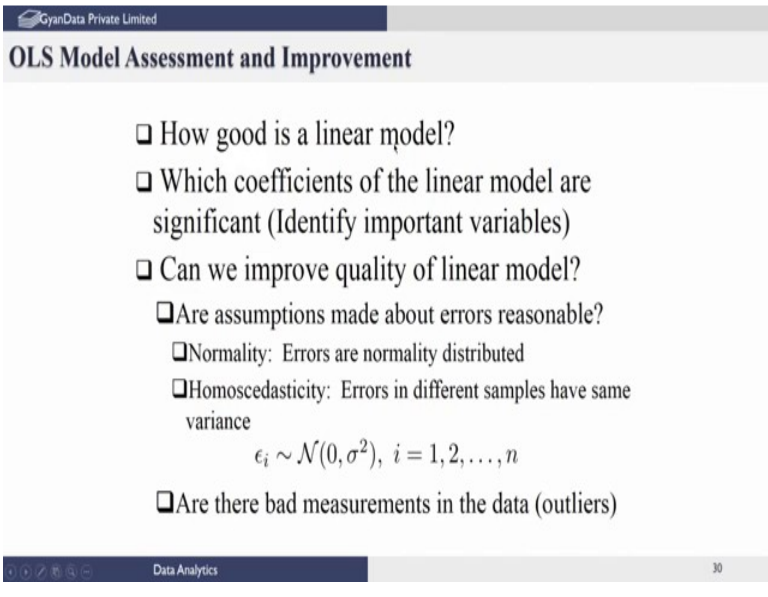
Data Science for Engineers
Prof. Shankar Narasimhan
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 33
Model Assessment

In the previous lecture, we saw how to fit a linear model between two variables x which is the independent variable and y which is the dependent variable using techniques called regression and in this particular lecture we are going to assess whether the model we have actually fitted is reasonably good or not. There are many methods for making this assessment, we will look at some of these. So, what are the useful questions to ask when we fit a model? The first question to ask is whether the linear model that we have fitted is adequate or not, Is good or not. If it is not good then perhaps we may have to go and fit an non-linear model. So, this is the first step that you will actually test whether the model is good or not?

(Refer Slide Time: 01:06)



Then even if you fit a model you may want to find out which coefficients of the linear model are relevant. For example in the one variable case that we saw one independent variable the only 2 parameters that we are fitting are the intercept term β_0 and the slope term β_1. So, we want to know whether we should have fitted the intercept or not whether we should have taken it as 0. When we have several independent variables in multilinear regression we will see that it is also important to find out which variables are significant, whether we should use all the independent variables or whether we should discard some of them.

So, this particular test for finding which coefficients of the linear model are significant is useful not only in the univariate case but more useful in multi linear regression, where we are would not identify important variables. Suppose, the linear model that we fit is acceptable then we would not actually see whether we can improve the quality of the linear model. When fitting linear model using the method of least squares we make several options about the errors that corrupt the dependent variable measurements.

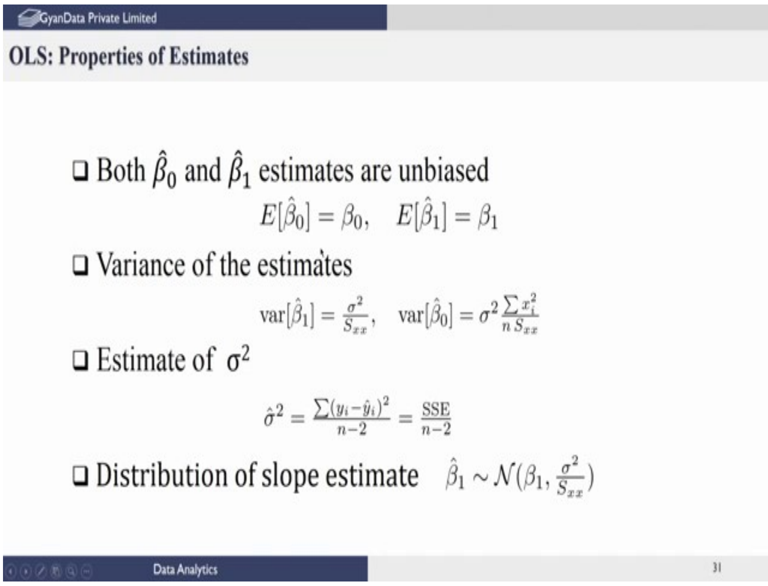
So, are these assumptions really valid? So, what are some of the assumptions that we make about the errors that corrupt the measurements of the dependent variable. We assume that the errors are normally distributed. Only this assumption can actually justify the choice of the method of least squares. We also assume that the errors in different samples have the same variance. So, this is called homoscedasticity assumption. So, we are assuming that the errors in different samples are also having the same variance.

In general, the these two statement assumptions about the errors that they normally distributed with identical variance can be compactly represented by saying that εi the error corrupting measurement i is normally distributed with zero mean and σ^2 variance. Notice the σ^2 is same and does not depend on i which means it is the same for all samples $i = 1$ to n that is the assumption we are making when we use the standard method of least squares.

Now, we also assume that all the measurements that we have made are reasonably good and there are no bad data points or what we call outliers in the data. We saw that even when we are estimating a sample mean, one that data can result in a very bad estimate of the mean. So, similarly in the method of least squares if we have one bad data point, it can result in a very poor estimate of the coefficients. So, we want to remove such bad data from our data set and improve maybe fit a linear model only using the remaining measurements and that will improve the quality of the linear model.

So, these are some of the things that we need to actually verify. These assumptions what we are made about the errors whether they are reasonable or not if there are bad data, can be remove them or not. And so, we will look at the first two questions in this lecture which is to assess whether the linear model that we are fitted is good and how do we decide whether the coefficients of the linear model are significant.

(Refer Slide Time: 04:33)



So, before we start, we need to derive some properties of these estimates that we have the derived. Remember, that the coefficients of the linear model that we are fitted which is the intercept term β_0 and the slope term β_1. These are obtained from data from the sample of data that you are given. We have indicated that these are estimates and not the true values by putting this caret term on top of each of these symbols which means that this is an estimate $\hat{\beta}_0$ is an estimate of the true β_0 and $\hat{\beta}_1$ is an estimate of the true β_1 which we do not know.

However, we can prove based on the assumption we have made regarding the errors that the expected value of $\hat{\beta}_0$ will be = β_0. What does it mean? If we were to repeat this experiment, collect an-other sample of n measurement and apply the method of least squares, we will get another estimate of β_0. Suppose, we do this experiment several times and we will get several estimates of $\hat{\beta}_0$ let me average all of them and average value of that will tend towards the true value. That is what this expression means that if we were to repeat this experiment several times the average of the estimates that we derive will actually, represent, be a very good representation of the truth.

Similarly, we can show that the expected value of $\hat{\beta}_1$ is equal to the true value β_1. Notice that β_0 and β_1 are unknown values we can we are only saying that the expected value of $\hat{\beta}_1$ will be true value and the expected value of $\hat{\beta}_0$ will be the true value and such statements are also known as if the estimates satisfy such properties we also call these estimates as unbiased, there is no bias in the estimate of $\hat{\beta}_0$ or $\hat{\beta}_1$.

The second important property that we need to derive about the estimates is the variability of estimates. Notice we get different-different estimates of $\hat{\beta}_0$ depending on the sample that we have derived and therefore, we want to ask what is the spread of these estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ if we were to repeat this experiment. We can show again through based on the assumptions we have made that the variance of $\hat{\beta}_1$ will be = σ^2 by Sxx. Sxx represents the variance of x or x - $\bar{x}$ the whole squared summed over all the samples. Whereas σ^2 represent the variance of the error that corrupts the dependent variable y. So, σ^2 is the error variance, S xx is the variance of the independent variable. So, this ratio we can show will be equal to the variance of $\hat{\beta}_1$.

Similarly, we can show that the variance of $\hat{\beta}_0$ is σ^2 which is the variance of the error multiplied by this ratio the numerator is the sum squared values of all the independent variables, while the denominator represents the variance of the independent variable. In this the S xx can be computed from data, σ of xi^2 can be computed from data, but we may or may not have knowledge about the variance of the error which corrupts the dependent variable that depends on the instrument that was used to measure the dependent variable.

If you have some knowledge of this instrument accuracy we can take the σ^2 from that, but in most cases data analysis cases we may not have been told what is the accuracy of the instrument used to measure the dependent variable. So, σ^2 also have to somehow be estimated from the data. We can show that we can derive a very good estimate of σ^2 by this quantity that is described here which is nothing but the difference between the measured value yi and estimated value ŷi which is obtained from the linear equation.

We have fitted a linear model, so, for every xi we can predict from the linear model what is the estimate of ŷi for every sample. Then we can take the difference between the measure and the predicted value of the dependent variable sum squared divided by n - 2 that is a good estimate of σ hat squared which is the error in the dependent variable. Now, why do we divide by n - 2 instead of n - n or n - 1? Very simple, ŷi was estimated using the linear model. It had 2 parameters β_0 and β_1 which represents means that 2 of the data points have been used to estimate β_0 and β_1 and therefore, only the remaining n - 2 samples are available for estimating this σ squared.

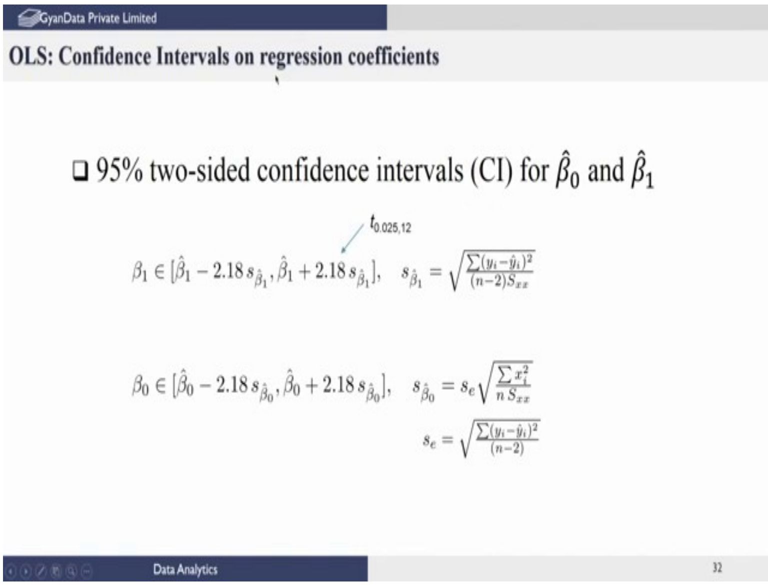
Suppose, you had only two samples then your numerator would be exactly 0, because you are more than two samples your variability and that variability is caused by the error in the dependent variable. That is one of the reasons that you are dividing by n - 2 because two data points have been used to estimate the parameters β_0 and β_1. Now, this particular numerator term is also called the sum squared errors or SSE for short and so, $\hat{\sigma}^2$ is nothing, but SSE divided by n - 2.

So, from the data after you have fitted the model you can compute this value and compute the SSE and obtain an estimate for σ^2. So, you do not need to be told the information about the accuracy of the instrument used to measure the dependent variable you can get it from the data itself.

So, now finally, not only we have got the first moment properties of $\hat{\beta}_0\hat{\beta}_1$ as well as the second moment properties which is variance of $\hat{\beta}_1$ and variance of $\hat{\beta}_0$, we can also derive the distribution of the parameters. In particular $\hat{\beta}_1$ can be shown to be normally distributed. Of course, with because the expected value $\hat{\beta}1$ is β_1 it is normally distributed with β_1. The true unknown value of β_1 is the mean and the variance given by σ if you substitute $\hat{\sigma}^2$ here you can finally, show that this is nothing, but σ oh, I am sorry. So, this is unknown σ^2 divided by S xx, σ^2 is essentially here we have derived this σ^2 by Sxx is the variance of $\hat{\beta}_1$.

Now, if you do not know σ^2 you can replace this σ^2 is with this $\hat{\sigma}^2$ SSE by n - 2. So, once you have derived the distribution of the parameters we can perform hypothesis testing on the parameters to decide whether these are significantly different from 0 and that is what we are going to do. We can also derive what we call confidence intervals for these estimates based on their distribution characteristics that is the mean and the variance.

(Refer Slide Time: 11:48)



Now, the first thing we will do is to develop confidence intervals. Confidence intervals simply says what is the interval within which the true value unknown values likely to be with 95 percent confidence or 90 percent confidence you can decide what size confidence interval size you need to have and correspondingly you can obtain the interval from the distribution.

So, if you want a 95 percent confidence interval also known as CI and it is two sided because it could be either to the left of this estimated value or to the right of the estimated value. So, we are obtaining the 95 percent confidence interval for β_1 from its distribution. You knowing it is normally distributed with some unknown variance. So, that we can actually derived from the from this particular range which is the estimated value of β_1, which is $\hat{\beta}_1$ + or - 2.18 times these standard deviation of b $\hat{\beta}_1$ estimated from the data.

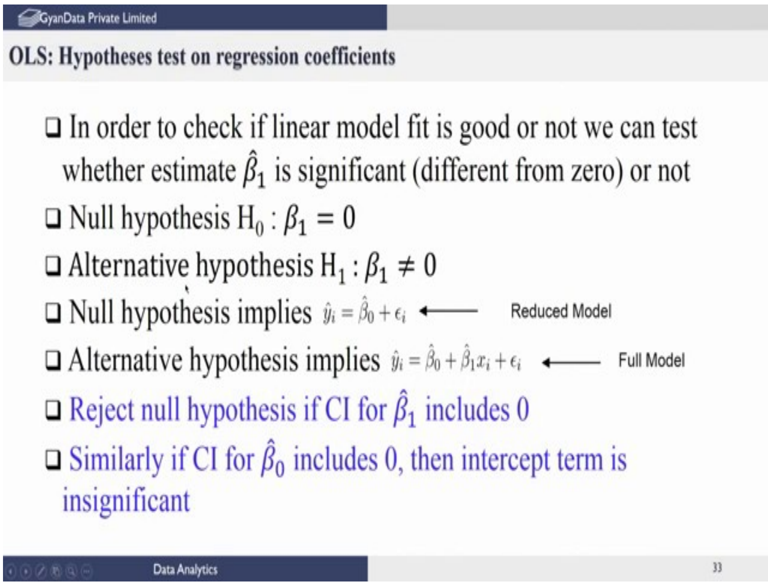
Notice this is very similar to the normal thing which says that the true value will between estimate + or - 2 times the standard deviation. The reason why we have 2.18 instead of 2 is because we are no longer obtaining the critical value from the normal distribution, but from the t distribution because σ^2 is estimated from the data and not known up priori.

So, the distribution slightly changes it is not the normal distribution, but the t distribution and that is what we are pointed out here. The steep one 2.18 is nothing, but the critical value 2.5 percent critical value, upper critical value with 12 degrees of freedom. Why 12 degrees of freedom because, you have in this particular example we had fourteen points and we used two of the points for estimating the two parameters. So, n - 2 is the degrees of freedom with represents 12. In general, depending on of number of data points this value 2.18 will change. So, that changes is the degrees of freedom of the t distribution from which you should pick the upper and lower critical value. So, lower critical value is - 2.18, the upper critical values 2.18, 2.5 percent. So, the overall is 5 percent. 90 no, this confidence interval represents the 95 percent confidence interval for β_1.

So, what all we are going to state is that the β_1 to unknown β_1 lies within this interval with ninety five percent confidence that is what we are saying. $\hat{\beta}_1$ can be estimated from data s $\hat{\beta}_1$ can be estimated from data. So, you can construct the confidence interval. Similarly, you can construct the 95 percent confidence interval for β_0 from its variance. So, we are doing the same thing $\hat{\beta}_0$+ or - 2.18 times standard deviation of $\hat{\beta}_0$ estimated from data which is what we call s $\hat{\beta}_0$.

Remember, s $\hat{\beta}_0$ is σ^2 which is estimated from data multiplied by this square root of σxi^2 divided by n times Sxx which is nothing, but the square root of what we have derived in the earlier thing with σ^2 replaced by the estimated quantity. That is all this these two terms represents s $\hat{\beta}1$ and s $\hat{\beta}_0$. So, having constructed this 95 percent confidence interval you can also use this for testing whether β_0 is unknown $\beta_0 = 0$ or the unknown $\beta_1 = 0$ or not which is what we will do.

(Refer Slide Time: 15:29)



So, let us look at why would want to actually do this hypothesis test. We have fitted linear model assuming that you know that there is a linear dependency between x and y and we have obtained an estimate of $\hat{\beta}_1$. Also we have also fitted an intercept term. We may want to ask should is the intercept term significant. Maybe the line should be pass through 0, 0 the origin, maybe the y variable that is not depend on x1 in a significant manner which means $\hat{\beta}_1$ is approximately equal to 0 that unknown β_1 is exactly equal to 0 although we have got some estimate for $\hat{\beta}_1$ non zero the estimate for $\hat{\beta}_1$.

So, the null hypothesis what we want to test is $\beta_1 = 0$, versus the alternative that β_1 is not = 0. If $\beta_1 = 0$ it implies that the independent variable x has no effect on the dependent variable, but on the other hand if you reject this null hypothesis we are concluding that the independent variable does have some effect on the dependent variable.

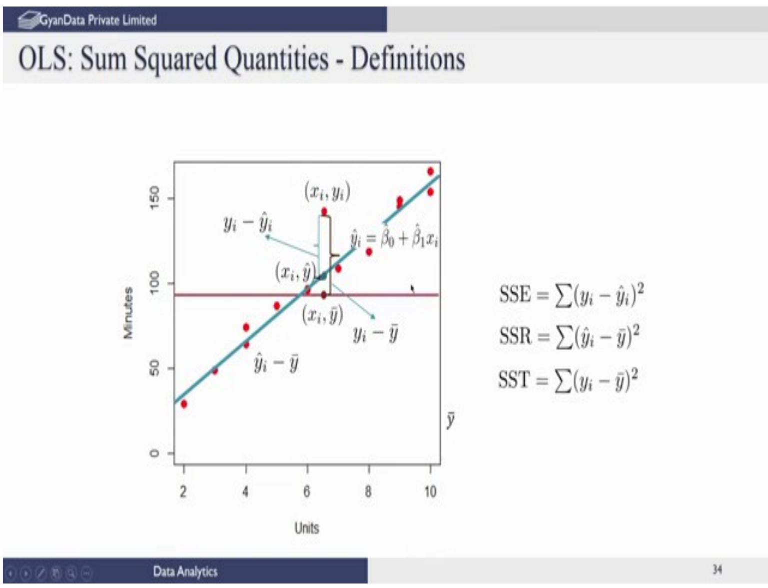
So, this particular hypothesis test can be also reinterpreted as the null hypothesis implies $\beta_1 = 0$ which means what we are doing is only at of yi = a constant whereas, if we accept in or reject the null hypothesis then we are actually fitting a linear model with β_0 and β_1 present. So, the null hypothesis represents the t of a reduced model which involves only a constant whereas, the rejection of the null hypothesis or the alternative hypothesis implies that we believe there is a linear model that relates y to x.

So, between these two models we want to pick whether the reduced model is acceptable or maybe the full model is to be accepted and the reduced model should be rejected that is what we are doing when we test this hypothesis $\beta_1 = 0$ versus β_1 not = zero Remember, the β_1 can be the positive or negative and that is why we are doing a two sided test.

So, we can do it 2 ways we can actually reject the null hypothesis if the confidence interval for $\hat{\beta}_1$ β_1 includes 0. So, notice that we have constructed the confidence interval for β_1. So, this term $\hat{\beta}_1$ - 2.18 maybe negative and this maybe positive in which case the interval includes 0 and then we have to definitely we will might make a decision that that β_1 is insignificant and actually true $\beta_1 = 0$. On the other hand if both these quantities if the interval is to the left of 0 which is completely negative or to the right of 0 which means both these quantities are positive, then this interval will not contain 0 and then we can make the conclusion reject the null hypothesis at β_1 equal is 0 which means β_1 is significant. So, from the confidence interval itself is possible to make the reject or the null hypothesis.

So, we can extend this kind of analysis to even test whether β_0 is 0 or not. So, if the confidence interval for β_0, this particular interval, includes 0, then we will say that the intercept term is insignificant otherwise we will say the intercept term should be is insignificant and should be retained in the model. So, let us actually when we do a final example we will see this. There are other ways of performing this test. And we will continue the we will do that also because that is very useful when we come to multilinear regression. In the univariate regression we have only these two parameters, but multilinear regression there are several parameters you will have one corresponding to each independent variable and therefore, there will be lot more hypothesis test you will look. Therefore, will extend this kind of an argument to test for $\beta_1 = 0$ or β_1 ! = 0 using what is called a F test which we will go through.

(Refer Slide Time: 19:37)



So, before performing the F test to check whether a reduced model is adequate or we should accept the full model we will use some definitions for sum squared quantities. Notice that let us say that we had set of data in this case we have the example of the number of units that were repaired and the time taken in minutes to repair the units by different sales person and we had fourteen such data points, fourteen such salesman, who have actually reported the data. So, the red points actually represents the data and the best, the linear t, using the method of least squares using all the data points we got something that is indicated by the blue line.

Now, suppose we believe that constant model is good, then we would have actually fitted this particular horizontal line would be the best t representing $\bar{y}$ the best estimate of constant model is the mean of y for all values of x are prediction best prediction for yi is the mean value of yi which means x has no relevance β_1 is 0, so, we will estimate the best constant t for yi is this mean value. So, the red line represents the best t when we ignore β_1 the slope, the blue line represents the best t of the data when we include the slope parameter β_1.

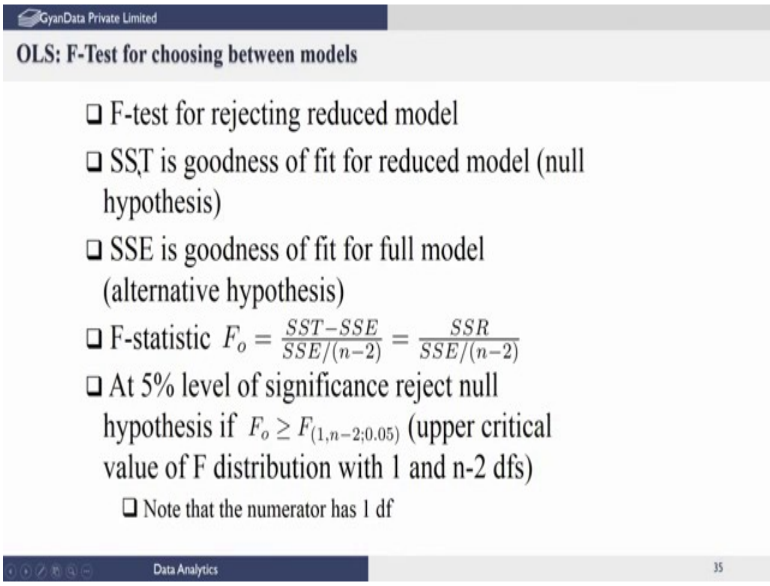
Now, let us look at certain sum squared deviation the deviation between yi and $\bar{y}$ which is the redline best t of the constant. This distance is yi - $\bar{y}$ and sum squared of all these vertical distances from the point to the red horizontal line constant line that is what we call the SS total or sum squared total which also represents the variance of yi - $\bar{y}$ the whole squared. All that we have not done is divided by n if we have divided by n or n - 1, we have got the variance of y, but this represents sum squared errors in yi when we ignore the slope parameter, that is another way of looking at it.

The distance between yi and ŷi. So, now suppose we assume that the slope parameters relevant then we would have fitted this blue line and for every xi let us take this xi yi corresponding to this independent variable, the predicted value of yi using this linear model would be the intersection point of this vertical line with the blue line which represented by the blue dot which is what we call ŷi. And therefore, this vertical distance between the measure and the predicted value is the sum squared errors, is called SSE, $(yi - \hat{y}i)^2$ and this is the total error if we include the slope parameters in the t.

So, the difference between these two quantities SS total - SS error will be equal to what is also called the sum squared residual which is nothing, but the predicted value - the mean value y bar sum squared over all the data points. Now, we can show that SST will always great be great greater than SSE because SSE was obtained by fitting two parameters there for you should be able to reduce the error ok, maybe marginally, but you will be always able to be able to reduce the errors.

So, SS total is the will always be greater than SSE and therefore, this difference SSR will also be positive all of these a positive quantities. Now, one can you separate SS total as the goodness of t if we assume a constant model, we can interpret SSE as the goodness of t of the linear model and therefore, we can now use this to perform a test. Literally intuitively we can say that if the reduction by including the slope parameter that is SST - SSE is significant, then we conclude it is worthwhile including this extra parameter, otherwise not. This can be converted into hypothesis test formal hypothesis test and that is what is called the F test.

(Refer Slide Time: 24:03)



So, what we are doing it as I said that SS total is a measure of how good the reduced model is which is reduced model here implies a constant model whereas, the SSE represents how good the linear model if we include this slope parameter. So, we are asking whether the reduced model should be accepted which is the null hypothesis or should be rejected in favour of this alternate which is to include the slope parameter. So, as I said the F-statistic for doing this hypothesis test is to compute the difference in the goodness of fit for the reduced model which is always higher - the goodness of fit SSE for the alternative hypothesis.

So, this represents the sum squared errors for the reduced order model fit, SSE represents the goodness of fit for the alternate hypothesis fit. This difference if it is large enough as I said then we can actually say maybe it is worthwhile going with the alternate hypothesis rather than null hypothesis. So, SSR which is the difference between this should be large enough.

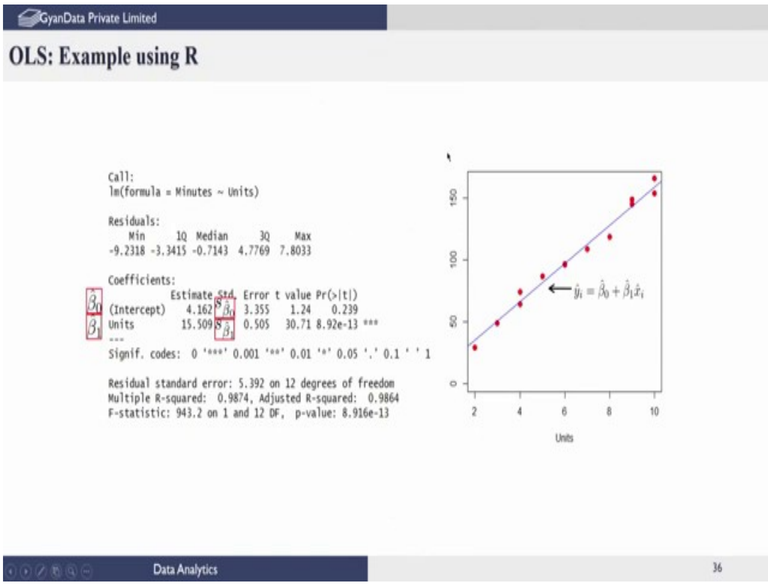
So, normalisation what the denominator represents in some sense of percentage SSE is the error obtained for the alternative hypothesis. Remember because of the different number of parameters used in the model we have to take that into account the numerator SST has n - 1 degrees of freedom because we are fitting only one parameter, this has n - 2 of freedom because you fitting 2 parameters. So, the difference actually means it is only one extra parameter. So, there is numerator which is SSR has only one degree of freedom which is n - 1 - n - 2 whereas, the denominator SSE has n - 2 degrees of freedom because it has 2 parameters which is fitted. So, we are dividing the SSE by n - 2 the number of degrees of freedom.

So, average sum squared errors per degree of freedom that is what we are saying, and that is your normalisation SSR divided by this normalised is this quantity and we can show formally that is an F-statistic because it is a ratio of two squared quantities and each squared quantities is itself a χ squared variable because it is a square of a normal variable. Therefore, this the ratio of two χ squared and we have seen in hypothesis testing the ratio of two χ squared variable is an F distribution with appropriate degrees of freedom. The numerator degrees of freedom is 1, the denominator degrees of freedom is n - 2.

So, if we want to now do a hypothesis test using this statistic F naught we compare F naught with the critical value from the F distribution. Notice F is actually a positive quantity. So, we do a one sided test if we choose the level of significance as 5, 5 percent then we choose the upper critical value from the F distribution with 1 and n - 2 degrees of freedom and 5 percent level of significance or what we call the upper critical values probability is 5 percent 0.05. So, once we get this from the F distribution we got this threshold and if the statistic exceeds the threshold then we will reject the null hypothesis and say the full model is better than the reduced model we will accept the full model or we say we reject to reduced model in favour of the full model that the slope parameter is worth including in the model we will get a better t. That is how we actually conclude.

So, there are now several ways for deciding whether the linear model that we have fitted is good or not. We could have used r squared value we said that if it is close to + 1 then we should that is one indicator that the linear model maybe good. It is not sufficient what I call sufficient to conclude, but it is good indicator we can also do the test for β significance of β_1 if we conclude that β_1 is not significant then maybe then a linear model is not good enough we have to find something else or we can do an F test and conclude whether including the slope parameter is significant. So, these are various ways by which we can decide that the linear model is acceptable or not or the t is good. We cannot stop this we have to do further test, but at least these are good in initial indicators that we are on the right track.

(Refer Slide Time: 28:38)



So, let us apply this to the example of repair of or the servicing a problem where we have fourteen data points and the time taken and the number of units repaired by different salesman are given. So, in this case we have this fourteen points which we have showed we have fitted the data using R. Remember, that l n is the is the function which we should call for fitting a linear model and here we are predicting the dependent variable is minutes and independent variables is units and once we have fitted this using the R function it gives out all of these output and it gives you the coefficient. The intercept term turns out to be 4.16 to the slope parameter turns out to be 15.501.

But, also it also tells you what is this standard deviation, estimated standard deviation, of this parameter which is S β_0 of the intercept it also tells you what is the standard deviation of this estimate for β_1 which turns out to be 0.501 505 all of this calculations from the data using the formulas we have described. Now, once it has given out that we can actually now perhaps construct confidence intervals and find out whether this these are significant not or R itself actually tells you something whether these if you run hypothesis test whether you can will conclude whether β_0 is significant or β_1 a significant and that is indicated by what is call this a p value that it is reported.

So, if you get a very high value, t values represents the statistic which have again described earlier. So, it has computed the statistic for you for β_0 and the statistic for testing whether $\beta_1 = 0$ or not and it has computed this statistic value and it has compared with the critical value while the distribution, t distribution with appropriate degrees of freedom and concluded that the upper critical or the probabilities 0.239

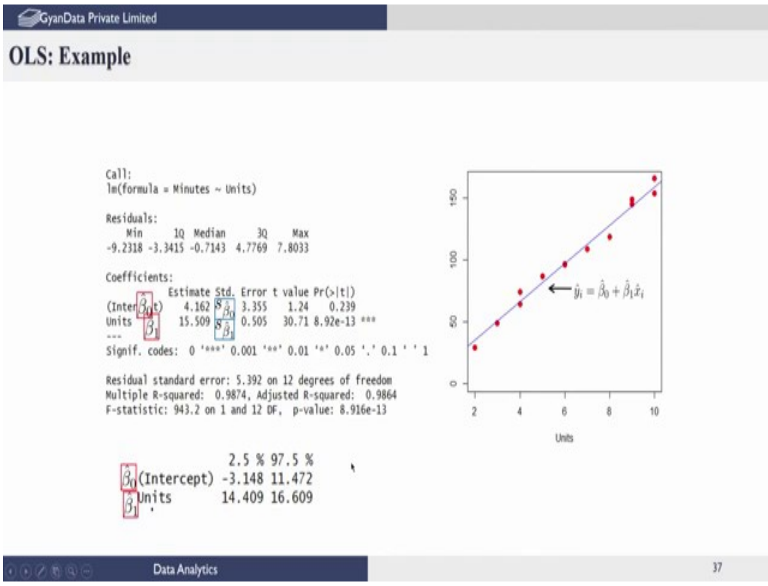
which means if you get very high value for this anything greater than 0.01 or 0.05, it means you should reject the you should not reject the null hypothesis on the other hand if you get a very low value it means you should reject the null hypothesis, with greater confidence you can reject the null hypothesis.

So, in this case all it is saying is, if you choose a level of significance 0.001 you would not reject the null hypothesis. If you choose 0.05 you will not reject the null hypothesis, if you choose 0.01 as your level of significance you will not reject the null hypothesis. So, that is what the star indicates. At what level of significance will you reject it. Whereas, in the case of β_1 you will reject the null hypothesis which means you will conclude that β_1 significant even if you choose very low significance value 0.05, 0.01, 0.001 or even lower value. In fact, up 10 power - 13 you will end up rejecting the null hypothesis. Very low type one error probability if you choose also you will reject the null hypothesis.

So, therefore, you can conclude from these values that $\hat{\beta}_0$is insignificant which means $\hat{\beta}_0$= 0 is a reasonable hypothesis, $\hat{\beta}_1$ is not = 0 is a reasonable hypothesis. Let us go and see whether this makes sense of this data. We know that if the no units are repaired then clearly no time should be taken by this sales repair person. Which means because you have not taken any time for servicing, yes because you have not repaired any units. So, this line technically should pass through 0, 0 and that is what you have said. But, however, we went ahead really and fitted a intercept term, but the tests for hypothesis says you can safely assume β_0 the intercept is 0 it makes physical sense also and we could have only fitted β_1 that is good enough for the data.

So, perhaps you should redo this linear t with β_0 and only using β_1 and you will get a slight different solution and you can test again. So, another way of deciding whether the significant whether the slope parameters significant or not this to look at the F statistic. Notice F statistic is very high and this p values very low which means you will reject the null hypothesis that the reduced model is adequate implying that you should use β_1 including β_1 is very good you will get a better t using β_1 in your modeling. So, the high value of the statistic indicates that it will reject the null hypothesis or a low value of p value for this F this F statistic indicates that you will reject the null hypothesis even at very low significance level.

(Refer Slide Time: 33:40)



You can also construct the confidence interval for β_0 and β_1 and from the earlier thing you say approximately it is estimated + or - 2.18 times the standard error and that is what it is deemed 4.1 + or - 2.18 times 3.35 and that turns out to give the that gives the interval confidence interval - 3.148 to 11.472. That means, with 95 percent confidence we can claim that the true β_0 lies in this interval. Similarly, we can construct the interval confidence interval for β_1 hat, 90 percent confidence interval and it turns out it is 15 + or - approximately two times 0.5 which is 14 and 16.6.

Now, clearly the interval confidence interval for β_0 includes zero and therefore, we should not reject the null hypothesis β_0 = 0. We should simply accept that β_0 perhaps = 0 whereas, interval for confidence interval β_1 does not include 0. So, we can reject the null hypothesis that β_1 includes 0 and the slope is an important parameter to retain in the model.

Now, all this we have done only for single thing. We will be extending it to the multi-linear case and we will also look at other assumptions, the influence of bad data and so on in the following lecture. So, see you in the next lecture.