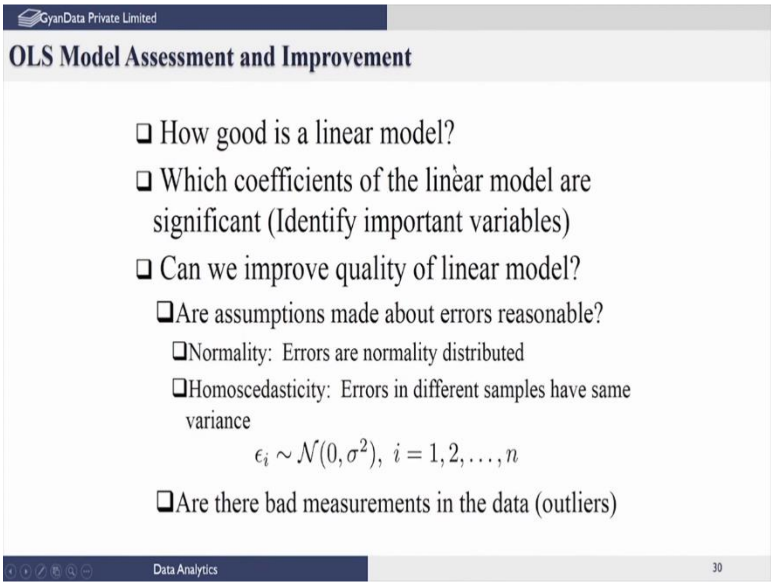
Python for Data Science
Prof. Raghunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 28
Model Assessment

Good morning everyone. 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 the 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 a non-linear model. So, thi`s is the first step that we will actually test whether the model is good or not.

(Refer Slide Time: 01:07)



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 two parameters that we are fitting are the intercept term $\beta 0$ and the slope term $\beta 1$.

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 multi linear 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 uni varied case, but more useful in multi linear regression where we want to identify important variables. Suppose the linear model that we fit is acceptable, then we want to 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 assumptions 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; now this is called the homoscedasticity assumption.

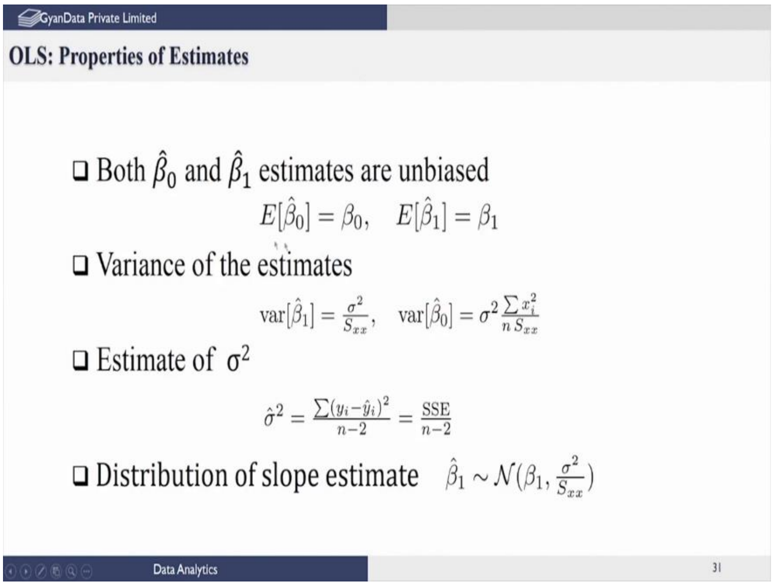
So, we are assuming that the errors in different samples are also having the same variance. In general the these two statements assumptions about the errors that they are normally distributed with identical variance can be compactly represented by saying that ϵ_i ; the error corrupting measurement i is normally distributed with 0 mean and sigma square variance. Notice that sigma squared is same and does not depend on i which means it is the same for all samples i equals 1 to n that is the assumption we are made 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 bad 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 may be fit the linear model only using the remaining measurements and that will improve the quality of the linear model. So, these are some of the things we need to actually verify. These assumptions what we have made about the errors, whether they are reasonable or not if there are bad data can we 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 have fitted is good and how do we decide whether the coefficients of the linear model are significant.

(Refer Slide Time: 04:35)



So, before we start we need to derive some properties of these estimates that we have derived. Remember that the coefficients of the linear model that we have 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 have given.

We have indicated that these are estimates and not the true values by putting this karat term on top of each of these symbols; which means that this is an estimate $\widehat{\beta_0}$ is an estimate of the true β_0 and $\widehat{\beta_1}$ is an estimate of the true β_1 which we do not know. However, we can prove based on the assumptions we have made regarding the errors that the expected value of $\widehat{\beta_0}$ will be=β_0. What does it mean? If we were to repeat these experiment collect another sample of n measurements 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 $\widehat{\beta_0}$; let me average all of them and the average value of that will tend towards the true value; that is what this expression means, that if we were to repeat these experiment several times the average of the estimates that we derive will actually be a very good representation of the truth. Similarly, we can show that the expected value of $\widehat{\beta_1}$=the true value β_1; notice that β_0 and β_1 are unknown values.

We are only saying that the expected value of β1 hat will be the true value and the expected value of β0 hat will be the true value and such statements are also known as if the estimates satisfies such properties, we also call these estimates are unbiased there is no bias in the estimate of $\widehat{\beta_0}$ or $\widehat{\beta_1}$. The second important property that we need to derive about the estimates is the variability of the estimates.

Notice, we get different estimates of β0 hat depending on the sample that we have derived. And therefore, we want to ask what is the spread of these estimates of $\widehat{\beta_0}$ and $\widehat{\beta_1}$; if we were to repeat this experiment; we can show again through based on the assumptions we are made that the variance of β1 hat will be=sigma squared by S_{xx}; S_{xx} represents the variance of x or $(x - \bar{x})^2$summed over all the samples; where as σ^2 represents the variance of the error that corrupts the dependent variable y.

So, sigma squared is the error variance, S_{xx} is the variance of the independent variable. So, this ratio we can show will be=the variance of $\widehat{\beta_1}$. Similarly, we can show that the variance of $\widehat{\beta_0}$ 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 sigma, of x_i^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 we have some knowledge of this instrument accuracy; we can take the sigma squared 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, sigma squared also have to somehow be estimated from the data; we can show that we can derive a very good estimate of sigma squared by this quantity that is described here which is nothing but the difference between the measured value yi and the estimated value yi hat which is obtained from the linear equation we have fitted the linear model.

So, for every x_i we can predict from the linear model what is the estimate of $\widehat{y_i}$ for every sample, then we can take the difference between the measured and their predicted value of the dependent variable; sum squared divided by n-2 that is a good estimate of sigma hat squared which is the error in the dependent variable. Now, why do we divided by n-2 instead of n-n or n-1?

Very simple $\widehat{y_i}$ was estimated using the linear model it has two parameters β_0 and β_1 which means the two 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 sigma square ok. Suppose, you had only two samples then your numerator would be exactly 0 because you have more than two samples you have 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.

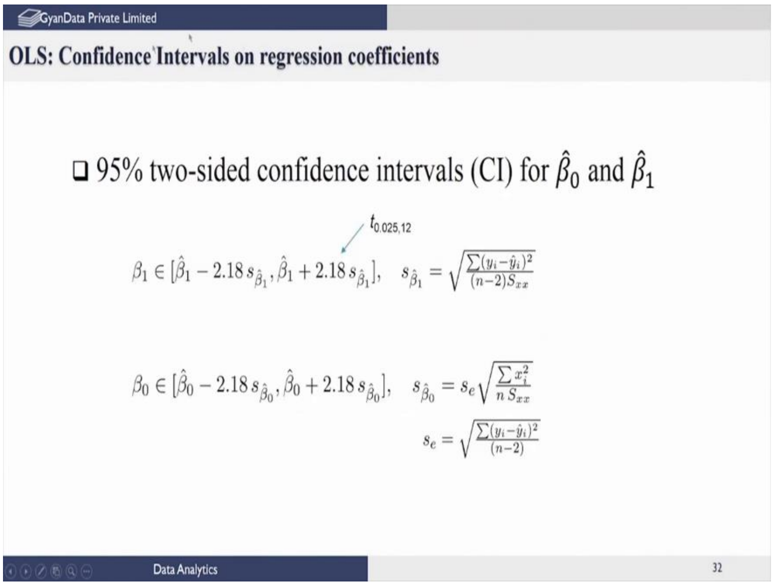
Now, this particular numerator term is also called the sum squared errors or SSE for short and so $\widehat{\sigma^2}$= SSE /(n-2). So, from the data after we have fitted the model we can compute this value and compute this SSE and obtain an estimate for sigma hat square. So, you do not need to be told the information about the; accuracy of the instrument used to measure the dependent variable, you can get it from the data itself ok.

So, now finally not only we have got the first moment properties of $\widehat{\beta_0}$, $\widehat{\beta_1}$ as well as the second moment properties which is variance of $\widehat{\beta_1}$ and the variance of $\widehat{\beta_0}$; we can also derive the distribution of the parameter in particular $\widehat{\beta_1}$ can be shown to be normally distributed. Of course, with because the expected value of $\widehat{\beta_1}$ is β_1; it is normally distributed with β_1 the true unknown value of β_1 as the mean and the variance given by sigma; if you substitute the sigma hat squared here, you can finally, show that this is nothing, but I am sorry.

So, this is the unknown σ^2divided by S_{xx} ok; sigma squared is essentially here we have derived this σ^2 by S_{xx} is the variance of β1 hat. Now, if you do not know sigma squared you can replace the sigma squared with this sigma hat squared SSE by n-2 ok.

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: 11:49)



Now, the first thing we will do is to develop confidence intervals; confidence interval simply says what is the interval within which the true value unknown value is 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 95 percent confidence interval also known as CI and its 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 knowing its normally distributed with some unknown variance. So, that we can actually derive from the from this particular range which is the estimated value of β_1; which is $\widehat{\beta_1}$ ± 2.18 times, the standard deviation of $\widehat{\beta_1}$estimated from the data.

Notice, this is very similar to the normal thing which says that the true value will lie between estimate ±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 not known apriori.

So, the distribution slightly changes it is not the normal distribution, but the t distribution and that is what we have pointed out here this 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 14 points and we used two of the points for estimating the two parameters.

So, n-2 is the degrees of freedom; which represents 12 in general depending on the number of data points this value 2.18 will change ok. So, that changes 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 value is 2.18; 2.5 percent. So, the overall is 5 percent this confidence interval represents the 95 percent confidence interval for β_1.

So, all we are going to state is that the β_1 true unknown β_1 lies within this interval with 95 percent confidence; that is what we are saying ok. $\widehat{\beta_1}$ can be estimated from data so you can construct this confidence interval. Similarly, you can construct the 95 percent confidence interval for β0 from its variance.

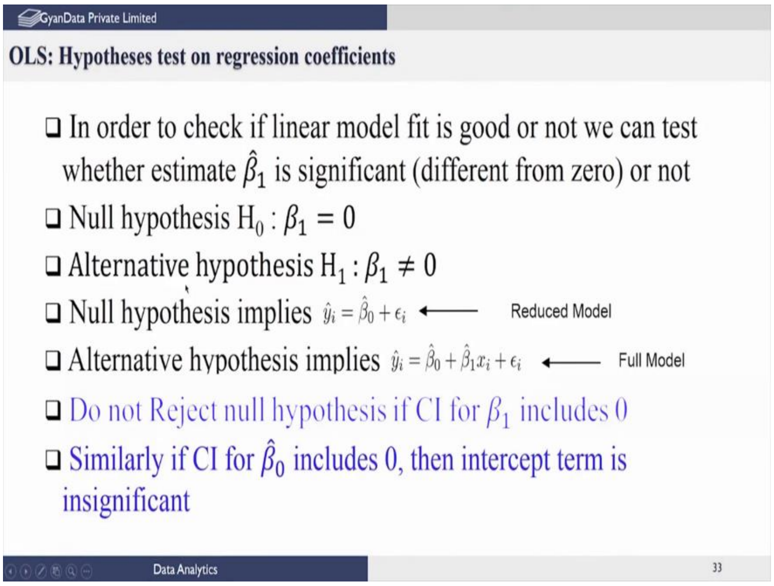
So, we are doing the same thing $\widehat{\beta_0}\pm 2.18$ times; standard deviation of $\widehat{\beta_0}$ estimated from data which is what we call $s_{\widehat{\beta_0}}$ remember

$$s_{\widehat{\beta_0}} = \sigma^2 \sqrt{\frac{\Sigma x_i^2}{n\, S_{xx}}}$$

which is nothing, but the square root of what we have derived in the earlier thing with sigma squared replaced by the estimated quantity; that is all this these two terms represents $s_{\widehat{\beta_0}}\ and\ s_{\widehat{\beta_1}}$.

So, having constructed this 95 percent confidence interval; you can also use it for testing whether β_0 is the unknown $\beta_0=0$ or not or the unknown β_1 is 0 or not which is what we will do.

(Refer Slide Time: 15:29)



So, let us look at why would we want to actually do this hypothesis tests. We have fitted a linear model assuming that you know that the there is a linear dependency between x and y and we have obtained an estimate of $\widehat{\beta_1}$.

Also we have also fitted an intercept term we may want to ask is the intercept term significant maybe the line should pass through 0,0 the origin. Maybe the y variable does not depend on x_1 in a significant manner which means $\widehat{\beta_1}$ is approximately=0 that unknown β_1 is exactly=0. Although we have got some estimate for $\widehat{\beta_1}$ non 0 estimate for $\widehat{\beta_1}$.

So, the null hypothesis what we want to test is $\beta_1=0$ versus the alternative that $\beta_1 \neq 0$. If $\beta_1=0$ it implies that the independent variable x has no effect on the dependent variable, but on the other if we reject this null hypothesis; we are concluding that the independent variable does have some effect on the dependent variable ok.

So, this particular hypothesis test can be also re interpreted as the null hypothesis implies $\beta_1=0$; which means what we are doing is only a fit of yi=a constant, whereas if we accept the or reject the null hypothesis; then we are actually fitting a linear model with β_0 and β_1 present ok. So, the null hypothesis represents the fit of a reduced model which involves only the 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 β_1=0 versus $\beta_1 \neq$0. Remember β_1 can be either positive or negative and that is why we are doing a two sided test.

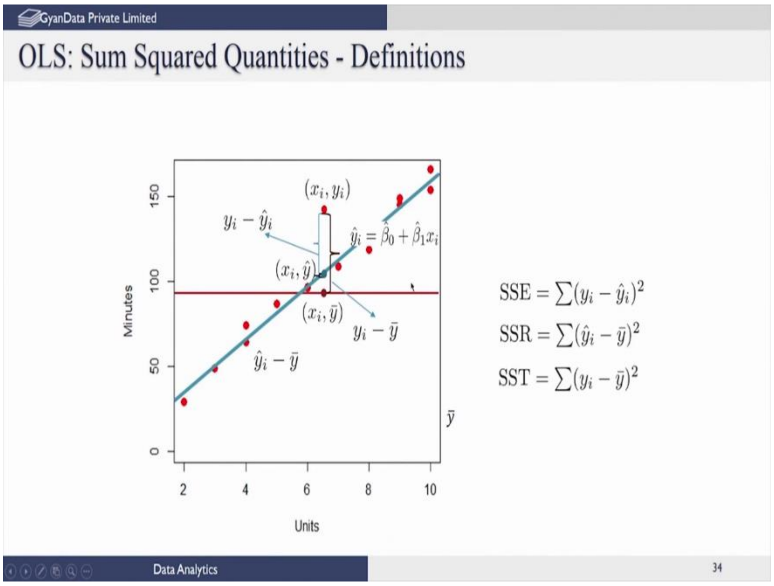
So, we can do it two ways we can actually reject the null hypothesis if the confidence interval for β_1 include 0. So, notice that we have constructed the confidence interval for β_1. So, this this term $\widehat{\beta_1}$-2.18 may be negative and this maybe positive in which case this interval include 0 and then we have to definitely; we might make a decision that that β_1 is insignificant and actually the true β_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 make the conclusion reject the null hypothesis that β_1 equal 0 which means β_1 is significant.

So, from the confidence interval itself; it is possible make the reject or not reject the null hypothesis. So, we can extend this kind of analysis to even test whether β_0 is 0 or 0. So, if the confidence interval for β_0; this particular interval includes 0, then we say that the intercept term is insignificant otherwise we will say that the intercept term should be is significant and should be retained in the model ok.

So, let us actually when we do a final example we; 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 multi linear regression. In the uni variant regression we have only these two parameters but multi linear regression there are several parameters we will have one corresponding to each independent variable and therefore, there will be lot more hypothesis test you will do therefore, we will extend this kind of an argument to test for β_1=0 or $\beta_1 \neq$0 using what is called a F test which we will go through.

(Refer Slide Time: 19:39)



So, before performing this F test to check whether a reduced model is adequate or we should accept a full model; we will use some definitions for some squared quantities. Notice that let us say that we have the 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 14 such data points 14 such salesmen, who have actually reported the data.

So, the red points actually represents the data and the best the linear fit 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 a constant model is good then we would actually fitted this particular horizontal line would be the best fit representing $\bar{y}$. The best estimate of the constant model is the mean of y for all values of x; our prediction best prediction for y_i is the mean value of y_i which means x has no relevance; β_1 is 0, so we will estimate the best constant fit for y_i is this mean value ok.

So, the red line represents the best fit when we ignore β_1 the slope the blue line represents the best fit of the data when we include the slope parameter β_1. Now let us look at certain sum squared deviation the deviation between y_i and $\bar{y}$ which is the red line; best fit of the constant this distance is y_i- $\bar{y}$ 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 y; $(y_i - \bar{y})^2$ all that we have not done is divided by n.

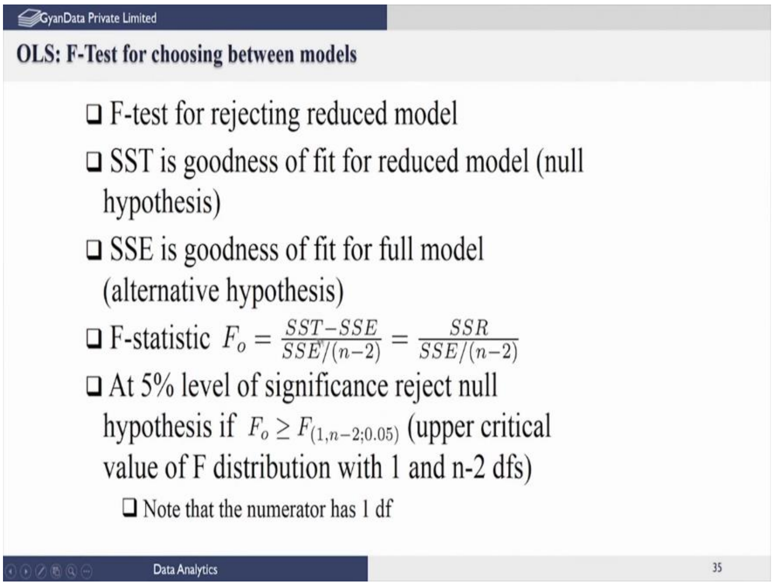
If we had divided by n or n-1 we would have got the variance of y, but this represents the sum squared errors in yi; when we ignore the slope parameter that is another way of looking at it. The distance between y_i and $\hat{y_i}$; so now suppose we assume that the slope parameter is relevant, then we would have fitted this blue line and for every x_i let us take this x_i; y_i corresponding to this independent variable, the predicted value of y_i using this linear model would be the intersection point of this vertical line with the blue line which is represented by the blue dot which is what we call $\hat{y_i}$.

And therefore, this vertical distance between the measured and the predicted value is the sum squared errors is called $SSE=(y_i - \hat{y_i})^2$. And this is the total error if we include the slope parameter in the fit ok. So, the difference between these two quantities SS Total-SS Error will be equal to what is also called the sum squared residuals which is nothing, but the predicted value-the mean value $\bar{y}$; sum squared over all the data points.

Now, we can show that SST will always be greater than SSE because SSE was obtained by fitting two parameters. Therefore, you should be able to reduce the error may be marginally, but you will be always able to be able to reduce the error. So, SS total is the we will always be greater than SSE and therefore, this difference SSR will also be positive; all of these are positive quantities.

Now, one can interpret SS total as the goodness of fit if we assume a constant model; we can interpret SSE as the goodness of fit 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 is 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, SSE represents how good the linear model; if we include the slope parameter.

So, we are asking whether the reduced model should be accepted which is the null hypothesis or should be rejected in favor of this alternative 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 sort of model fit; SSE represents the goodness of fit for the alternative hypothesis fit. This difference if it is large enough as I said, then we can actually say may be it is worthwhile going with the alternative hypothesis rather than the null hypothesis. So, SSR which is the difference between this should be large enough.

So, normalization what; what the denominator represents in some sense a percentage SSE is the error obtained for the alternative hypothesis. Remember because of the difference in the 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 degrees of freedom because we are fitting two parameters; so the difference actually means its only one extra parameter.

So, there is numerator which is SSR has only 1 degree of freedom which is n-1-(n-2) whereas, the denominator SSE has n-2 degrees of freedom because it has two 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 normalization SSR divided by this normalizes the quantity.

And we can show formally that it is an F static because it is a ratio of two squared quantities and each squared quantity is itself a chi squared variable because it is a square of a normal variable. Therefore, this is the ratio of 2 chi squared and we have seen in the hypothesis testing the ratio of 2 chi 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.

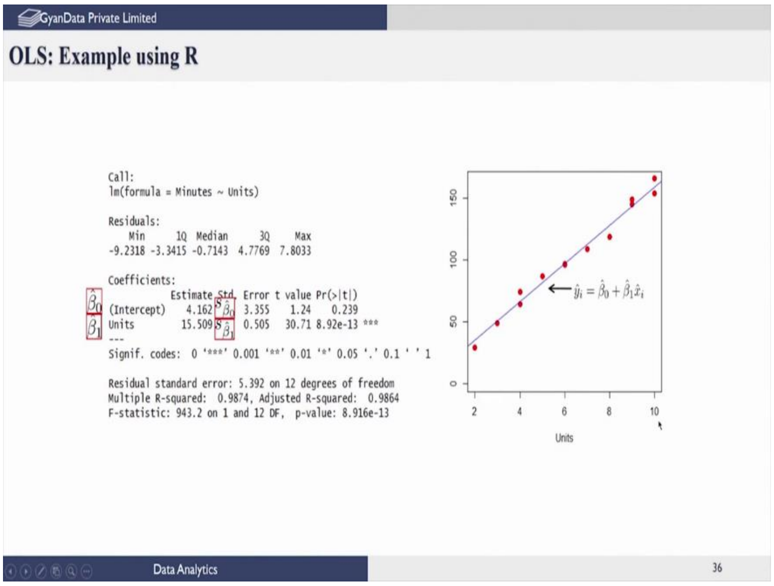
So, if we want to now do a hypothesis test using this statistic F 0 we compare F 0 with the critical value from the F distribution. Notice F is actually a positive quantity, so we do one sided test; if we choose the level of significance as 5; 5 percent; then we choose the upper critical value from the 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 F distribution; we got this threshold and if the statistic exceeds and threshold, then we will reject the null hypothesis and say the full model is better than the reduced mode. We will accept the full mode or we say the; we reject the reduced model in favor of the full model that. The slope parameter is worth including in the model we will get a better fit that is how we actually conclude.

So, now there are several ways for deciding whether the linear model we have fitted is good or not. We could have used the r squared value we said that if it is close to plus one then we should that is one indicator that the linear model may be good it is not sufficient what I call sufficient to conclude but it is good indicator. We can also do the test for βsignificance of β1; if we conclude that β1 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 the 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 fit is good. We cannot stop at this we have to do further tests, but at least these are good initial indicators that we are on the right track.

(Refer Slide Time: 28:39)



So, let us apply this to the example of repair of or the servicing problem, where we have 14 data points and the time taken and the number of units repaired by different salesmen are given.

So, in this case we have these 14 points which we have showed, we have fitted the data using r remember that lm is the function which we should call for fitting a linear model and here we are predicting the dependent variable is minutes and the independent variable is units and once we have fitted this using the R function; it gives out all of this output and it gives you the coefficient, the intercept term turns out to be 4.162; the slope parameter turns out to be 15.501 ok.

But also it also tells you what is the standard deviation; estimated standard deviation of this parameter which is $s_{\widehat{\beta_0}}$ of the intercept. We also tells you what is the standard deviation of this estimate for β1 which turns out to be 0.505 all of this calculated from the data using the formulas we have described. Now, once it has given out we can actually now perhaps construct confidence intervals and find out whether these are significant or not or itself actually tells you something whether these if you run a hypothesis test; whether you can we will conclude whether $\widehat{\beta_0}$ is significant or $\widehat{\beta_1}$ is significant and that is indicated by what is called this P value that it has reported.

So, if you get a very high value t value is represents the statistic which you have again described earlier. So, it has computed the statistic for you for β0 hat and the statistic for testing whether β_1=0 or not and it has computed this statistic value and it has compared with the critical value the distribution t distribution with the 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 we 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 its 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 β1; you will reject the null hypothesis which means you will conclude that β1 is significant even if you choose very low significance value 0.05, 0.01, 0.001 or even lower value. In fact, upto 10^{-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.

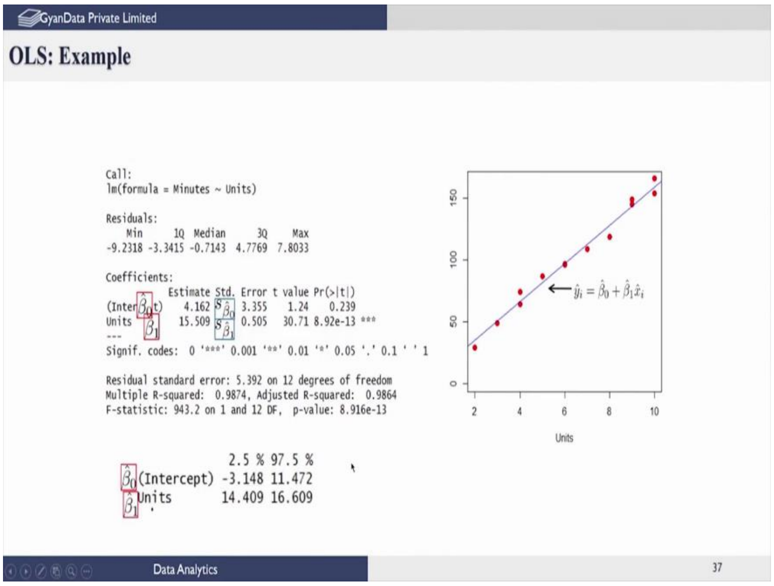
So, therefore you can concluding from these values that β0 hat is insignificant which means $\widehat{\beta_0}$ =0 is a reasonable hypothesis, $\widehat{\beta_1}$ is ≠0 is a reasonable hypothesis. Let us go and see whether this makes sense for this data. We know that if there no units are repaired then clearly no time should be taken by the sales repair person; which means because you have not taken any time for servicing because he has not repaired any units.

So, this line technically should pass through 0,0 and that is what he has said, but; however, we went ahead merrily and fitted an intercept term but the test for hypothesis says you can safely assume β0 the intercept is 0; it makes physical sense also and we could have only fitted β1 that is good enough for this data ok.

So, perhaps you should redo this linear fit with β_0 0 and only using β_1 and the; you will get a slight different solution and you can test again. So, another way of deciding whether the significant whether the slope parameter is significant or not is to look at the F statistic. Notice the F statistic is very high and this p value is very low which means you will reject the null hypothesis that the reduced model is adequate; implying that you should use β1, including β1 is very good you will get a better fit using β1 in your modeling.

So, the high value of test statistic indicates that you reject the null hypothesis or a low value of p value for this F statistic indicates that you reject the null hypothesis even at a very low significance level.

(Refer Slide Time: 33:41)



You can also construct the confidence interval for β0 and β1 and from the earlier thing you say approximately it is estimate ±2.18 times the standard error and that is what is seen 4.1 ±2.18 times 3.35 and that turns out to give that gives the interval confidence interval-3.148 to 11.472; that means, with 95 percent confidence, we can claim that the true β0 lies in this interval.

Similarly, we can construct the interval confidence interval for $\widehat{\beta_1}$ 95 percent confidence interval and it turns out it is 15 ±approximately two times 0.5 which is 14 and 16.6. Now, clearly the interval confidence interval for β_0 includes 0 and therefore, we should not reject the null hypothesis β_0=0 ok.

We should simply accept that β_0 perhaps=0, whereas the interval for confidence interval for β_1 does not include 0; so, we can reject the null hypothesis that β_1=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 lectures.

So, see you in the next lecture.