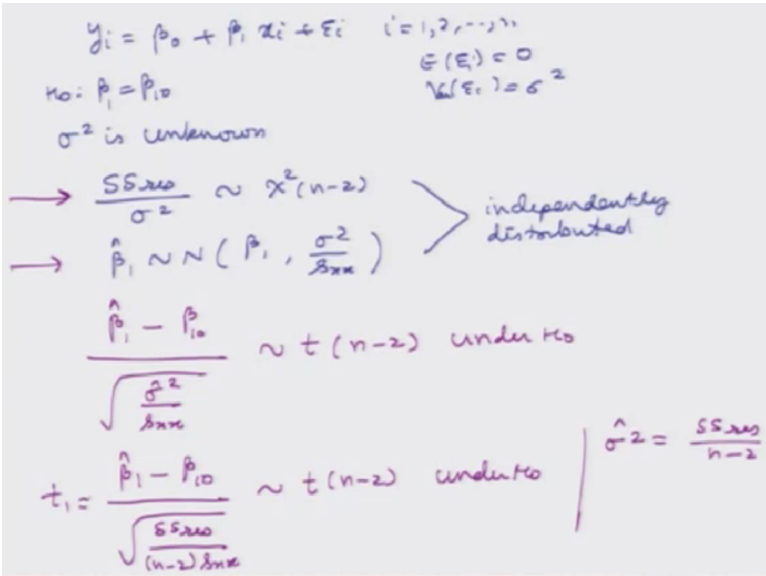
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology – Kanpur

Lecture 08
Testing of Hypothesis and Confidence Interval Estimation in Simple Linear Regression Model

Welcome to lecture number 8 you may recall that in the earlier lecture we had considered a model.

(Refer Slide Time: 00:20)



yi=beta0+beta1 xi+f salient i goes now 1 to n and we were trying to develop a test of hypothesis for h0 beta1=beta10, and we had consider the case how to test the hypothesis when sigma square is known to us. Now we are going to assume that sigma square is unknown.

Now if you try to see in practice actually this situation is more important or say more applicable than the earlier case where we assumed that sigma square is known to us. We had considered that while doing the linear regression modeling we have a only a set of data xi and yi and we have to infer everything on the basis of that sample of data, when I say that sigma square is known that is the value of sigma square has to be known from some past experience or from some other relevant sources.

In real life usually the value of sigma square is not known to us and we need to estimate it from the sample. We had earlier discussed that we can estimate the value of sigma square by sum of square due to residual that ss res/n-2. Now we are going to consider here a situation which is more applicable in real life data set, where we have a set of data on xi and yi and we try to estimate all the parameters only from the data set including beta1 as well as sigma square.

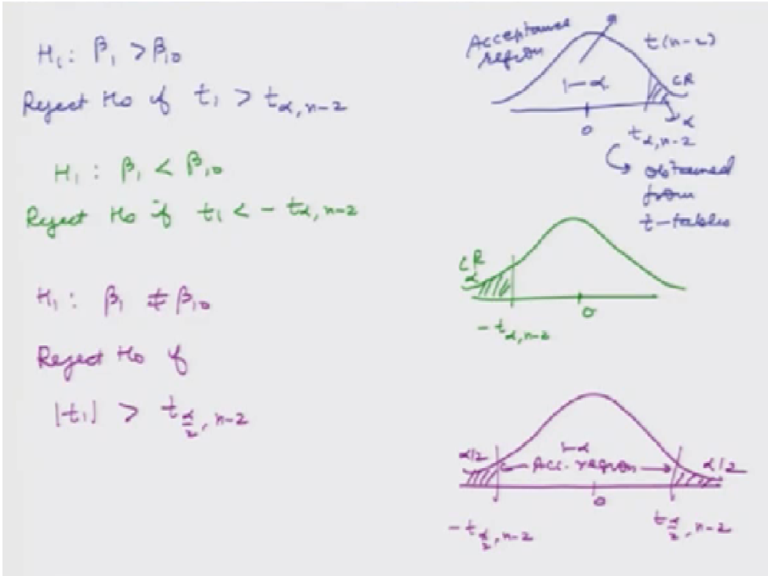
Where we assumed that accepted value epsilon i=0 and variance of epsilon i=sigma square. So in this case our next objective is how to find out a test statistic, so we know that ss res that is the residual sum of square divided by sigma square follows a chi-square and -2 degrees of freedom and we also have seen that this beta1 hat follows a normal distribution with mean beta1 and variance sigma square upon as sxx, and we can also prove that these 2 are independently distributed.

Now I can use a result from the statistical inference that if x is the random variable which is following a chi-square distribution and y is another random variable which is following a normal distribution then y upon square root of x by n this follows a t distribution. So in this case I can see here that this beta1 hat is a normal random variable this is following a chi-square random variable.

So, I can write down here beta1 hat – beta10, which is known to us divided by square root of sigma square hat upon sxx this follows a t distribution with n- 2 degrees of freedom under h0, so this can be further written as beta1 hat – beta10 upon square ss res divided by n - 2 sxx, this is follows a t distribution with n-2 degrees of freedom under h0.

Because if you try to recall we had estimated sigma square by ss res over n-2, now this is statistics we can denote by here say t1. Now we try see depending upon on the nature of alternative hypothesis we can have different types of decision rule, for example if my alternative hypothesis is h0 beta1 greater than beta10.

(Refer Slide Time : 04:26)



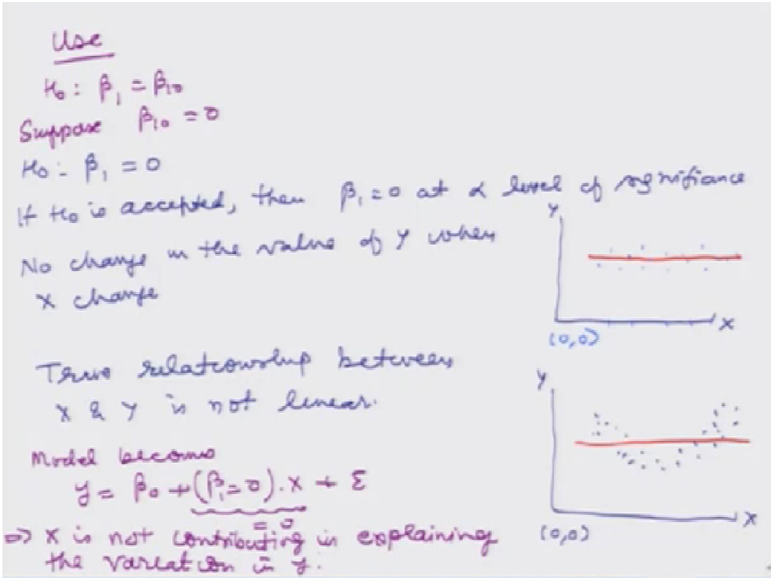
Then in this case as we have done in the earlier case this is the probably density function of t distribution with say here n-2, which is freedom and this is going to be the critical region of size alpha and this is going to be here expectance region of size 1- alpha and this is the critical value t alpha and -2 and this value t alpha n-2 because obtain from the t tables.

In this case I would say that reject h0 if t 1 is greater than t alpha n-2 on the other hand if my alternative hypothesis is h1, beta1 let than beta10 where beta10 is some known value then in this case this t distribution will be like this and the critical region will be on left hand side of size alpha and this value will be -t alpha and -2 and somewhere here will be 0, and in this case I would say that reject h0 if t1 is less than -t alpha and -2.

The third case will be of h1 beta10 =beta10, in this case again the t distribution will have critical reason on both the sides of size alpha by 2 and alpha by 2 and this will be the value -t alpha by 2 and n-2 and this will be the value t alpha by n-2, so I can say here that reject h0, if absolute value of t1 is greater than t alpha by 2n-2.

So this will again be our acceptance region of size 1- alpha, so this how we proceed to test the hypothesis about the h0, beta1=beta10. Next the question is what is the interpretation, and what is the utility of making this test of hypothesis, so let us try to understand the use.

(Refer Slide Time: 07:09)



Now when I am trying to do this hypotheses like h0 beta 0 = beta 0 and suppose beta10 =0 so essentially we are trying to now test the hypothesis beta1=0, now there are 2 options either our h0 is accepted are not so in case if I say if h0 is accepted then we say that beta1=0 hat alpha level of significance on the bases of given sample of data.

What is interpretation of this thing? We had learnt earlier that beta1is the slop of the line so when I am saying that h0 beta1=0 is accepted that means the slop of the fitted line is 0 hat means the fitted line is going to be something like this and all the points which are scattered around this line will look like this, so this means here what? This means the slop of the line is 0 that means x is not contributing in explaining the variation.

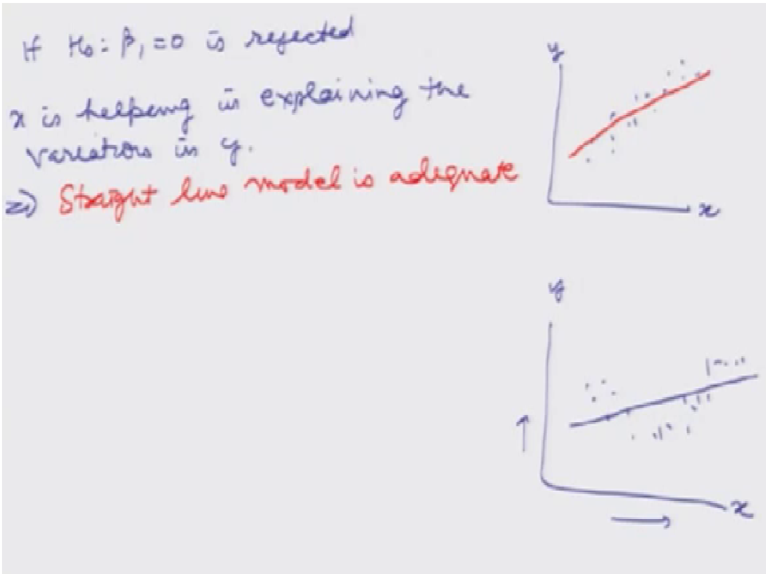
In y whatever happens here either here or here or say here or say here whatever the values x takes there is no change in the value of y, so what we can see here is the following there is no change in the value of y when values of x change. This means x is not contributing in my model building x 1 is such a variable which is not helping us in explaining the variation in the value of y.

The second situation can be it can be like this suppose we have a data like this on, which is really nonlinear and this case when we are trying to fit here a linear model this well look like this So what is this mean, that means there exists some relationship which is actually nonlinear and our x is trying to help the variation, but still it is not capable of explaining the variation.

This case I can say here in simple words the true relationship between x and y is not linear in either of the case h0 beta1=0 is indicating the true situation, when the relationship between x and y is nonlinear why should we fit a linear model. Well, in this case in this incase if somebody want to proceed than we have the setup of polynomial regression model that can be used.

Now we can see here was h0 beta1=0 is accepted then in that case our model now becomes say y=beta0+beta1 which takes active value here zero into x+ epsilon, so this quantity becomes =0 this implies that x is not contributing in explaining the variation in y on the other hand there can be another situation that suppose h0 beta1=0 is rejected.

(Refer Slide Time: 11:42)

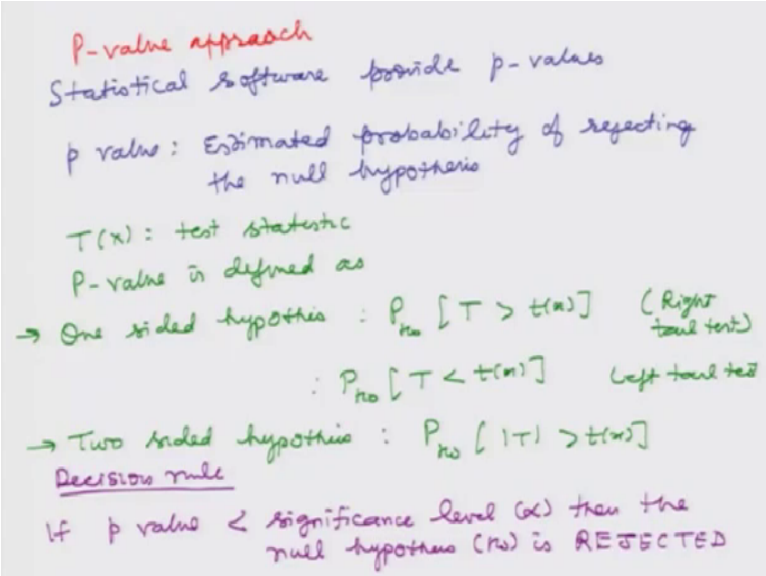


So what are we trying to say, we are simply trying to say that in this 2 dimensional graph x and y axis, yes there are some observation s like this and we are trying to fit here a model like this 1 that makes that sense, that means x is helping in explaining the variation in y or in simple words this implies that a straight line model is adequate, and that is essential what we wanted.

On the other end if I try take the second case over here where we had a nonlinear relationship like this 1 and suppose if we are trying to fit here a linear model this may look like a this, even in this case we can see here that x is changing the values of y's are changing, but that is possibly indicating a nonlinear relationship so it is suggested that by considering a polynomial regression model of say higher order like 2 or three that name work in this situation.

This is about the test of hypothesis using the classical approach, now there is another issue usually when we do this regression modeling we try to use a statistical software and nowadays statistical software gives us the outcome about that test of hypothesis using another criteria what is call us p value approach. So here I would like to explain you that what is this p value incase if you are working with the software and how to interpret it.

(Refer Slide Time: 13:52)



So statistical software provide p values, and in this case we do not look on that critical reason, but we simply try to look on the p value and based on that we try to take a decision whether my null hypothesis is accepted or not. This p value is nothing but this is the estimated probability of rejecting the null hypothesis, so in simple words if I say if tx is the my it test statistic which is for example in case of testing h0 beta1=beta10.

When we know the value of sigma square than tx is nothing but z, and incase if sigma square is unknown to us we try to estimated from the sample and then we use the t statistics So in this case the p value is defined as for 1 sided hypothesis this is defined as probability under h0 that t is greater than some given value small tx, this is for the right tail test and incase if I am considering left tail test than this is the probability and their h0 t less than tx.

So we are simply trying to compute the probability of a t being greater than the given value tx or smaller than tx under h0 that is when h0 is true. And when I have a 2 sided hypothesis then this is defined as probability under h0 absolute value of t is greater than say tx. Now the software will give us this value and that simply given by so called p value, what is the decision rule.

If p value is smaller than the significance level, which we had denoted has alpha then the null hypothesis, which we have denoted by h0 is rejected. This gives us a very simple solution just look at the software try to look at the p value and whenever we are doing the modeling and consequently we are doing test of hypothesis we have to define the level of significant usually it is 5% or 1% depending on the situation.

So alpha =0.05 or 0.01 depending on whether we are using 5% level of significant or 1% level of significance, so you simply try to compare the p value and value of alpha and based on that we can take a correct design about the test of hypothesis. This approach we need not look into the tables of either normal probability or say t probability so that is the advantage of this approach.

We stop here now and would I request all of you to just have a close look whatever we have covered in this lecture and the earlier lecture and try to understand how we have constructed the test statistics and how we are going to use it. Now our next objective will be first of all to develop that test statistics for testing hypothesis for intercept term and for the sigma square, and after that we will concentrate on the confidence interval estimation of all these parameters till then good bye.