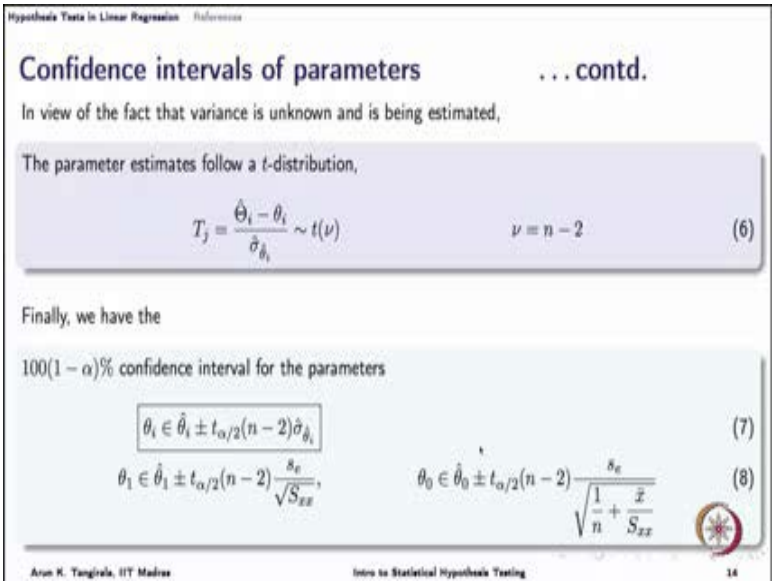
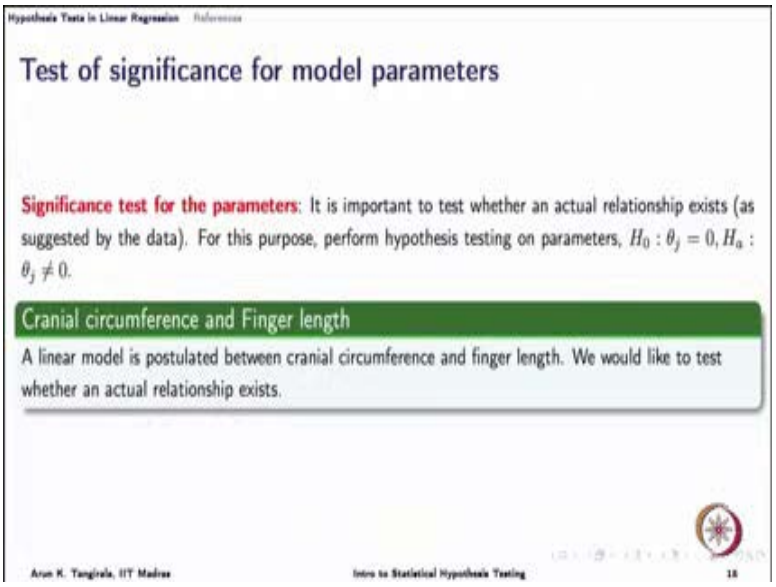
Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 17
Hypothesis Testing in Linear Regression

(Refer Slide Time: 00:09)

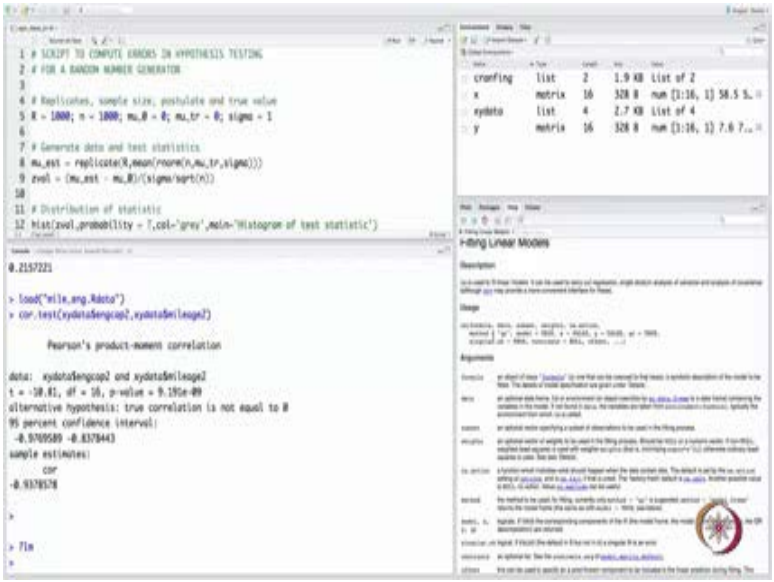


(Refer Slide Time: 00:12)



That is it. So, we are all set now to conduct significance tests on the model parameters. Let us go back and re-run this example that we worked out before in the context of correlation. Remember I said that, we could also infer the same thing, what is a same thing? That is the correlation being absent between the cranial circumference and the finger length based on observed data using linear regression as well and I will show you how to do this in r and we will also work out the highway mileage and engine capacity exercise and then close the discussion with some closing remarks. Let us get now to r and quickly work out the linear modeling example.

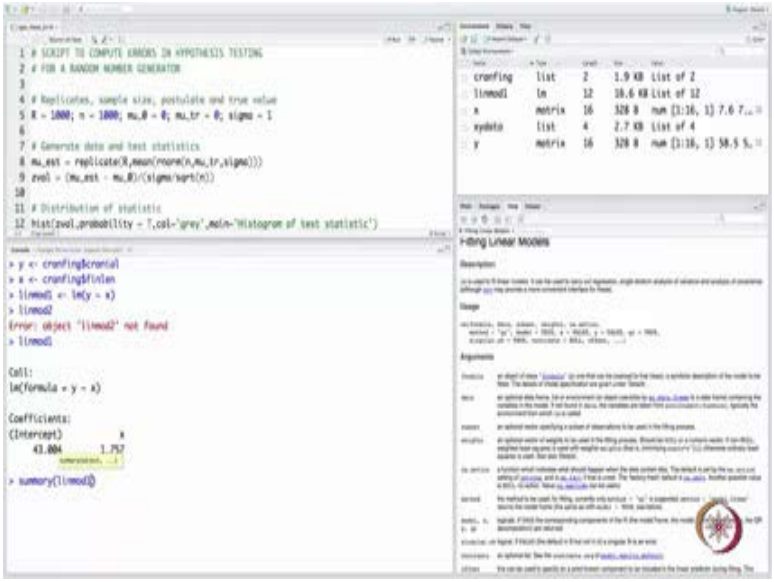
(Refer Slide Time: 01:00)



In r, a linear model is fit using the least squares approach using the routine l m. So, you can, I strongly encourage you to go through the help on this l m it is a very beautiful routine, it allows you to specify formula type that is you could write y till the I will show you how to write the formula. You can directly write the formula, if the variables are present in the work space or you can refer to a data frame and you can also run different variants of least squares which we have not discussed here. There are so many things that you could do with l m, I am going to only show you a plain vanilla implementation of this all right. In the correlation analysis really x and y did not matter so much because row x y is same as row y x, but in a linear regression problem it does matter even theoretically as to what you call as x and what you call as y because, remember we

assume x to be independent or x to be free of error, where as y is an observed one. Of course, in this cranial circumference and finger length problem both are measured quantities, measured variables. Both the cranial circumference and the finger length they may have both of them may have errors, but for the sake of discussion we will keep the finger length as the fixed or the regressor variable and the cranial circumference being the y variable. So, that is what we are going to do now.

(Refer Slide Time: 02:38)

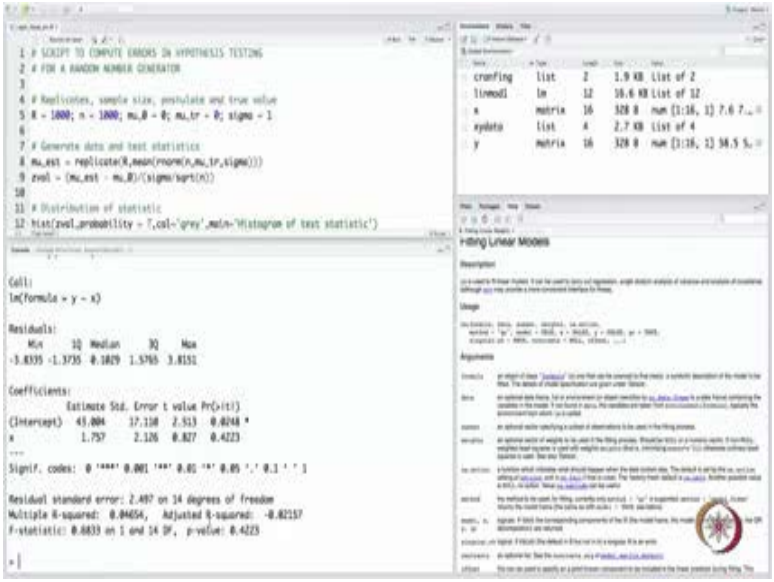


All right so now we are assigned y and x variable. Now working as I said with l m is a charm in r, let us calls this model as lin mod 1 and now we say y till the x. So, it say l m bracket open y till the x that is the way you write the formula or you supply the formula. It understands that there is also an intercept term, if you want to omit the intercept term then you would write x minus 1. Of course, you can also include other regressors in the multiple linear regression column that I am not going to show you that and that is it the linear model has been obtained. Now, we can display the coefficients from the linear model if you wish and here you have the intercept term and the slope that you have for this thing.

You can also build a model the other way around as I said you could assume finger length to be the dependent variable and the cranial circumference to be the regressor

nothing prevents you in this case because both are measured variables anyway. So, let us ask if the both the intercept and the slope estimates are statistically significant. There are several ways of doing that but let me actually show you a very nice way of displaying the summary of the linear regression that we have just gone through.

(Refer Slide Time: 04:29)



What you do is you ask for a summary of the model that you have identified, it tells you the formula that has been used; that means, a relationship that you have postulated and gives you some basic statistics for a residuals. We are interested in the coefficients or the estimates of the coefficients so it reports in the first column the estimates of the slope and intercept and in this case, it says that the standard error in the; it reports the standard error in the intercept and in the slope and reports t value as well and finally, gives you the p value.

This is the most important thing then one has to watch out for, we have again to summarize; the estimates of the intercept and slope in the first column, the standard errors in the second column, the t statistic that I showed you on the slide in the third column and the p value in the final column. We know now we can use this to conduct our hypothesis test, we can also look at confidence intervals. I will show you how to do that, but using the p value alone we can now conduct a hypothesis test of significance. What

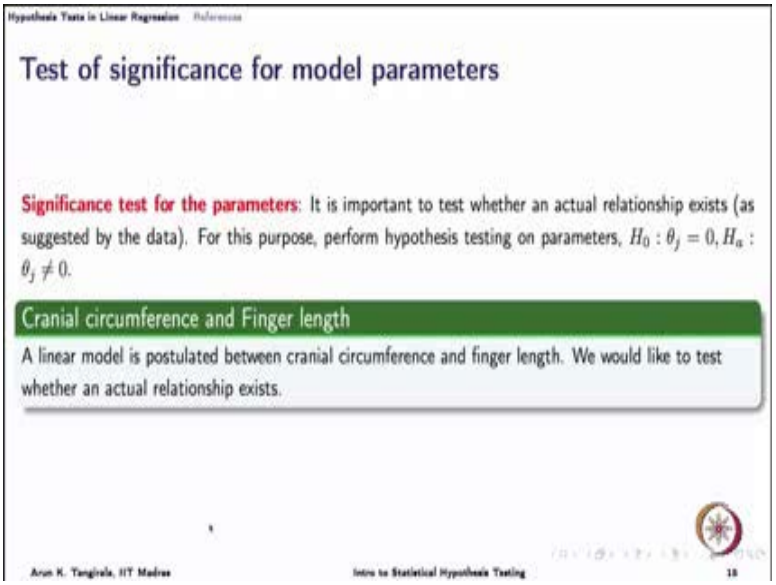
do you mean by test of significance? That the null hypothesis is theta 1 is 0; that means, a slope is 0, no linear relationship between y and x and intercept is 0; that means, when the finger length is 0, the true value of the or the average is 0, the average of the cranial circumference is 0. What do this p values tell me now? If I look at the intercept term, the p value that I obtained is less than the 0.05 significance level. Look at the notes or the legend that is given here, there is a star given next to the 0.0248. What does that star mean? You just have to look at the significance codes at the bottom and it says, at a significance level of 0.05 this intercept term is statistically significant; that means, a null hypothesis that the intercept is 0 is rejected because the p value is less than alpha.

On the other hand if the significance level is 0.0one then the p value is greater than alpha which means that at that significance level the null hypothesis that the intercept is 0 is also failed to be rejected, it is not rejected. What by mean by also is the slope in any cases, if you look at the p value it is quite high. You choose any of the standard significance levels; 0.05, 0.01 and 0.001 at all the 3 standard significance level that one chooses the p value is high. So, we failed to reject the null hypothesis that the true linear relationship is absent; that means, true theta 1 is 0 which means truly this data does not provide us any sufficient evidence to believe that there is a linear relationship between cranial circumference and finger length. The intercept term is not of so much of an importance, it is because it has got to do with the averages and we are not so worried about it. Generally we are worried about the slope because if that is theta 1 because if theta 1 is 0; that means, a linear relationship is absent; that means, there was no case for fitting a linear model which is what we saw earlier when we computed the correlation and performed significant tests on the correlation estimates remember that.

The summary of this exercise is that this data does not provide any evidence of a linear relationship between a cranial circumference and finger length. Now, there are also other statistics that are given out here at the bottom of the summary, the summary actually gives you a wealth of information about the model. It gives you residual standard error, remember. What is this here? This is s e; earlier we had given an expression for s square e, this is s e and you should cross check that the expression given in the slides will give you the same value that summary is reporting for you, 14 degrees of freedom because we have 16 observations and ignore the multiple R squared, I will quickly talk about what is

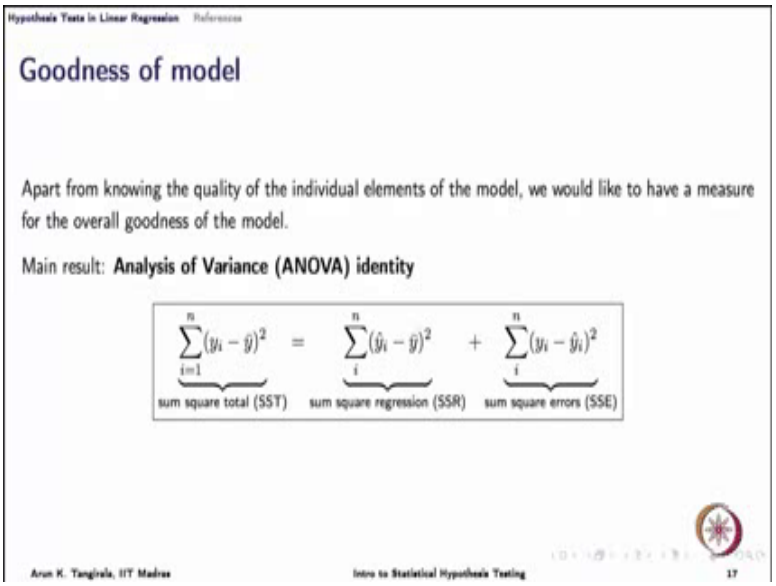
adjusted R squared and s statistics. So, that will kind of complete your understanding of what the summary actually throws out. We can also do a similar thing on the highway mileage and engine capacity problem, I welcome you to do that we will come to that is shortly once I discuss the adjusted R squared and the f statistic with that will close the discussion on linear regression, at least the hypothesis test for linear regression.

(Refer Slide Time: 09:46)



Both correlation analysis and linear regression analysis have shown us that, there is no linear relationship between cranial circumference and finger length based on the data that I have. May be if you collect a different data, a different conclusion may arise.

(Refer Slide Time: 10:03)



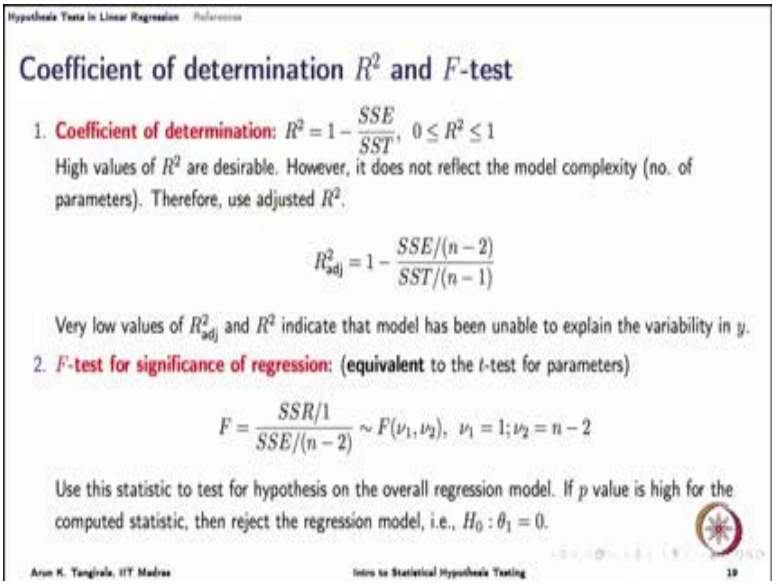
Now, we will come back to this highway mileage and engine capacity example shortly, before that I promised to discuss or explain what is this adjusted R square and f statistic. It is very straight forward to follow this until now we have talked of parameter estimates, but now we want to ask as a whole; that means, if I take an overview of the model, how well has the model managed to explain the relationship between y and x. Specifically, how well has the model managed to explain the variability in y as a linear function of the variability in x or as a function of the variability in x. Remember when I write y equals theta 1 x plus theta naught plus epsilon. I am postulating a linear model between y and x agreed, but that also means that the variability in y is also a function of variability in x. So, how well has a model managed to split y into 2 terms, one due to x and the other due to error because y equals theta 1 x plus theta naught plus epsilon means there is a part of y that is coming from x and that there is a part of y that is coming from epsilon, that is called a signal model or the data model. Now, we want to ask how what does the translate to in terms of decomposition of variance, y equals theta 1 plus theta naught plus epsilon is a decomposition of y, the signal or the data itself, variable itself.

Now, what about the variance? It turns out that, when you use least squares method to estimate the parameters and make predictions, it is very important only when you use a least squares method to compute the parameter estimates and then make the predictions,

you can decompose the so called sum squares total, which is a measure of the variance in y. In fact, if you write 1 over n, it gives you a it is an estimate of the variance of y, can be decomposed in to 2 terms, one due to x or due to regression which is sum of y hat minus y bar square, is also known as sum square regression and the other term being y i minus y hat square, that is the summation of that which is called naturally the sum squares errors.

What are least squares method has done is, although we have asked it to split y into x and epsilon; it has actually also split the variance of y into variance due to x and variance due to epsilon, this is called the analysis of variance because we are analyzing the variability in y, we are breaking it up into 2 terms and this kind of an expression is only possible when you use a least squares method to construct your predictions. To summarize model fitting is nothing but variance decomposition and we can use this result to set up hypothesis test for the goodness of fit, that is whether the regression was necessary or not, how good is a regression itself and also propose or define a measure known as the R squared measure.

(Refer Slide Time: 13:37)



Let me go through the R square measure first and then come back to the significance of regression test. So, this R square is known as a coefficient of determination or you know

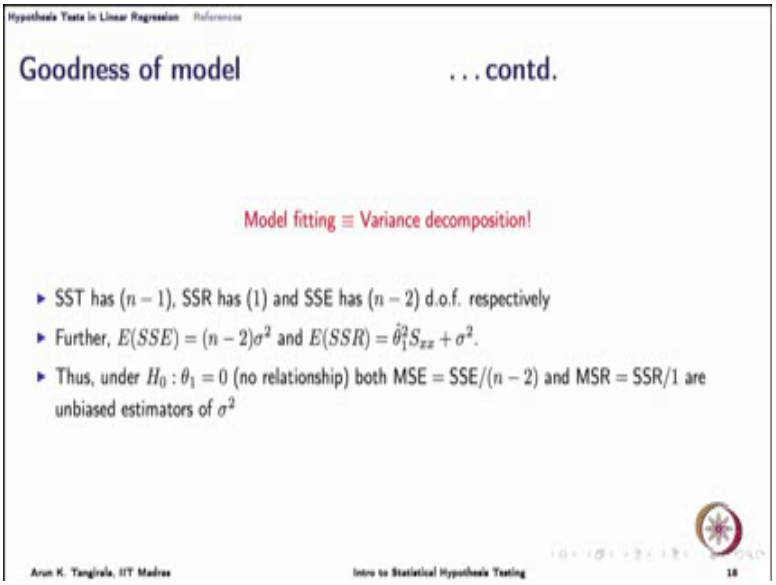
goodness of fit measure. It tells us how well x has managed to explain y using the linear model and it is natural to define this way 1 minus SSE sum square errors by sum square total for, if sum square error is 0; that means, x has fully managed to explain y then R square it is a value of 1. If x has absolutely not managed anything to explain in y, then explain meaning variance in y then R square hits a value of 0. So, R square is always between 0 and 1, again this is only true for least squares approaches only because only under least squares methods sum square error is always less than or equal to sum square total.

Now what happens is; obviously, we want high value of R square because we want the model to explain y or predict y as good as much as possible. In that process we can become greedy, what we mean by greedy is, we considered a linear model, but this concept of R square can be extended to non-linear models also as long as I am using least square approach. For example, I could have considered a polynomial model, I could have said y is theta 2 times x square plus theta 1 x plus theta naught plus epsilon so that I can actually explain more of y due to x. In fact, if I have n observations, I can construct a polynomial of n minus 1th degree so that sum squares error is 0 because if I have n observations and I fit an n minus 1th polynomial between y and epsilon I have exactly n parameters and n data points and therefore, I will obtain an exact fit, but having done that on the training data if you take this model to a fresh data set, it will fail miserably in terms of predicting because what we have done is the reality is that y contained epsilon which cannot be explained by x remember we have assumed co-variance between x lawn and epsilon to be 0. Which means that there is a part of y that cannot be explained by x, but we in a bit to explain everything in y have forced the model to explain only using x. As a result, it becomes highly conditioned on the data, it is like this person who is been trained only to answer a set of questions and when a fresh question is presented to this person, the person gives some terrible answer.

That is not how it should be, modeling in general you should remember this principle modeling should be such that when you are modeling y as a function of x, the art and skill in modeling is to be able to explain attribute whatever is due to x in a rightful way. You should not confuse a contribution of x and epsilon in any manner neither should be under fit nor should be over fit. Under heavy under fits would mean R square is 0, over

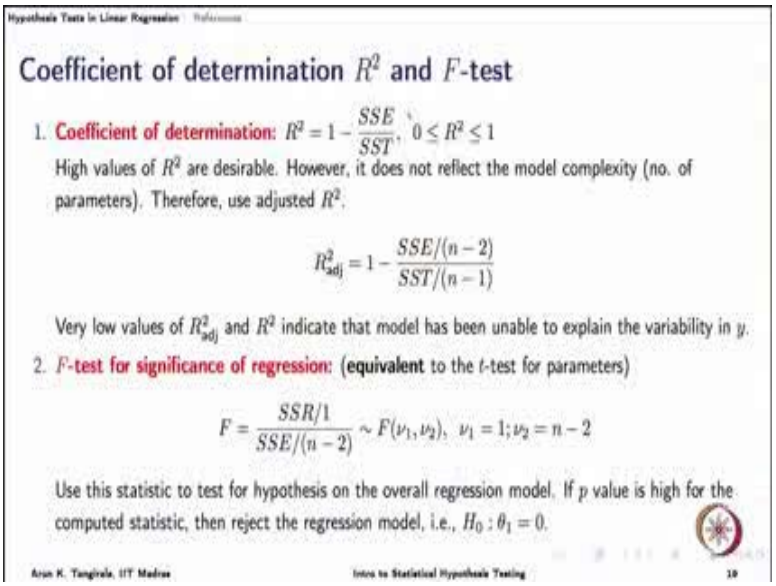
fitting, extreme over fitting would mean R square is 1. We want to avoid this the situation, unfortunately in R square there is no penalty for over fitting; that means, I can include any polynomial functions of x and take the R square to a value of 1 and that is why generally one does not recommend R square rather one recommends what is known as adjusted R square, which accounts for the or kind of includes a penalty for over fitting using what is known as a degrees of freedom. Here is where we go back to the sum square, total sum square regression and sum square error. The sum square total has n minus 1 degree of freedom naturally because we have used 1 degree of freedom to estimate y bar and sum square r has only 1 degree of freedom.

(Refer Slide Time: 17:52)



It can be proved, but we will not go into that and sum square error we have already discussed has n minus 2 degrees of freedom. The total degrees of freedom on the left and the right remain the same. Why are we talking about this, degrees of freedom? Because we want to now work with, we want account for the number of observations that are available or the degrees of freedom that are available to estimate the parameters. When I fit an n minus 1th polynomial to explain n observations then I do not have any degrees of freedom at all. I am solving an exact problem there is no degree of freedom so then I will run into problems error is 0.

(Refer Slide Time: 18:42)



Now, the other way of looking at it is replace a sum square error and sum square total in R square with their mean square values and the mean is obtained generally by dividing with the actual degrees of freedom available and that is what is being done in adjusted R square and that is what was reported for the cranial circumference and the finger length example. Let me take you back to that r summary, it reports the adjusted R square and it is pretty low indicating that this regression has not really managed to explain this model has not managed to explain anything in a cranial circumference using a linear function of the finger length so that is why this is a very valuable piece of information.

Now the adjusted R square unlike R square can be here the R square has is property that it lies between 0 and 1. This adjusted R square need not be exactly satisfying that, but the bottom line is very low values of adjusted R square would mean the regression was not so great. In fact, it is only a qualitative statement and very high values of adjusted R square means that yes that you have managed to explain in y using x very well without over fitting it. In fact, you can try actually fitting a very high parameter model or you know high order polynomial and see what happens to R square and adjusted R square, you will see the difference. Now the statistical way of testing for the significance of regression which is again equivalent to the t test for parameters, is to ask to set up a test statistic which compares two things, one it the sum square regression and sum square

error. In fact, the mean square regression and mean square error that is to be more precise.

Under the null hypothesis that theta 1 is 0, that is what we are saying is let us assume that the truth is theta 1 is 0, like we do in a null hypothesis test; that means, that the truth is there is no linear relationship and if it a model then what would happen to the mean square error and mean square regression. Now, if you look at the expressions for sum square error and sum square regression, we know that we can show that expected value of sum square error. That is a true mean of the error is sum square error is n minus 2 times sigma square and that expected value for sum square regression is theta 1 hat square times s x x plus sigma square. Ideally it should be theta 1 square so it should be theta 1 square times s x x plus sigma square; that means, that the sum square regression is reflecting the model that you are fit and the sum square error is only referring to the original sigma square that is a true sigma square.

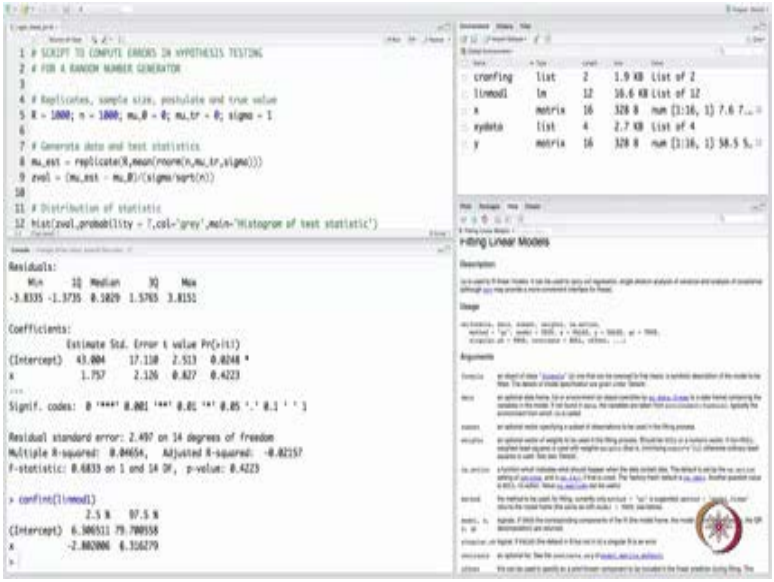
Suppose the true theta 1 is 0, suppose that is the case, then the sum square regression that the average value of sum square regression expectation of that is sigma square. Whereas regardless of whether the true theta 1 is 0 or not the expected value of sum square error is n minus 2 times sigma square. What does this tell us? When the null hypothesis that theta 1 equals 0 is true, whether I estimate the variance of epsilon using sum square regression or the sum square error by n minus 2, I should get identical estimates; that means, the ratio of sum square regressor by 1; divided by sum square error by n minus 2 should be ideally 1. What is a numerator here? It is a means; it is an estimate of sigma square obtained from sum square regression. Whether I look at essentially the sum square regression to estimate sigma square or sum square errors to estimate sigma square, the means or mean square error should be the same. So, under the null hypothesis therefore, this s statistics which is a ratio of the variances, variance estimates should ideally be 1.

Which is now essentially what we have done is, we have converted the test of significance on theta, to a test of ratio of variances that is what we have done. Of course this f test on ratio of variances is used for something else also, but will not go into that. I repeat the significance test for theta 1 has been now converted, earlier we did it using the p value or the confidence interval approach and so on. Now, we have now converted that

significance test for theta into a hypothesis test for the ratio of variances explained by the regressors, obtained by the regressors and obtained from the errors.

So, the bottom line is now we are going to conduct the hypothesis test on f the usual way on the ratio of variances. Let us go back to the regression summary that we had from r and at the bottom you see the s statistic being reported with the p value. Now the s statistics has been calculated with the appropriate degrees of freedom, remember in the slide we said the numerator has 1 degree of freedom and the denominator has n minus 2 degrees of freedom and the p value is reported here quite high, well higher than the significance level saying that the null hypothesis which is theta 1 is equal to 0 cannot be rejected.

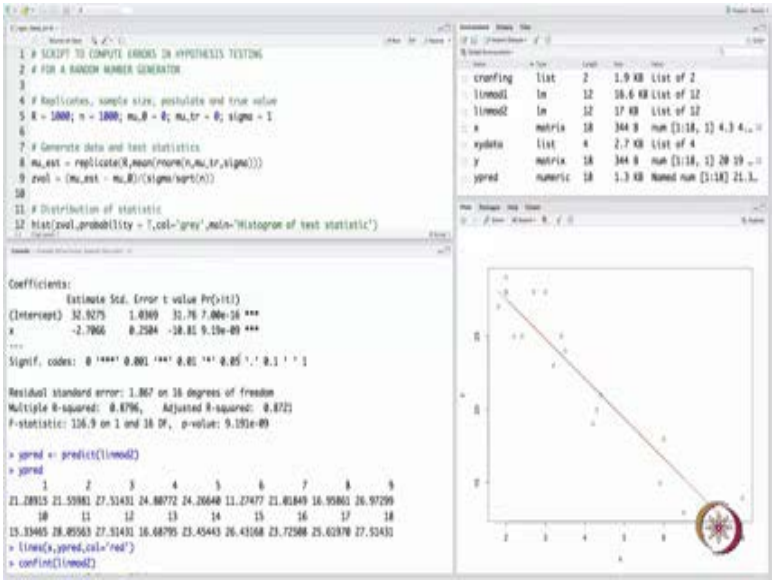
(Refer Slide Time: 25:37)



Therefore, again we come to the same conclusion that we made earlier using the p values. In fact, we can also do this using the confidence interval, there is a routine called conf int in r to which we can directly supply the model that you have estimated and it reports the confidence intervals for the slope and the intercept. If you look at the confidence interval for the slope clearly 0 is a part of the confidence interval, once again confirming that 0 has a true value for theta 1 is possible. So, we have learnt 3 different ways of looking at significance test for regression either I can look at the p value. In fact,

4 we can look at correlation, we can look at p value for the estimates that we have obtained, we can look at the f test and we can look at the confidence intervals, all of them should give me the same answer if everything has gone right. Now, we can turn to the other example, the highway mileage and engine capacity and then will close the discussion with a few closing remarks on residual analysis.

(Refer Slide Time: 26:18)



Let us now look at the highway mileage engine capacity example, where the engine capacity is a regressor and here we have x y data. Let us call this as x sorry, and y being the mileage so in the session on correlation coefficients we used engine cap 2 and mileage 2 and the reason is as follows, I will explain look at the a plot a scatter plot of y versus x and of course, what you notice straight away is, yes there is a possibility of a linear fit when you look at the scatter plot with the negative slope, which is something that we had commented on even when we perform the correlation access that the correlation may be negative. On top of it you look at most of the data points they are all in this region here whereas, there are this two data points that seem to be away from the general trend. What is happening here is as one increases the engine capacity, the highway mileage drops, that makes sense because as the size of the engine increases the highway mileage generally comes down that is the smaller the engine capacity better the mileages that you would get, which is kind of understandable physically one can

understand that. But if you look at the general highway mileage that one gets for these kinds of engine capacities here in the region 7 and 8, these 2 data points here show the unusually high mileage for the class of engines that have that engine capacity around that point.

Now, this data has been is available all over the web and I have borrowed this data set from the book by Ogunnaike. If you read the explanation in book by Ogunnaike, a clear explanation is provided on as to why these two data points and a fly away from the general trend, they correspond to vehicles which have been made by a lighter material. Therefore, the weight of the entire vehicle is much lower than the weight of the these the vehicles in this class of engine capacities. As a result of which one gets higher mileage; obviously, lighter the vehicle you can afford to get higher mileage because the power required is not so much to drive a lighter vehicle.

Therefore, there are other factors that go into determining the highway mileage, we have only considered engine capacity, in order not to allow this two data points to drive away drift our analysis, we omit this two data points and that is what is contained in the mileage 2 and the engine cap 2 so you can look up those data sets. In fact, now when we plot this x and y, you will see that those two data points have vanished now there seems to be a general trend. We have already computed the correlation coefficient between these 2 variables; we will now simply fit a linear model between y and x and ask for a summary.

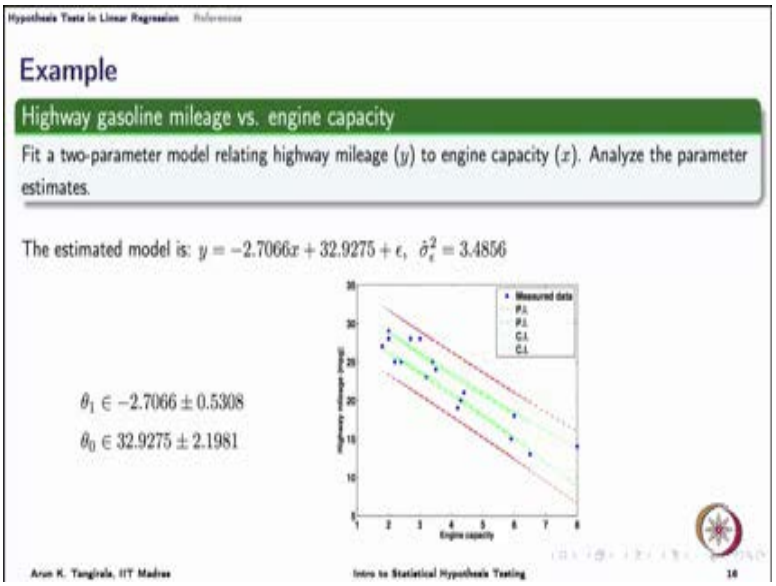
As usual we look at the estimates here of course, here also we have asked for an intercept. If I find there is no if I believe there should be no intercept model, we can go back and tell that information to l m. You look at the p values here for both the estimates extremely low, clearly lower than any of this standard significance levels telling us straight away that both theta 1 and theta naught are not in reality 0; that means, we reject the null hypothesis that theta 1 is 0 and theta naught is 0; that means, the regression model is significant it was necessary, not necessary, but the regression linear regression model has done a fairly good job of predicting it better than not fitting anything that is what it means and we can turn to even adjusted R squared, the adjusted R squared is quite high 87 percent which means it is done a good job. One can even predict we can

actually ask why predict that is on the data set, we can even use fitted there are so many other commands we can see how the fit has come out to be. So, when you have, when you supply this then it would give you the prediction and we can now super impose. Let us look at what y pred consists of, it is just bunch of predictions corresponding to the values of x that we have given in the data while training so we can now ask for sorry so there you go that is the fit that we have passing which is around which the data points are scattered and we can see once again that slope is negative. In fact, if you look on the at the confidence intervals on the parameters, we have here the confidence intervals of course, not including a 0.

You can see that both the lower and upper bounds that we have here are of negative sign for the slope, clearly indicating that the slope is negative and that is to be expected as per our discussion because we did say that it makes sense for the highway mileage to decrease as the engine capacity increases. The confidence intervals for both do not include a 0, once again confirming that the null hypothesis of 0 value theta 1 and theta naught have to be rejected and we can also look at the f test here to ask if the regression was significant and the f statistic has a very low value of p value, definitely lower than the standard significance levels again confirming the same thing.

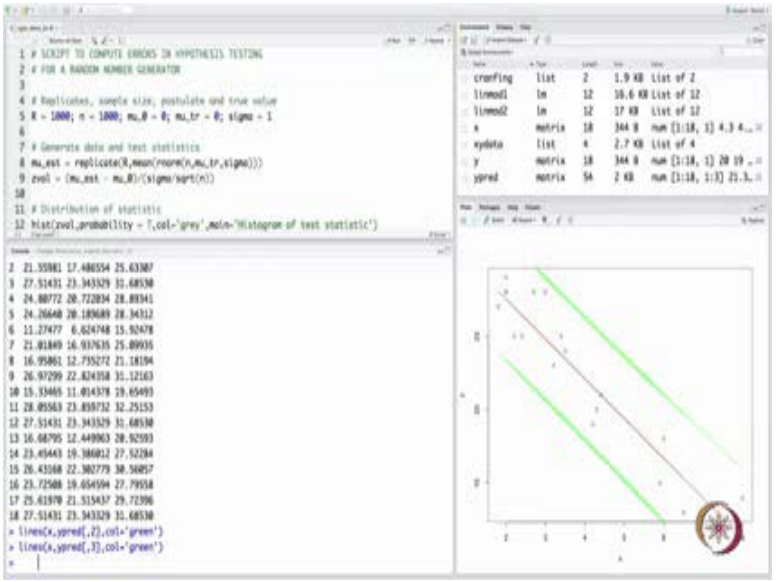
This is a proper way of performing a linear regression in a statistically meaningful manner. From here of course when you can; once you have the confidence regions that is it, I mean you are reporting a set of possible truths with of course, the 95 percent confidence is what your (Refer Time: 33:35) does, but you can change the significance level as well; either you report a confidence interval or you report a standard error, the standard error would be the sum square, sorry the square root of the variance or these values that are reported in the second column for the intercept estimate this is a standard error, for the slope estimate this is standard error. So, if everyone who fits everybody whoever fits a linear model, should either report the confidence interval or should first perform a significance test and then determine whether those terms should be include in the model, after having determined one should then go ahead and report the estimates with this standard errors and also show the line of fit.

(Refer Slide Time: 34:28)



Of course, in the plot that I show here, I also show the so called confidence intervals and prediction intervals that I talked about early on. One can construct these confidence intervals and the prediction intervals using the same predict routine and if you look up the help on predict routine; it would give you not only the fitted values, but the intervals.

(Refer Slide Time: 34:51)

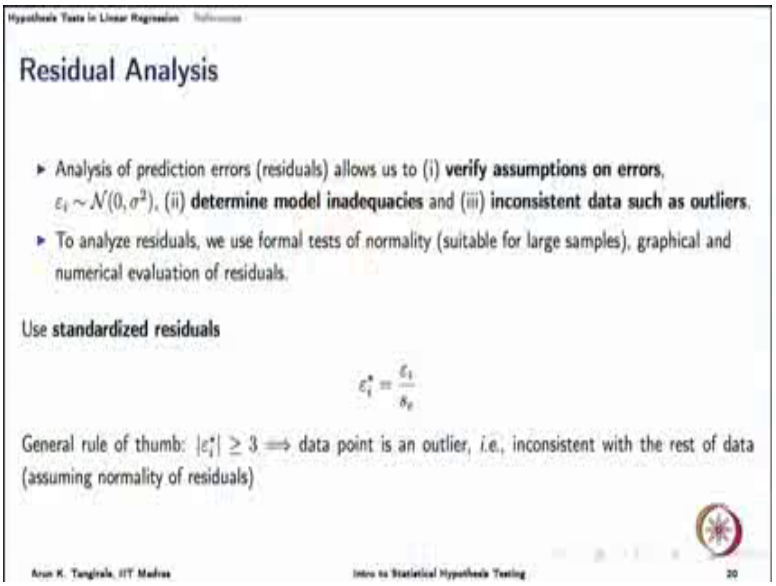


For example, earlier we predicted the, we made a prediction here where it simply took the data from the data contained in the linear model that is the training data. I can provide a new data, only when I provide a new data will it make so called predictions and then construct your prediction, intervals or confidence intervals. Suppose I wanted to make the prediction on the same training data set and give me confidence intervals and prediction intervals, what you do here is to construct the prediction intervals. You supply the new data which is nothing, but the training data itself of course, if you have a new data you can supply that as well fresh data and specify the type of interval that you want and if you look up the help you will see there are 3 type of intervals that it constructs the default if none.

You can ask for a confidence interval which would be for the mean response, mean prediction and or a prediction interval which would be for the prediction itself. So, if I ask for the interval here and let us say I asked for prediction one can even use a short forms for this. Now, you see y pred would have 3 columns, here the fit being in the first column and then the lower bound on the prediction interval in a second column, the upper bound on the prediction interval being in the third column. Earlier we had this plot of the fitted values, we can actually show, draw these predictions intervals so let us say here we can say lines y pred 2 color equals green. So, that is your lower prediction interval and then one can draw the upper prediction interval so this is what you see in the plot.

Of course, I have used different colors here do not get confused. I have used the red here in the plot for prediction intervals and green for the confidence intervals. Now if you look at this I have reporting the estimated model and also one needs to report the estimate of the variance of epsilon, which I am reporting here. The theta 1 and theta naught have laid in this confidence region, they are in this confidence region; minus 2.7066 plus or minus 0.5308 and the other one here 32.93 roughly plus or minus 2.2. This is how one reports the results from a linear regression exercise, hopefully now you have understood all the hypothesis tests that that are involved in a standard regression. I want to close this discussion or this entire lecture on the hypothesis tests and linear regression with some remarks on residual analysis which form a very important and integral part of regression, but I do not intend to elaborate on this.

(Refer Slide Time: 38:06)



Let me just conclude briefly here by saying that, residuals carry wealth of information they tell you how to improve the model, whether there is something in the model that requires more improvement. Of course, your adjusted R square or the f statistics is telling you whether the regression is significant, but whether there is any further scope for improvement is a question. What we have done until now is to determine whether it was even worth fitting a linear model, but we have not really verified, if the linear model has done the best in terms of there is nothing left to be explained.

We have not done any such test at all, all we have asked is; has a linear model explained anything at all? Yes, and then the next question is if then has a linear model managed to explain everything that was possible to be explained using x. Then let us assume that suppose y and x have a quadratic relationship, but I have fit only a linear model then what if there is a non-linearity that is missed out in it; obviously, that goes and sits in epsilon. Then of course your assumptions on epsilon that you have made may not be correct for your linear model and so on.

Let us say we do that and now we come to the residual analysis and verify if the assumptions on errors that we have made are correct because in calculating, in conducting our R square sorry calculating on adjusted R square, conducting the sorry s

statistic and constructing the confidence intervals, we have made certain assumptions on the errors. Now we are asking, if those assumptions are correct and therefore residual analysis is important. In fact, in a proper course on linear regression analysis what I generally teach is, first you should go through the residual analysis and then only you should construct look at the confidence regions, but what I have done here is because of the nature of the course, I have not followed that very strict procedure because then it involves me, it requires me to explain the various tests that are involved in a residual analysis.

So, the summary is whatever tests that we have performed until now rests on certain assumptions of the residuals. You should actually perform tests to determine whether those assumptions have been met on the residuals on the errors through the help of residual analysis and then only go back and construct the confidence regions, perform your f test and the rest of the things. What are the assumptions that we have made, Gaussianity we have assumed errors to be at least uncorrelated if not independent; that means, there is nothing no pattern in epsilon or whether that there are no outliers, for example extreme values are not present.

The residual analysis is a formal way of testing the normality assumption, the uncorrelatedness or the independence and the well behaveness of the data; that means lack of outliers. Generally, one works with standardized residuals for example, to test for outliers or uses a q q plot for normality assumption or looks at also makes a visual analysis of epsilons to see if they are random. There are other formal tests for asking if epsilons are uncorrelated using what are known as auto correlation tests. You can look at auto correlation of epsilon that is the serial correlation between epsilons and ask, epsilons meaning for each observation and put conduct a test of significance of the correlation between observations to be 0 and so on.

But that requires lot of other explanations and quite a bit of theory, other theoretical details that we have not talked about in this course therefore I will skip that part. The objective of this exercise was to show you what kind of hypothesis test typically arise in a linear regression. Of course, even in residual analysis you do get hypothesis test, for example, you can set up a hypothesis test that epsilons follow Gaussian distribution, that

is a hypothesis tests or you can say that the epsilons are uncorrelated that is another hypothesis test. But we have not talked about those more advanced hypothesis tests which are of course, critical in a linear regression problem.

Therefore, I strongly recommend that you read through the remaining portions of that particular chapter in the book by Montgomery and Runger or by Ogunnaike and get the complete picture. Hopefully then now that you are more comfortable with linear regression you know what kind of hypothesis tests are involved, what to watch out for, how to report the results of a linear regression problem and that it is very important to analyze errors. Although I am not shown you that, in fact I leave it as an exercise to you to go back and to the examples, the cranial circumference and finger length or the highway mileage and engine capacity.

Go back and check the residuals, plot the residuals and see if those residuals are meeting the assumptions of course, you have very few observations there but still you can plot the residuals and see if they look random or if there is a pattern. You can for those of you are familiar with q, q plots just use a q, q plot routine and see if the assumption of normality is reasonably satisfied and see if epsilons have any outliers and so on. So, go through that to complete the linear regression exercise all right then. Remember therefore that the first test that one should perform is correlation and of course, the test of correlation rests on the bivariate normality assumption and also confirm the results of your correlation test with the linear regression. But up front performing a correlation test is very useful because it gives you a lot of insights, even in cases where non Gaussianity assumption is not satisfied strictly.

With those words, we come to a conclusion on this lecture, there is a short lecture that is left out which talks about the power of hypothesis tests and in general what are the factors affecting the goodness of a hypothesis tests and with that we will close the course itself. Hopefully, you have been enjoying the course and learnt a lot.

See you in the last lecture very soon.