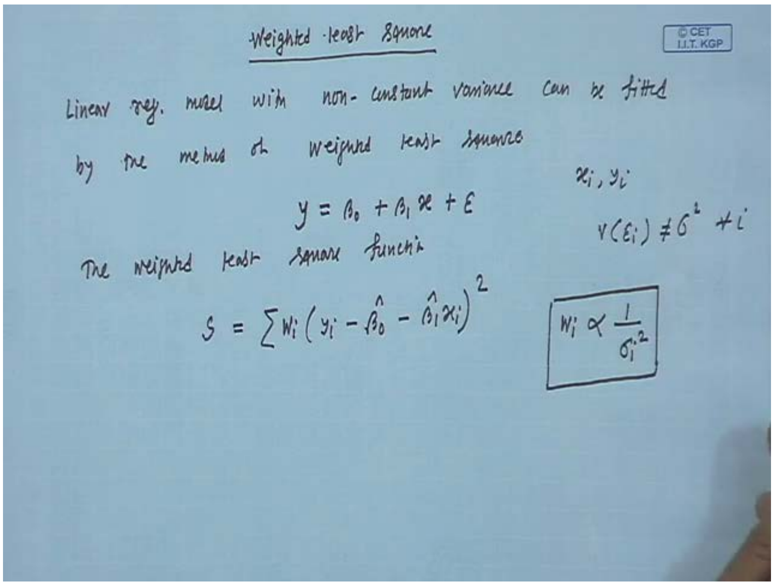
(Refer Slide Time: 00:32)



And here is the content of this topic, we already talked about, variance stabilizing transformation and transformation to linearize the model and also it talked about generalized and weighted least square. So, today I want to give an example to illustrate you know, this weighted least square technique, we talked in the previous class and also I am going to talk about this, analytical methods to select atoms formation. Ok, so, let me repeat once more that, in simple linear regression model, or in the multiple linear regression model. We make several assumption; on the error terms and given set of observations, say x i and y i. You do not know whether, your data set satisfy those assumptions or not.

So, you have started you have learn several techniques to check, whether your data set satisfied the model assumption or not, in module called module adequacy checking, most specifically you know this residual plot, that is residual against a fitted response is an effective technique to test whether, your data set satisfy the model assumptions or not. Now, in this module what we are doing is that, you know we are supposed, your data set does not satisfy the module assumption, then how to correct the module inadequacies. So, we have learned about two techniques, like one is called variance stabilizing transformation and also we learned generalized and weighted least squares to correct model inadequacy.

(Refer Slide Time: 03:17)

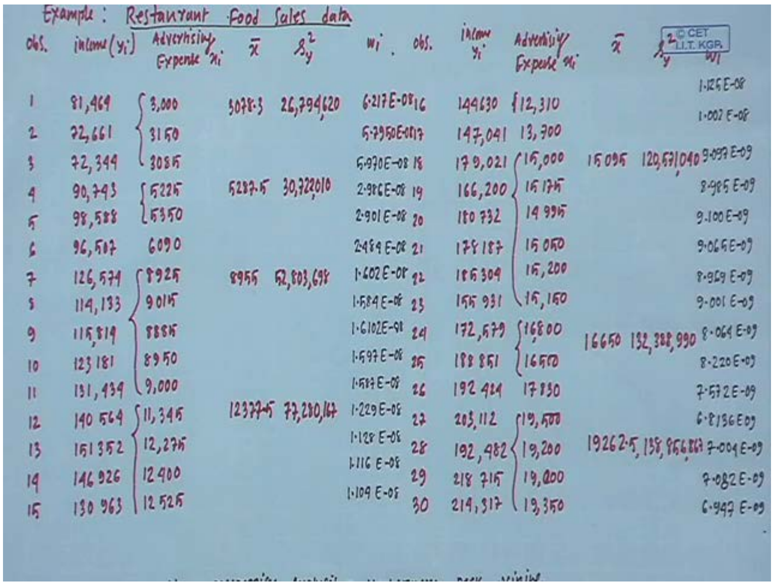


So, today what I will do is that now, I will sort of repeat, weighted least square technique and then I will give an example to illustrate, weighted least square technique. So, weighted least square, so, linear regression model, with non constant variance, can be fitted, by the method of weighted least square. So, this is the particular case of generalized, least square technique and here, the variance is are not equal, but, the observations are sort of uncorrelated, that means the coherence terms are equal to 0. So, suppose you are your given the data sets x i, y i and your trying to fit a model between this two variables, between x and y, a simple linear regression model.

So, y equal to Beta naught, plus Beta 1, x plus epsilon and you know here that variance of epsilon i, is not equal to sigma square for all i, there is no constant variance. So, what we do here is that, we consider the weighted least square function. The weighted least square function is equal to S, which is y i minus, y i had the basically so, that is Beta naught hat minus Beta 1 hat x i. This is the function we minimized to estimate Beta naught and Beta 1, in simple linear in ordinary least square. But, here what will do is that we give weighted is W i to the ith observation.

And what we started in the previous lecture is that, this W i is proportional to 1 by sigma i square, ok. And I already explained this part in the previous class, why this weight is proportional to 1 by sigma i square. So, today what I am going to do is that the main problem you see here, you are the you are just given the observation x i and y i. So, you do not know what are, what is this sigma i square for the i-th observation. So, I will illustrate, how to estimate this sigma i square for a given set of observation, x i and y I, ok.

(Refer Slide Time: 07:03)



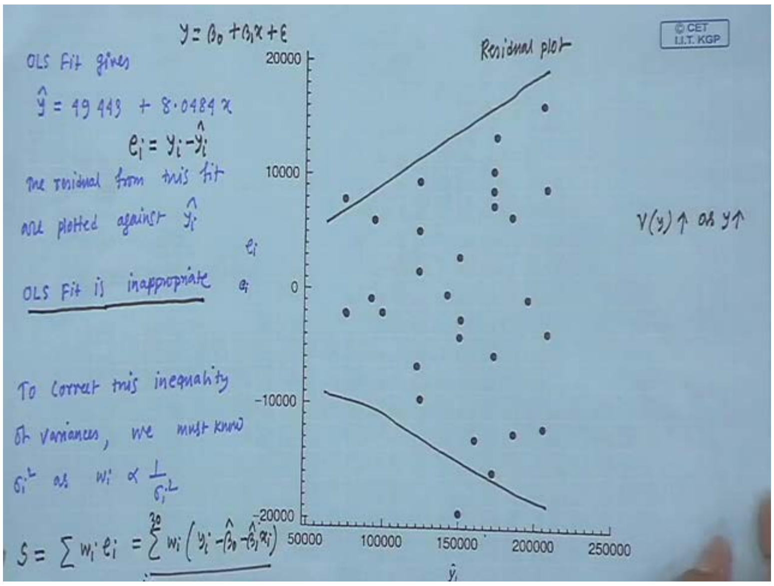
Example : Restaurant Food Sales data

obs.	income (y_i)	Advertising Expense x_i	$\bar{x}$	s_y^2	w_i
1	81,464	3,000	3078.3	26,794,620	6.217E-08
2	72,661	3150			5.7950E-08
3	72,344	3085			5.970E-08
4	90,743	5225	5287.5	30,722,010	2.986E-08
5	98,588	5350			2.901E-08
6	96,507	6090			2.484E-08
7	126,574	8925	8955	52,803,698	1.602E-08
8	114,133	9015			1.584E-08
9	115,814	8885			1.6102E-08
10	123 181	8950			1.597E-08
11	131,434	9,000			1.587E-08
12	140 564	11,345	12377.5	77,280,167	1.229E-08
13	151 352	12,275			1.128E-08
14	146 926	12 400			1.116E-08
15	130 963	12 525			1.109E-08

obs.	income y_i	Advertising Expense x_i	$\bar{x}$	s_y^2	w_i
16	144630	12,310			1.125E-08
17	147,041	13,700			1.002E-08
18	179,021	15,000	15095	120,571,040	9.097E-09
19	166,200	15 175			8.985E-09
20	180 732	14 995			9.100E-09
21	178 187	15 050			9.065E-09
22	185 304	15,200			8.959E-09
23	155 931	15,150			9.001E-09
24	172,579	16800	16650	132,388,990	8.064E-09
25	188 851	16500			8.220E-09
26	192 424	17830			7.572E-09
27	203,112	19,500			6.8136E-09
28	192,482	19,200	19262.5	138,856,867	7.004E-09
29	218 715	19,000			7.082E-09
30	214,317	19,350			6.947E-09

So, let me take an example, of this restaurant food sales data. So, here we have observation 30 observations here and this is the response variable y, this is income on food sell, per month and this is the advertising expense x I, for the whole year and so, here is your response variable y I, which is stands for the income per month and the regressal variable x i, which is cost on advertising per year. And we are trying to sort of find a relationship between, these two variables x i and y i and so, first what we do is that, well your given x i and y i for i equal to 1 to 30.

(Refer Slide Time: 08:32)

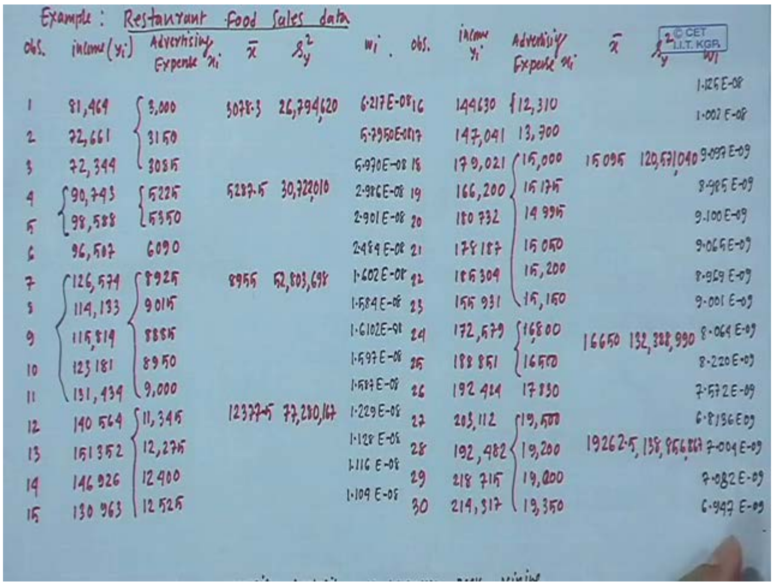


And why do not you just first fit a simple linear regression model, say using the ordinary least square technique. You fit a model, y equal to Beta naught, plus Beta 1 x, plus epsilon. So, you know how to estimate this parameter Beta naught and Beta 1, that you have learned in the 1st module, call simple linear regression model. So, here is the fitted model and then once you have the fitted model, you compute the residuals. So, residual means, you compute e i, e i is equal to y i minus y i hat. So, this is the original observation, or response value and this is the fitted response value.

And, the new plot, this is what I call I know, this is call the residual plot, this is call residual plot, residual plot. So, here residual is plotted against the fitted response of y i hat and look at this plot here, it looks very similar to outward, open final to me, right. So, that means here, the constant variance assumption is highlighted. So, what happen here is that, here variance of y increases, or sigma square increases as, y increases. So, the constant variance assumption is highlighted here. So, this implies that the ordinary least square is fit is in appropriate here. So, you cannot go for of course, the ordinary least square fit is the starting point and then you have realized that, from the residual plot that ordinary least square fit is inappropriate.

So, now to correct this inequality of variances, we will go for weighted least square technique and for weighted least square technique, to use that we need to know sigma i square, because, there in the weighted least square technique in minimize this quantity S, which is W i, e i. That means, w i y i minus Beta naught hat, minus Beta 1 hat x i. We minimize this quantity. So, we need to know, the weight w i for the i-th observation from i equal to 1 to 30 here, ok. Now, I will talk about given a set of observation x i and y I, how to how to estimate, sigma i square, for the i-th observation. So, sigma i square is the population variance, from where the i-th observation is coming.

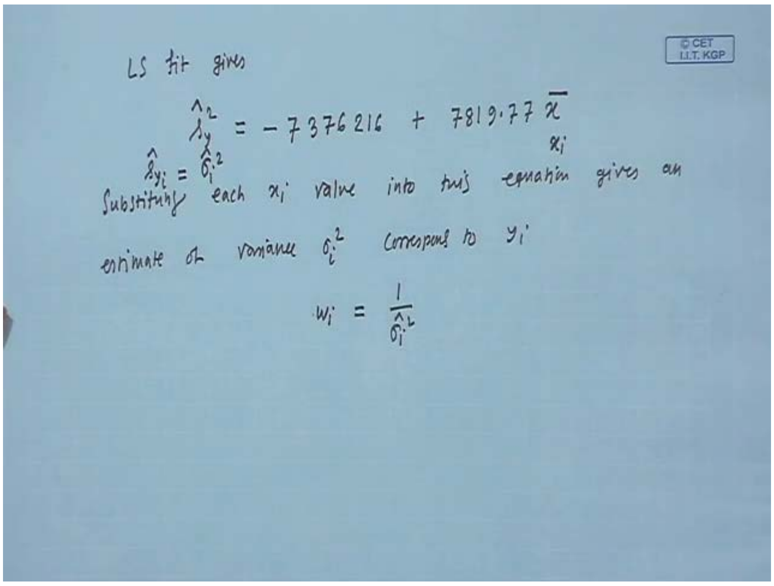
(Refer Slide Time: 12:33)



Well, so, here is the observation again; this is the income and the cost on advertising. Now, look at this data here. You can see that, this 3 x values there sort of near equal. So, what we do is that, will put them in one cluster. So, this is one cluster, these two values are again near equal. So, we put them in one cluster. Here you can see this five observations, or five regressal values there near equal, will put them in one cluster, so on. Now, here you compute the average of this cluster, x bar and the idea here is that you know, this 3 points are enough near equal to consider them as a single point. So, corresponds to a single response, corresponds to a single radix.

What is call what you call x value? single x value. We have three responses and so, these are the three response values, in this cluster and what we do is that, will compute the sample variance of this response values. So, here is the sample variance. I am show that you know what is sample variance, correspond to this cluster here is the sample variance correspond to, this is sample variance correspond to these two observation and here is the sample variance. Correspond to these five observations, ok. So, on you compute the sample variances. Now, if you look at the x value, x bar value and the sample variance, you can see that the sample variance of the response variable y that increases, as x increases and if you sort of plot scatter plot between x bar and the sample variance. You will see that there is a that scatter plot will indicate sort of linear relationship between at these two, ok.

(Refer Slide Time: 16:03)

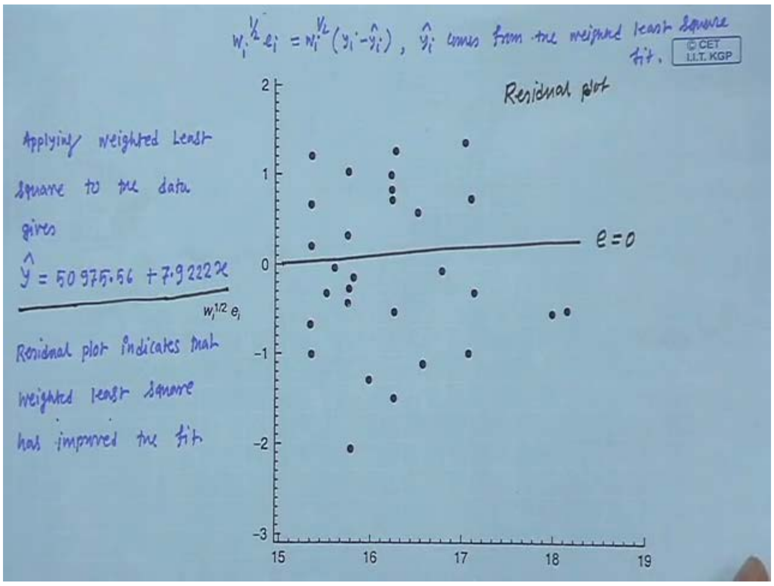


So, what we do is that, then, we fit a linear relationship between x and the sample variance. So, least square fit gives, s y square equal to, this is the estimated value, equal to minus 7376216 plus 7819.77 x bar, ok. So, what we are doing is that, we try to find the regressal variable, or x values. Which are near equal and then we consider them as a single a point, correspond to the regressal variable and then correspond to that single point. You have several response values and you compute the sample variance of those response values.

So, you can think of like, that particular observation example, (Refer Slide Time: 12:33) this one. Now, you can think of that, this observation, if this is my advertising cost. Then the response for the income, whatever I get that is coming from a population. Which has sample, which has variance this much: ok and that we do for all the clusters. And then, we see the relationship between the sample mean for the, for advertising cost and the sample variance for the response variable and we fit a linear relationship between these two. Why we do that?

Now, what you can do is that, you can substitute, each x i value, into this equation, gives an estimate of variance sigma square correspond to y i. Now, here you put x I, you have 30 x i values and then, you will get some estimate of the variance. Correspond to that x i, that means you will get here, the estimate of sigma. So, sigma i square hat, that is you know basically, s y i hat ok. So, this square once you have, sigma i square, for all i equal to I mean estimate of sigma i square for i equal to 1 to 30. You can compute w I, which is equal to 1 by sigma square. (Refer Slide Time: 12:33) So, here you can see that, you compute the sigma i square, corresponds to I mean estimate sigma i square correspond to this point and then you take 1 by sigma a square will get the weights here. So, the weight has given here.

(Refer Slide Time: 20:28)

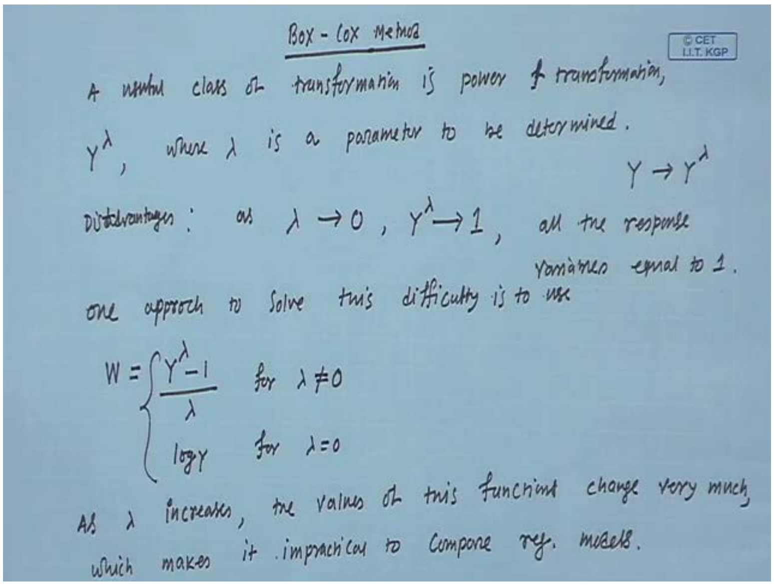


And then so, your now in a position to apply weighted least square technique. So, you apply weighted least square technique, because, you know the weights. All the weights w i from i equal to 1 to 30 and here is the model, obtain using weighted least square technique, ok. And now, to see whether this fit has any improvement, what the previous one. Again what you do is that, you have the fitted equation and then you compute the residual and you know the fitted response. So, you again what you do is that, you again plot draw the residual plot.

So, this is plot of e i against y i hat and here instead of just simple residual, against the fitted response. You take the weighted residual and here also, you multiple that by the weight. So, here is this scat, here is the residual plot and this is the line, e equal to 0 and this sort of the residual plot indicates that, weighted least square has improves the fit. Because, before the weighted least square it was the sort of outward open panel, but, now it is, it has improve I mean here the residual are, all most like you know centered about the line e equal to 0. So, this is how we applied weighted least square technique to, whenever given a set of data.

So, given a set of data so, the final massage so, given a set of data. You check whether, that satisfy the model assumptions. If it is not, then if you are willing to apply, weighted least square technique, you find out sigma i square. Because, the weight w i is proportional to 1 by sigma i square and we talked about how to find or how to estimate sigma i square just now. And then you fit a linear regression model using the weighted least square technique and finally, after fitting the after once you have the fitted model using the weighted least square technique. You again draw the residual plots and see the improvement ok. So, well this is what about the generalized least square and weighted least square technique.

(Refer Slide Time: 24:15)

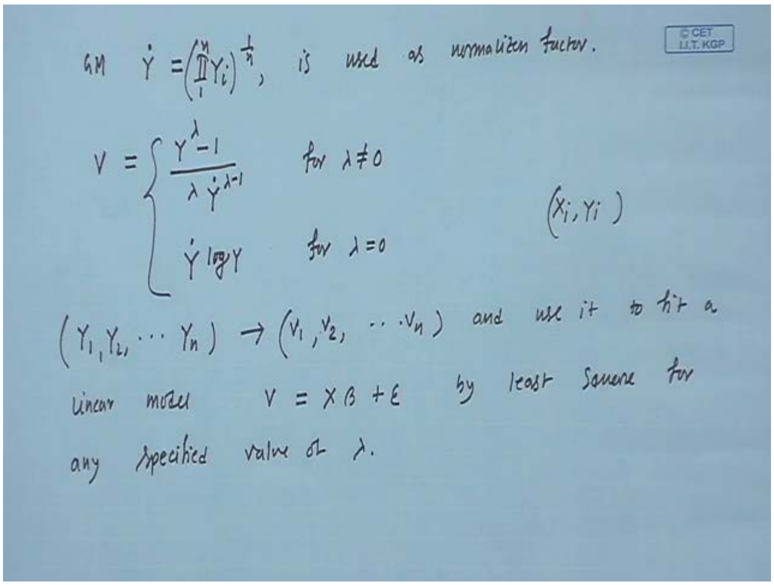


And finally, will talk about one more technique, which is again to collect the model inadequacy. So, this technique is called Box Cox method. So, this is called Box Cox method and this one is a technique to correct the model inadequacy, by transforming the response variable ok and it says that. So, this one is in I mean this is, this one correct the model in inadequacy by transforming the response variable y to something else, ok and it says that useful class of transformation is power transformation, is power transformation. That is, you transform Y to, Y to the power of lambda, where, lambda is a parameter to be determined, ok.

So, what you are doing is that. S o, you try to correct the model inadequacy by transforming Y to Y lambda and lambda is a parameter and which lambda whether it is 2, 2.5 something, or minus 2 ok. Now, the problem with this power transformation or this particular power term transformation is that. So, the disadvantages is that as lambda approaches 0 Y to the power of lambda, approaches 1 right. So, that means, this is sort of meaningless, because, here all the response variable equal to 1. Irrespective of what is the value of regressal variable, all that response variable is equal to 1. So, this is and disadvantage of this particular power transformation, I mean transforming Y to Y lambda.

So, it says that, the method says that, one approach to solve this difficulty is to use this transformation. Instead of Y transforming to Y lambda, you take this transformation W, which is equal to Y to the power of lambda, minus 1, by lambda for lambda equal lambda not equal to 0. And as you know that this function, tense to log Y, as lambda tense to 0, for lambda equal to 0. So, this solve the problem of all the response variable transforming to 1, as lambda tense to 1 as lambda tense to 0. But, the problem I mean problem with this one, or even this one is that, as lambda increases, the values of this functions change very much. Which makes it in practical: to compare regression models. As you see, you know of course, if lambda is large then W value will be very large and this makes it and in practical to compare the regression model.

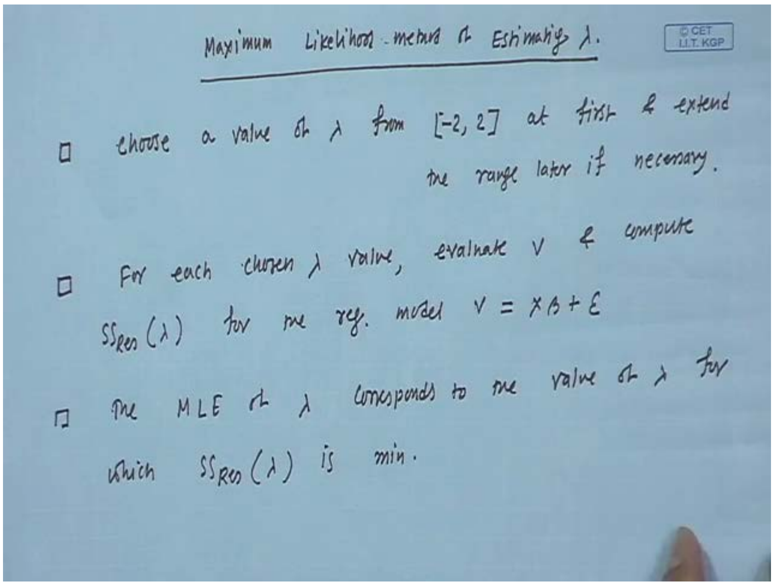
(Refer Slide Time: 30:47)



So, we need to use some normalization factor here. So, the geometric mean of the response variable, which we denoted by Y dot, which is equal to Y i, n is used as normalization factor. So, what we do is that, we transform Y to V, where V is equal to Y to the power of lambda minus 1, by lambda into Y dot lambda minus 1. For lambda not equal to 0 and transform Y to Y dot log y. This is base e for lambda equal to 0. So, this is the transformations suggested by Box Cox method and the question is, now how to get this lambda value? So, given set of observation X I, Y i. What you are doing is that? a transforming all the response variable Y 1, Y 2, Y n to V 1, V 2, V n and use it to fit a linear model, between V and x. V equal to x Beta plus epsilon, by least square for any specified value of lambda.

So, what Box Cox method does is that, you know it suggest some transformation from, some power transformation of course. From thorough response variable Y. So, you transform Y to V, for all i. I mean for transform Y i to V i for i equal to 1 to n. And then you fit a linear regression between the transform variable, V and the regressal variable X, by using the ordinary least techniques, for the specified value of lambda. Because, see still I did not talked about how to decide, how to fix the value of, how to decide the determine the value of lambda.

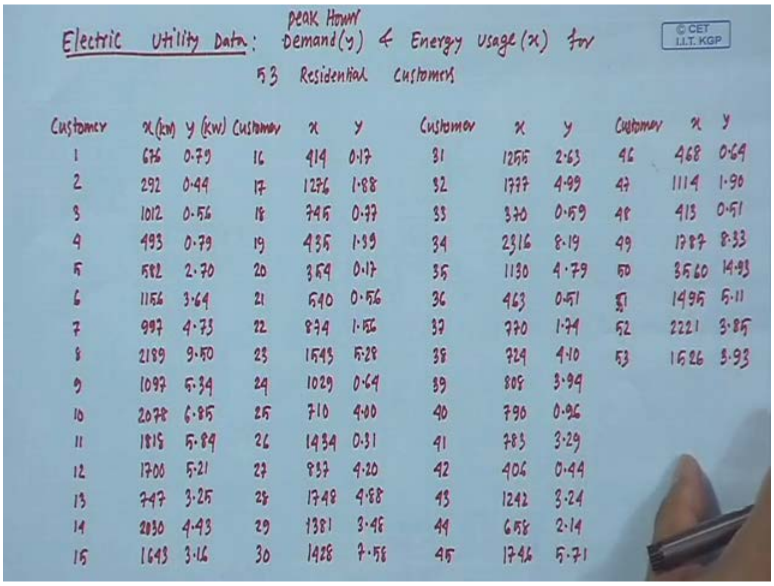
(Refer Slide Time: 34:29)



So, here is the method to determine the value of, to estimate the value of lambda. This is called maximum likelihood method of estimating lambda ok. So, what are the steps you know? It is suggest there, you know choose, a value of lambda from this interval, minus 2, 2 it is a closed interval. At first and extend the range later if necessary, and then for each chosen lambda value, evaluate V and compute, S S residual correspond to that lambda. For the regression model, V equal to X Beta plus epsilon. For chosen value of lambda you fit this model, between X and Beta and then between X and V and then you compute the s s residual. You know what is this S S residual.

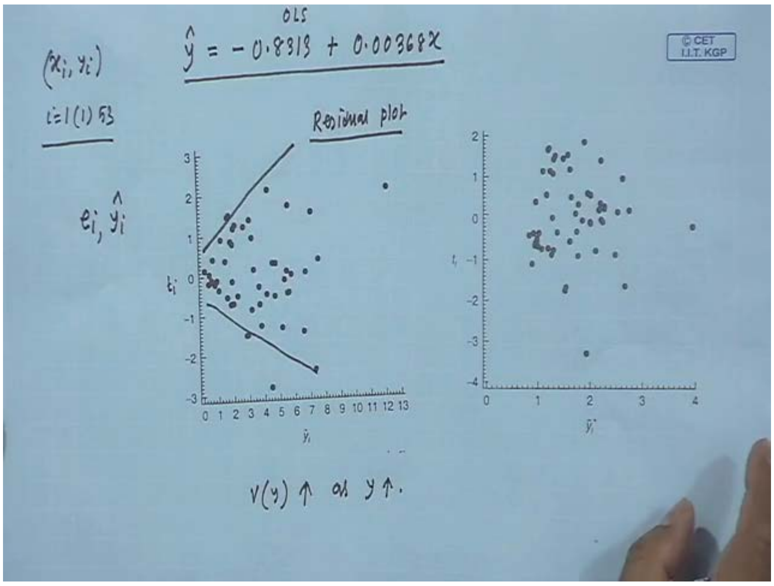
And then it says that, the maximum likelihood estimator of estimating lambda, corresponds to the value of lambda. For which S S, residual lambda is minimum. So, this is you know of course, you compute you take different value of lambda between this, in this interval you compute, you fit the model between Y between V and X simple linear regression model, or of course, multiple linear regression model, if number of regresses is more than 1. And then you compute S S residual, for each lambda and you then you see for which lambda, this S S residual is minimum and that value of lambda is the maximum likelihood estimator of lambda ok.

(Refer Slide Time: 37:58)



Now, I given example, to illustrate this Box Cox method. So, here, this is called you know electric utility data. So, I have 53 observations total and this y the response variable is, stands for the peak our demand and the regressal variable x stands for the energy usage per month. So, this is for a foe first family, this is the energy usage in the particular month and hers is the pick our demand. And what you will interested to do is that, you are interested to do find a relationship between, the monthly usage and the pick our demand. So, here we have only one regression variable and the response variable and of course, what you will do is that. Will fit a, will start with a simple linear regression model using ordinary least square technique.

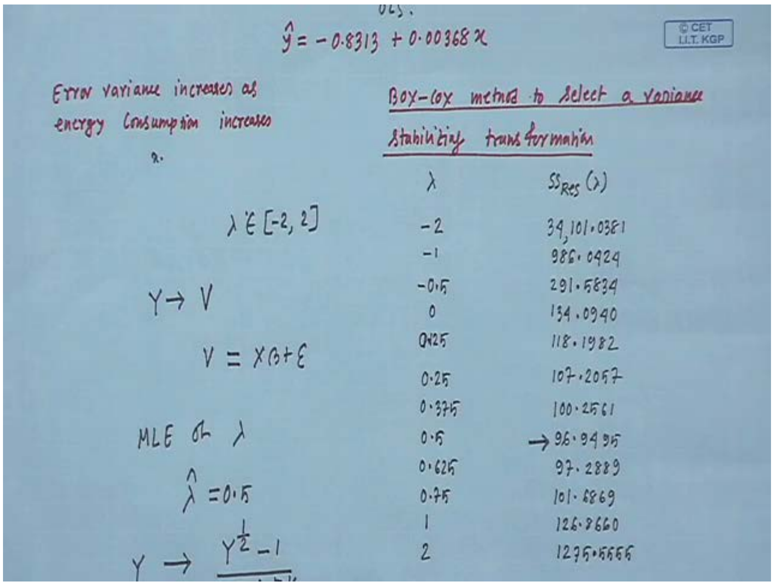
(Refer Slide Time: 39:31)



And here is the fitted model between y and x, y hat is equal to minus 0.8313 plus 0.00368 x. So, this is the fitted model between, or the relationship linear relationship between x and y. And of course, you know we need to check whether, this fit is good, or in other sense, I mean whether this data set here given a data set x i y i, for i equal to 1 to 53 I believe. 1 to 53 and you need to check whether this data set satisfied the basic assumption or not. For that what we do is that, you fit the simple linear regression model, does not matter whether this data set satisfied the basic assumptions or not. You fit a linear regression model using ordinary least square.

And then you compute the residuals, you compute the residual and you have the fitted response, you plot them. So, this is what is called the residual plot and this residual plot will suggest, or will say whether the data set satisfy the basic assumption or not. So, here instead of I think instead of e I, we have standardized, we have use standardized residual does not matter. So, if you see the residual plot here. So, t i is plotted against estimated response. So, you can see here that again, this residual plot is sort of outward open panel. That means, this residual plot indicates that, ordinary least square fit is not appropriate and because, of the fact that here, the variance of y increases, or sigma square increases as y increases, ok. So, this data sort of highlight the constant variance assumptions. So, we cannot continue with the ordinary least square fit, here what we I will do is that? will try to apply the Box Cox technique here.

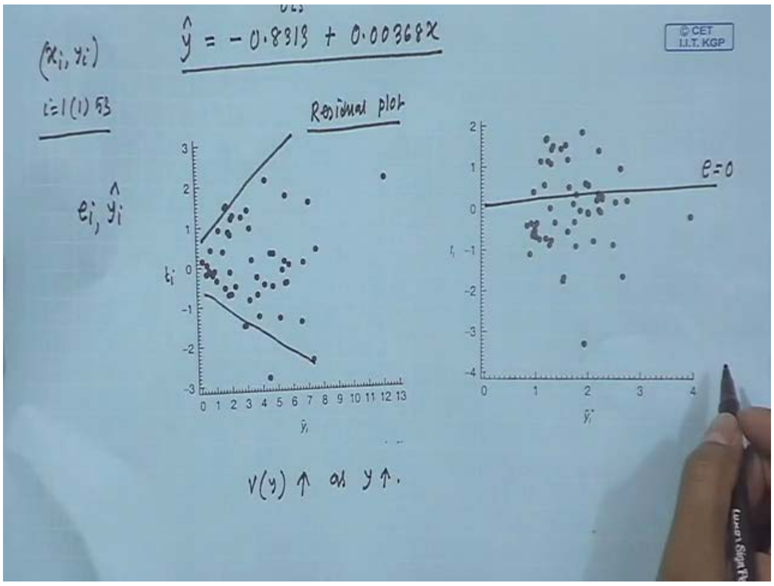
(Refer Slide Time: 42:41)



So, the residual plots suggest that the error variance increases at as you know, energy consumption increases. That means, I think the this is the x value y. So, this is the ordinary least square fit and now, what we do is that will take a lambda, from this interval minus 2, 2 plus 2. So, first we start with minus 2 and you compute V, you know, what is that V for lambda equal to 2. So, you transform here Y to V and then you fit a model between V and X, V equal to X Beta plus epsilon, ok and here is the residual value.

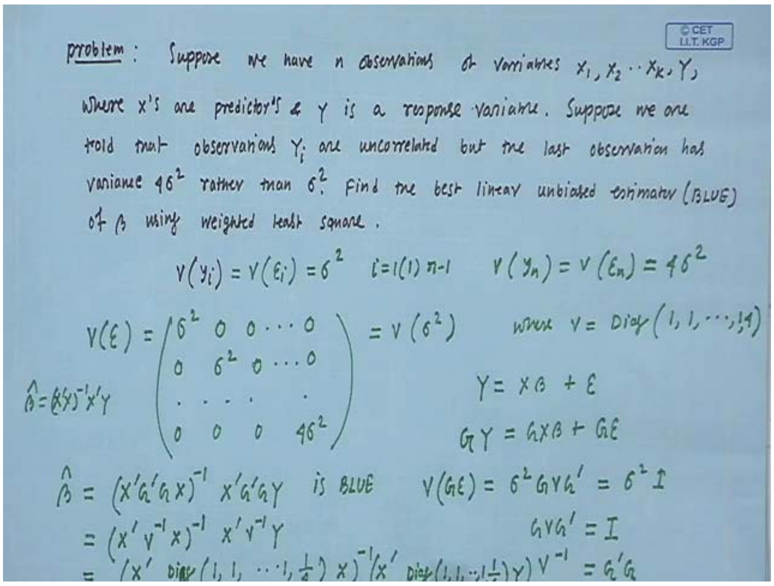
So, you do it for different lambda value and compute the corresponding S S residual. So, here you can see that, S S residual is minimum for lambda equal to 0.5, ok. So, this is called the maximum likelihood estimate of lambda. So, maximum likelihood estimate of lambda is; lambda hat is equal to 0.5. So, the transformation finally, we go for is this one. You transform Y to, Y to the power of half that is Y to the power of lambda, minus 1 and lambda Y to the power of lambda minus 1. So, lambda minus 1 is again half year. So, this is the final transformation, this is the geometric mean of the response variable. So, you transform Y to this. So, this is the suggestion from the Box Cox technique and if you check the residual plot for this transform data, for lambda equal to half.

(Refer Slide Time: 45:45)



Now, you can see the line e equal to 0. So, the transform fit here, I mean has improved because, this residual here, the standardized residual are all most central about the line i equal to 0, ok.

(Refer Slide Time: 46:12)



So, this is what the Box Cox method is and now, we have some time. So, what we do is that, will talk about some problem. Ok see the problem here, suppose, we have n observations of variables X 1, X 2, X k and Y. Where X's are of course: regressor of weighted variable and Y is response variable. You have n observation for this variables, suppose, you have told that, observation Y i are uncorrected. But, the last observation has variance, 4 sigma square rather than sigma square. When the problem is that find the best linear unbiased estimated blue, of Beta using weight least square. Because, you know this sort of fit the weighted least square assumption, because, weighted least square is the particular case of generalized least square and in weight least square we assume that the observations are uncorrelated. But, the variances are non constant, I means, they are unequal.

So, what the data we have here is that variance of Y I, is equal to, which is equal to of course, variance of epsilon i. Which is equal to sigma square, for i equal to 1 to n minus 1 and variance of Y n, which is n as variance of epsilon n is given to be 4 sigma square well. So, the what is the variance co efferent of matrix here, variance of epsilon is equal to n sigma square and the data are uncorrelated. So, the variance terms all equal to 0, 0 sigma square 0 0 0. And then finally, here it is, 4 sigma square, ok. So, this I can write as V sigma square, where V is diagonal 1, 1, 1 and then finally, it is 4 ok.

So, what you have to do you have to find, best linear unbiased estimated of Beta. That is the regression co efficient. So, if I forget the formula for the estimated best linear unbiased estimated of Beta. You can derive it of course, let us start with the model Y equal to X Beta, plus epsilon and this variance of epsilon is not of the form sigma square i. So, what you do is that, in the generalized this technique, what we do is that we take a transformation of this model. You multiple by a matrix called G, G Y is equal to G X Beta, plus G epsilon. So, G Y are the transform data now and we need to choose a correct G right.

So, and also we need to make this variance of G epsilon, which is equal to sigma square G V G prime, I hope you understand this one. I want this to be sigma square, i which is equivalent to G V G prime is equal to I, right; which is same as V inverse equal to G prime G, ok. So, V inverse is equals to we need to choose G such that V inverse is equal to G prime G and we know that, Beta hat is equal to X prime, G prime, G X inverse, X prime, G prime, G Y. I hope you understand this, because, in the simple linear regression model it is, Beta hat equal to just X prime X, inverse X prime Y.

So, what I am doing is that I am replacing X by G X and Y by G Y. Then you get this formula. So, this 1 is equal to this is no this is known as this is blue, ok. So, what I have to do here is that, this G is nothing but, V inverse. So, the blue is, final blue is X prime, V inverse X, inverse X prime, V inverse Y and I know my V, this is my V and then no I know what is V inverse. So, the final best linear unbiased estimator of Beta is, X prime, diagonal of inverse of this. So, that will again diagonal matrix and it is 1 1 1 by 1 4 inverse of this elements. This is my V inverse, X whole inverse and then X prime again the inverse that is diagonal 1, 1, 1, 1 by 4 Y, right.

So, this is what the blue of, best liner unbraided estimator, of Beta hat. When you have this sort of restriction for the variance of Y i ok. So, that is all for today and we have to stop now. So, just let me conclude the whole module once more. So, we know that the simple linear regression model, or in the multiple regression model. There are some basic assumptions. Now, given a set of data you would not know, whether your set satisfy those basic assumptions or not. So, what you do is that, you start with the simple you fit a simple linear regression model, using the ordinary least square technique.

And then you compute the residuals, you go for the residual plot, which is plot between residual and the fitted response. From the residual plot we say, whether your observations satisfied the basic assumptions or not. Well, so, if your data does not satisfy the basic assumption, then what we have learned in this module is that. How to correct those module inadequacy, using different techniques like, variance stabilizing transformation, weighted least square technique. Generalized least square techniques and also finally, we talked about, regarding transformation of response variable by using Box Cox method. So, that is all for today.

Thank you.