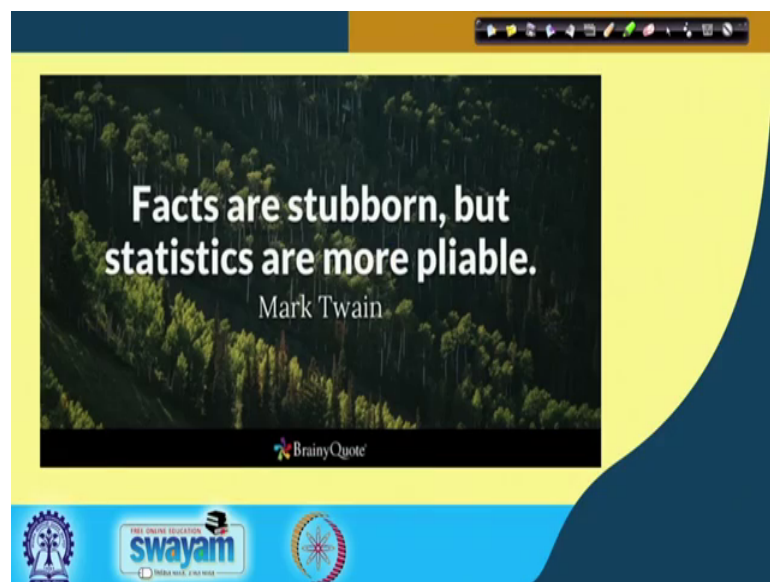


Six Sigma
Prof. Jitesh J Thakkar
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kanpur

Lecture - 33
Regression Analysis: Model Validation

Hello friends, hope you are doing well in your Six Sigma journey and step by step we are advancing in our DMAIC six sigma cycle. So, we are discussing analyze phase of six sigma cycle and as a part of that in the last lecture we had seen the concepts basics related to correlation and Regression Analysis. This lecture this is lecture 33 we will focus on Regression Analysis Model Validation.

(Refer Slide Time: 00:55)

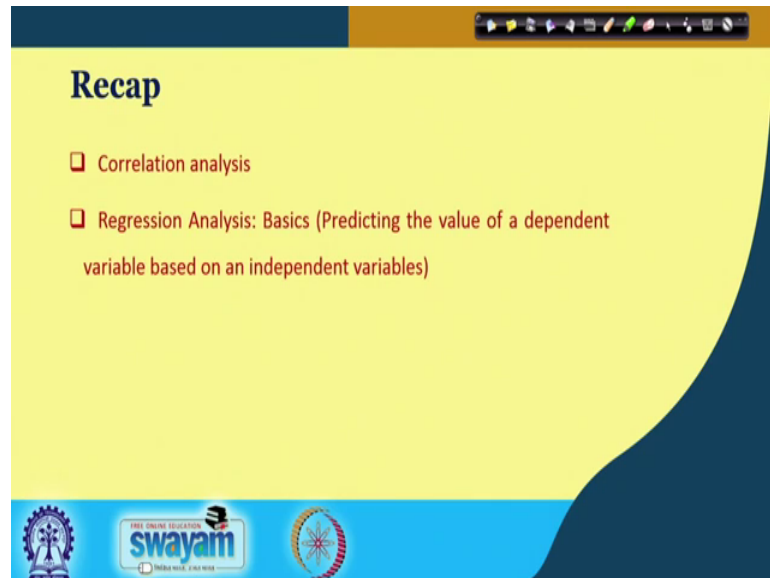


So, if you just see this beautiful quote given by Mark Twain that; 'Facts are stubborn, but statistics are more pliable. So, here we should appreciate that when we want to be scientific in our analysis, pliable means; statistics are flexible. The way you said you are null hypothesis, alternate hypothesis, the investigation that will move your investigation in that particular direction.

And many a times you do not really reveal the meaningful insightful information results because the considerations whatever you have taken into account are not valid like; length of most rack and increase in sales there is no relationship. Like divorce rate and

maybe the sales of the butter there is no relationship. So, facts are stubborn they do not change, but statistics are more pliable and we should be cautious.

(Refer Slide Time: 02:02)



So, with this quote I would just like to give your small recap and introspection that; we discuss the correlation analysis, regression analysis, basics about predicting the value of dependent variable based on set of independent variable. But if you recall I mention that you need to say understand the quality of the regression model and this is about validating my regression model. So, this particular lecture is dedicated on validation of regression model by conducting performing various kind of taste. And to see that to what extent my model is really capable of predicting what it should predict.

(Refer Slide Time: 02:52)

CONCEPTS COVERED

Concepts Covered:

- Validation of Regression Model
- Autocorrelation
- Durbin-Watson statistic
- The t-test and F test
- Pitfalls of regression analysis

The slide features a dark blue background on the left and a yellow background on the right. At the bottom, there are logos for 'swayam' and other educational institutions.

So, I will try to focus on validation of the regression model and as a part of that we will see autocorrelation, Durbin-Watson statistics, t-test f-test, some of the pitfalls of regression analysis. And I would also like to present the simple minitab application for conducting regression analysis.

(Refer Slide Time: 03:17)

Measures of Variation

➤ Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \bar{Y})^2$	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

where:

- \bar{Y} = Mean value of the dependent variable
- Y_i = Observed value of the dependent variable
- \hat{Y}_i = Predicted value of Y for the given X_i value

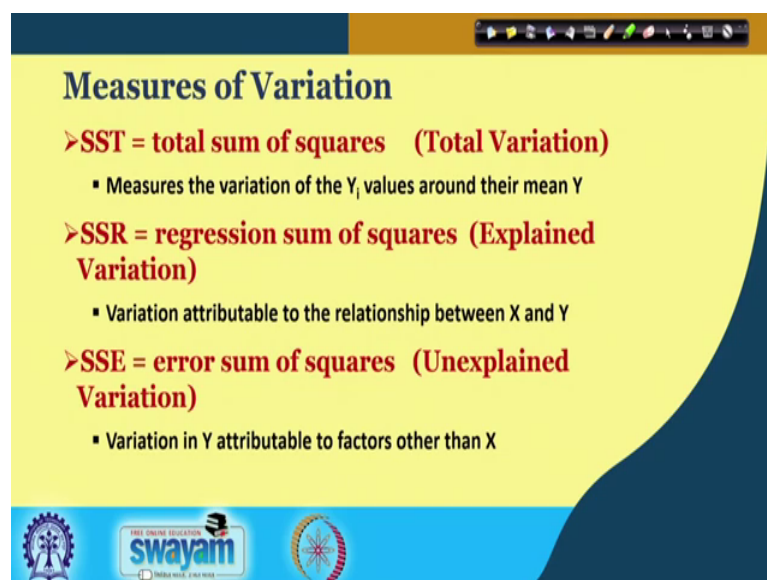
The slide includes a video inset of a presenter in the bottom right corner. The background is yellow with blue and orange accents.

So, just try to appreciate that; when we are trying to minimize the sum of square of error which we did last trying. And it was basically this particular component this is what we try to minimize and this is what we did in the last lecture. So, basically you have SST

this is SST, this is SSR and you have SSE. So, basically your total sum of square if you just decompose then it is SSR that is sum of square of regression and SSE that is error sum of square or sum of square of error and you have \bar{Y} .

So, \bar{Y} is mean value of the dependent variable; very well intuitively you can understand that individual observation minus \bar{Y} mean value of the dependent variable that is the total say some of square variability. Y_i is observed value of the dependent variable and you can see here that Y_i minus \bar{Y} whole square this is typically related to your predicted value \hat{Y}_i minus \bar{Y} . So, this will be the error specific to sum of squares specific to your regression. And when you see this is your SSE sum of square of error. So, this is the expression that typically relates your SSR SSE and SST.

(Refer Slide Time: 05:07)



Measures of Variation

- **SST = total sum of squares (Total Variation)**
 - Measures the variation of the Y_i values around their mean \bar{Y}
- **SSR = regression sum of squares (Explained Variation)**
 - Variation attributable to the relationship between X and Y
- **SSE = error sum of squares (Unexplained Variation)**
 - Variation in Y attributable to factors other than X

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there are logos for 'THE OPEN UNIVERSITY', 'swayam', and the Indian national emblem.

Now, what I said I just summarize here. That SST means total sum of square that is a total variability and measures the variations of the Y_i values around their mean \bar{Y} so that is \bar{Y} . SSR is regression sum of square explained variation and this is variation attributed through the relationship between X and Y how well the relationship between dependent and independent variable is defined.

So, this particular phenomena is captured through SSR. And SSE basically error sum of square it is the unexplained variation. So, variation in Y attributable to factors other than X. So, typically you may consider some of the exogenous factors, external factors,

unknown factors, that may have impact on your regression model which you are not able to actually capture and this is your SSE.

(Refer Slide Time: 06:12)

Coefficient of Determination, r^2

- Coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.
- It is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Note: $0 \leq r^2 \leq 1$

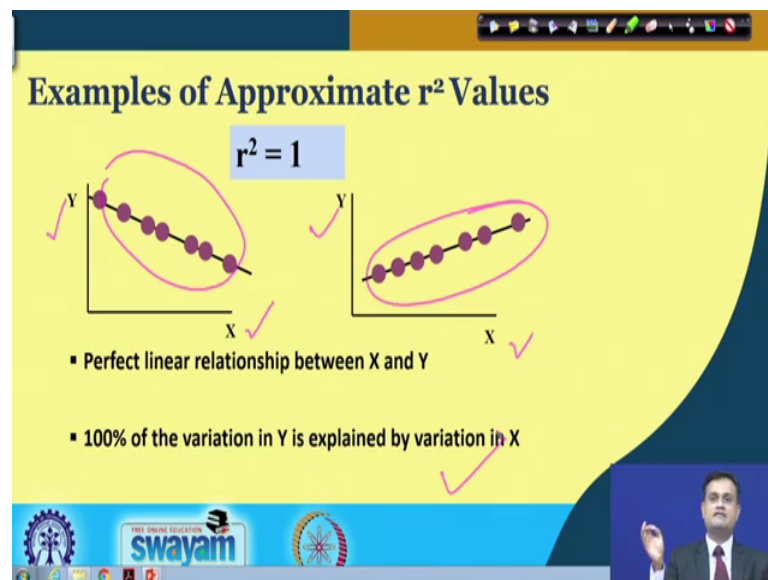
So, this r square basically we have seen that coefficient of determination r square is a very very important measure and this is basically the ratio of your SSR and SST. So, regression sum of square and total sum of square so; this particular measure is the very very important measure. And this measure is basically used to explain the proportion of the total variation in the dependent variable your left side variable. That is explained by the variation in the independent variable.

So, again I would just give you the example to make it clear; that I want to judge, I want to predict that what are the factors that contribute towards the productivity of my manufacturing process. So, let us say productivity is my dependent variable and it depends on couple of independent variables like; process capability, my training and skill of the operator, state of the art technology, management policy.

Now, where I am trying to figure out the relationship between productivity and this other independent factors. Then I want to see that how well the portion of the total variability in the dependent variable here it is productivity for the manufacturing process is explained by the independent variables like worker and training skill, say training skills, and the worker skills or the management policy, maintenance schedule, state of the art technology and so on.

So, I have r^2 is equal to SSR divided by SST and typically this should fall within the range of say 0 to 1. And this term is very important to understand that to what extent my regression model is capable enough to capture the variability. And this variability should be explained by the various independent variables in the regression model.

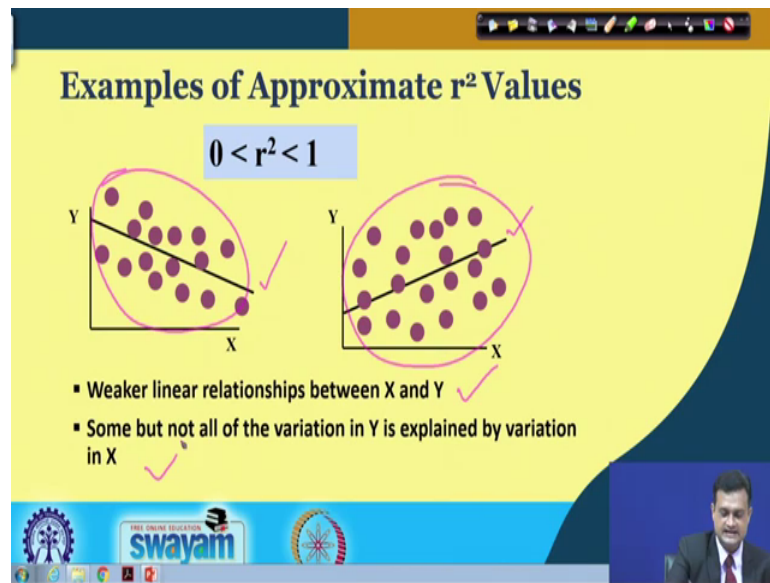
(Refer Slide Time: 08:37)



So, with this little understanding just try to see because visual say display can help us better. So, you can see that approximate r^2 values and you are trying to say see the relationship between X and Y, X and Y. And what you can see that there is a perfect linear relationship whether positive or negative, but there is a perfect linear relationship 100 percent of the variation in Y is explained by variation in X.

So, in my words I will say whatever independent factors we have included in your model they are 100 percent capable enough to explain the total variability in the Y. And there are no the factor or exogenous factor which really have an impact on the quality or prediction ability of my regression model.

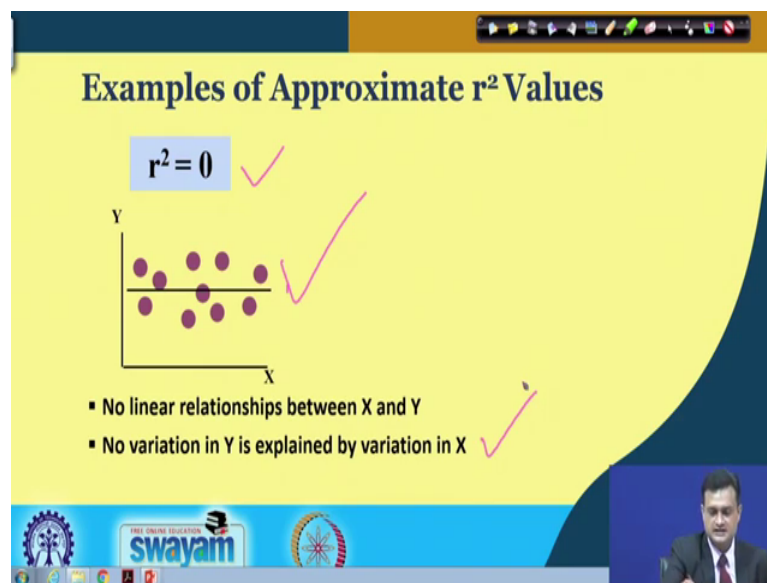
(Refer Slide Time: 09:41)



So, this is a very important term and we must try to you appreciate this term in the validation of the regression model. Now, you just see here this is the situation where you have the scatter plot you have the scatter plot, you have a line, explaining the relationship. Now in both the case you will see that with linear relationship.

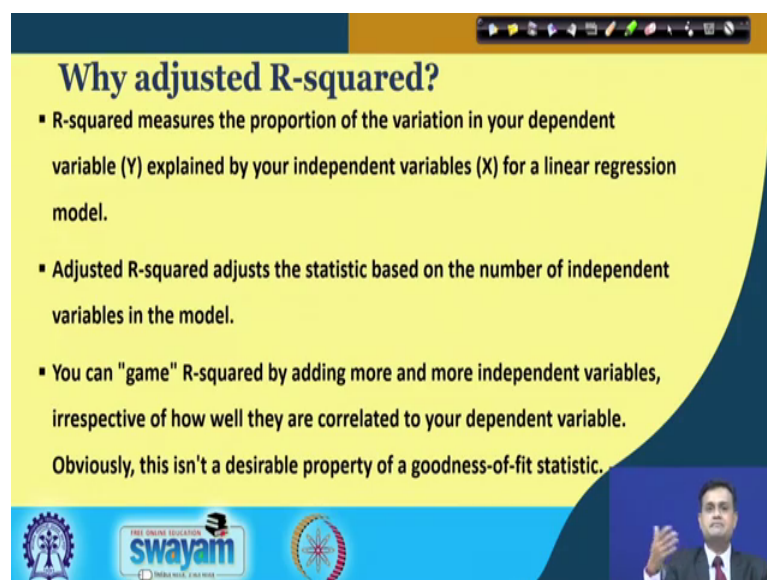
So, I would say that some, but not all variation in Y is explained by variation in X. So, either you have not considered adequate number of independent variable which can explain the variability in Y that is your dependent variable. Or there are some exogenous factors which you are not aware of affecting your model. And hence your variability is very high and the relationship is weak.

(Refer Slide Time: 10:45)



So, this is the importance of r square in judging the quality of my regression model. And this is the case where you can see that your r square is equal to 0 so no linear relationship between X and Y. You cannot predict the Y using some independent variable X or X_i and no variation in Y is explained by X. So, this is the third case where no variation in Y is explained by X.

(Refer Slide Time: 11:14)



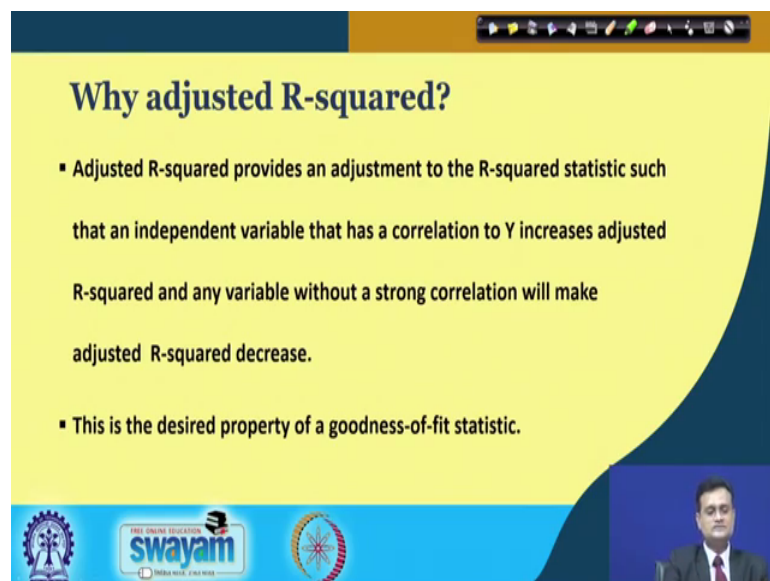
Now, there is another term I am just trying to familiarize you with some fundamental concepts in validating the regression model checking the quality of my regression model.

There is another term which is called adjusted R square. So, many a times what happens; that an analyst in its a desire to improve the value of R square may keep adding many independent variables regardless of their relevance to the model. And by doing such kind of manipulation you may end up with very high R square value.

So, there is a way and there is a there is an another measure which is called adjusted R square. So, this typical measure adjusted R square is basically use to check even the truth explained by the R square and this particular say measure adjust the statistics based on the number of independent variables in the model. So, I would say you can game R squared by adding more and more independent variables irrespective of the relevance or how well they are correlated to your dependent variable. Obviously, this is not a desirable property of a goodness of fit statistics.

So, when you are trying to check the quality of your regression model you will not only see the R square value you will also see the R adjusted R square. And many a times say if the adjusted R square is very low then you need to revisit your assumption about the independent variables or the relevance to the dependent variable may modify your model. And see that the r square value and adjust adjusted R square value both should be flow a favorable adequate enough to express the goodness of fit of my regression model.

(Refer Slide Time: 13:20)



Why adjusted R-squared?

- Adjusted R-squared provides an adjustment to the R-squared statistic such that an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease.
- This is the desired property of a goodness-of-fit statistic.

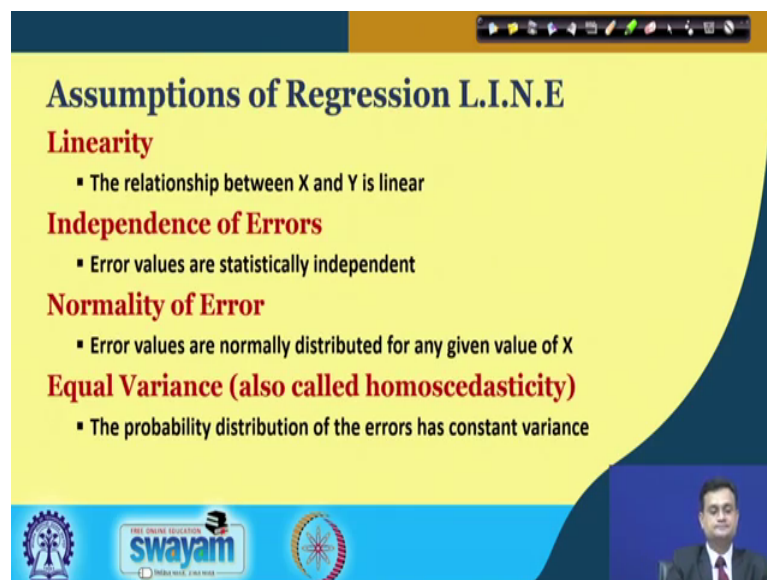
swayam

So, as I mentioned that adjusted r square provides an adjustment to the R squared statistic such that; an independent variable that has a correlation with Y increases

adjusted R square. And any variable that is a strong correlation will make R adjusted R square decrease. So, this is where say I would like to make the adjustment to my R square. And see that what is the real R square and simply I will not get governed by the value of R square.

So, this is a property of goodness of fit to the test. And if you have a relations, if you have a variable without a strong correlation then; it will decrease the adjusted R square. So, there is a trade off if your dependent variable has a good relationship with the independent variable it will increase the R square it means; adjusted R square if not it will decrease. So, this trade off is taken care by adjusted R square and that will help you to even verify your R square value.

(Refer Slide Time: 14:34)



Assumptions of Regression L.I.N.E

- Linearity**
 - The relationship between X and Y is linear
- Independence of Errors**
 - Error values are statistically independent
- Normality of Error**
 - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)**
 - The probability distribution of the errors has constant variance

THE OPEN UNIVERSITY
swayam
INDIAN INSTITUTE OF TECHNOLOGY

So, now we talk about the model validation of regression model. And we must check the regression model against the fundamental assumptions and these assumptions are called line assumptions. Line assumptions tends for line l stands for linearity the relationship between X and Y is linear. If you have say non-linear relationship then the kind of model linear regression you are using will not really have the capability to predict what you are trying to say investigate or predict through this analysis.

Independence of errors; so errors where statistically independent it means there should not be any hidden phenomena because of which there is some kind of relationship or dependence within the error. And if my errors are independent then this is the second

assumption which is satisfied that is the independence of error as a part of line. The third one is normality of error that error values must be normally distributed for any given value of X. And finally, equivalence of variance also called homoscedasticity and this is the probability distribution of the error that has the constant variance. So, this four assumptions must be checked, must be verified, if I have to accept my regression model has a good predictor for the phenomena under consideration.

(Refer Slide Time: 16:18)

Residual Analysis

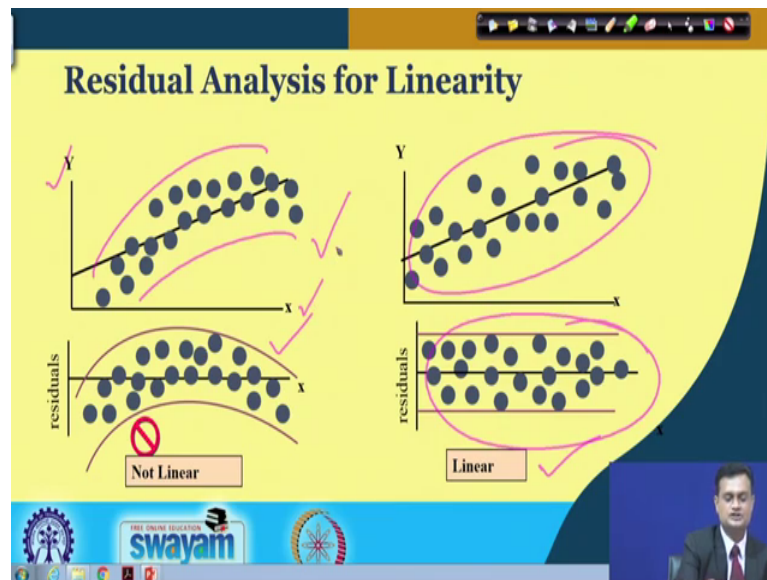
$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i , e_i , is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
 - Examine for linearity assumption
 - Evaluate independence assumption
 - Evaluate normal distribution assumption
 - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals
 - Can plot residuals vs. X

Logos: Swamyam, Free Online Education, and a university emblem.

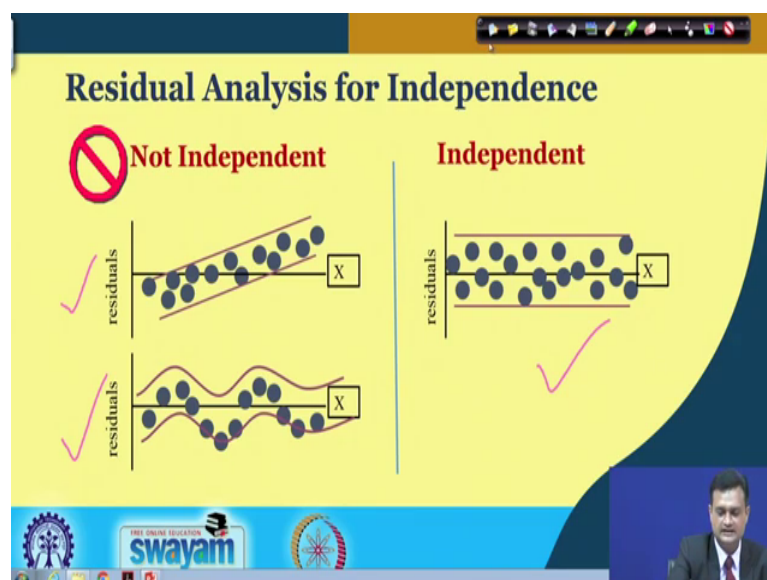
So, you see that how can I compute the residue? So, it is very simple there is an observed value there is a predicted value when I take the difference between these two observed and predicted value I get the error component for a particular value of i so Y_i . Now we want to check this assumptions as I mentioned these are the line assumptions. And we can first do the graphical analysis of the residue to immediately get an idea that to what extend my line assumptions are valid or they are satisfied. So, just see that how we can interpret the line assumptions.

(Refer Slide Time: 17:08)



This is my residual analysis for linearity and this is the first one where I can see some pattern. So, this pattern is depicted here and this is the curvature so not linear. But if I see this scatter plot and if I see this then I can say that linearity is present. So, when I plot my residual with respect to say my observed value or X_i and when I plot my X Y value with respect to X and see the plot I get an idea about the linearity assumption.

(Refer Slide Time: 18:03)



The second one is independence inline. So, you can see here very well that; you can see here that the first case these are not independent. And here also you can see that it form

some pattern and there is ups and downs in my error component residual component. And these are independent because there is no pattern available this is just the random variation. And my independence assumption can be visually verified.

(Refer Slide Time: 18:56)

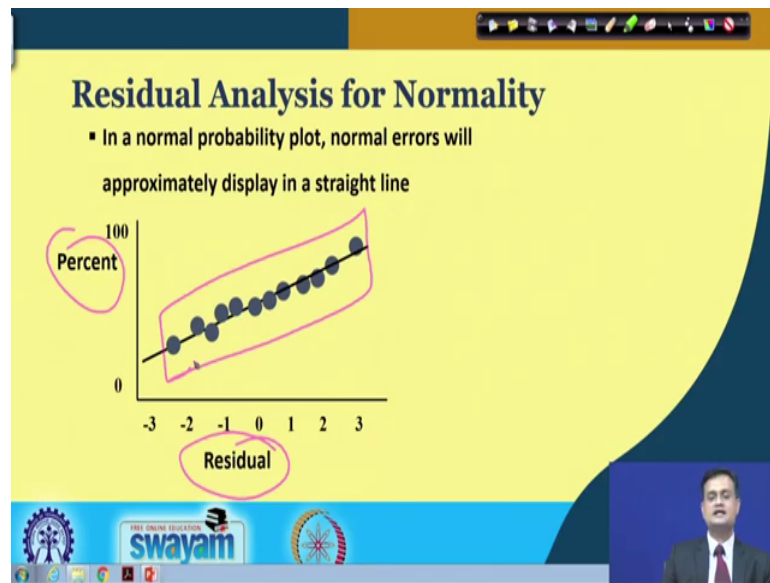
Checking for Normality

- Examine the Stem-and-Leaf Display of the Residuals ✓
- Examine the Boxplot of the Residuals ✓
- Examine the Histogram of the Residuals ✓
- Construct a Normal Probability Plot of the Residuals ✓

swayam

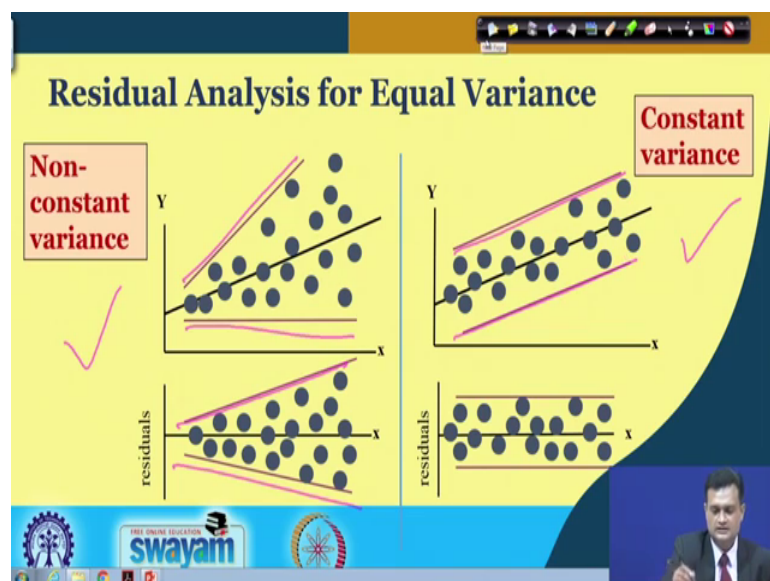
Now, the third one which is very very important is checking for normality. You have various options available that you can examine the stem and leaf display, you can examine the box plot of the residual, you can examine the histogram, and most widely used approach is the normal probability plot of the residual. So, these are the various approaches for three we have already seen; e 1 in process capability analysis non normality plot we have plotted for the normal and non normal data to check that whether my data set is normal or not. So, this methods can equally be applied for checking my normality assumption for regression model.

(Refer Slide Time: 19:43)



And you can see here that the normality plot when you put or plot the values of residual with respect to percentage. So, basically it is a classic classically designed normality plot where you try to plot the values of residual with respect to say your z value probability value. And if this points are covered by a pencil or pen then majority of the points you will say that normality is present. So, this is the approach which is widely use to check the normality.

(Refer Slide Time: 20:29)



And the final one of line that is the e and equivalence of variance. So, if you see the first case then non constant variance and you have the constant variance. Here you will find something like a funnel approach for residual again you will find a funnel approach. Here it is a constant variance over a particular range the variance is not changing significantly it is just random phenomena and I have the idea that whether my equivalence of variance is satisfied or not. So, this is the how I can check my line assumption. And when you are using excel or minitab your software we will generate the different display that can help you to understand that whether your line assumptions are satisfied or not.

(Refer Slide Time: 21:21)

Measuring Autocorrelation

- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period

swayam
INDIA RISE, TIME RISE

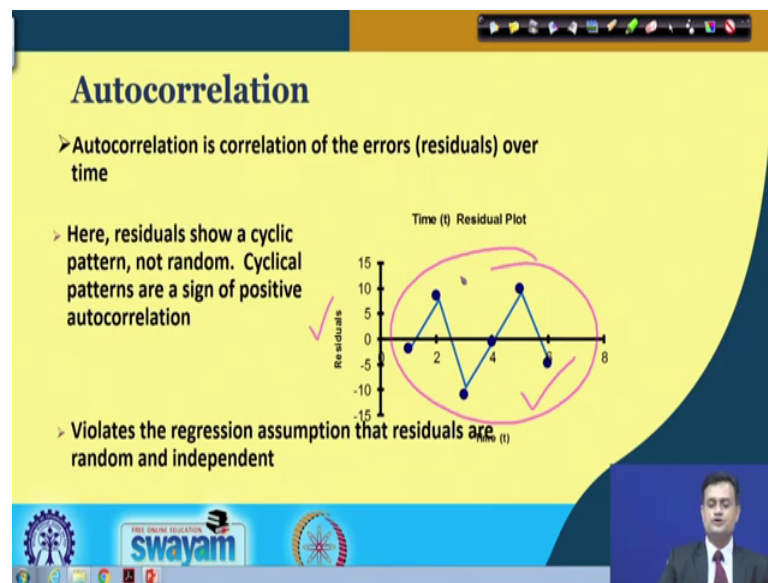
Now, let us try to go into detail because whatever I presented it is mainly the visual say observation about my line assumption, but I also want to do some inferential analysis to check the prediction ability of my regression model or in the simple word quality of my regression model. So, the first one is to measuring autocorrelation. So, this is particularly the concept or autocorrelation is used when data are collected over a period of time to detect if autocorrelation is present.

Is there any correlation among the data is present maybe because of some factor and if autocorrelation exist. If residual in one time period are related to residual in another time period then we can say that autocorrelation exist. And this is not desirable so far the regression model is developed based on this kind of data set. So, you can just think about

a situation that there is a continuous wear and tear because of some miss adjustment in your process.

And whatever data you are getting which you are using for predicting or developing the regression model is some way auto correlated. And if your data set used for developing the regression model itself is some kind of biasness phenomena prevailing then it will have an impact on the quality of your regression model.

(Refer Slide Time: 23:14)



So, I would like to check this that whether autocorrelation is present or not. So, you can just see that when I am plotting the residual with respect to time I can see that I can just a little bit shift this particular graph above. And you can see that this fits like this just you can put it like this would be better I think and you can get an idea that; yes, there is some cyclic pattern which is present. So, here your residual they capture this the pattern when this residual are plotted over a period of time.

And they show the cyclic pattern not random and there is a sign of positive autocorrelation. So, this violates the regression assumption that residuals are say random and independent. So, previously we have checked tested this assumption simply by having the graphical say display. Now I am trying to say go I am trying to see also the an alternate approach inferential approach to be more accurate about my say assumption. So, the first thing that you plot and just a visual display can help you to appreciate whether there is any kind of autocorrelation prevails or not.

(Refer Slide Time: 25:01)

The Durbin-Watson Statistic

➤ The Durbin-Watson statistic is used to test for autocorrelation

H0: residuals are not correlated ✓

H1: positive autocorrelation is present ✓

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

➤ The possible range is $0 \leq D \leq 4$ ||

➤ D should be close to 2 if H_0 is true ✓

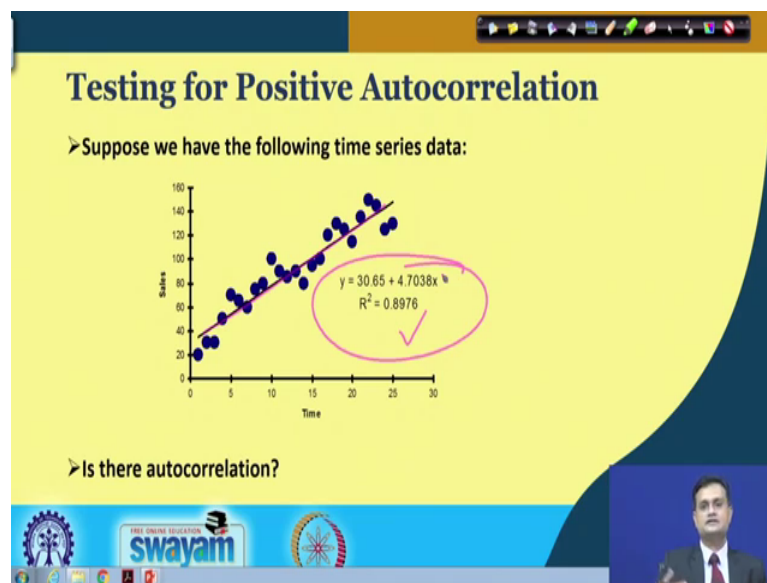
➤ D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation ✓

swayam

Now, if we going to the inferential testing of autocorrelation then there is a test called Durbin-Watson test. And this test basically like all other taste we must have a test statistics and we must set initially the null and alternate hypothesis. So, this particular taste use to infer about the auto correlation among the data says that; residuals are not correlated H 1 there is a positive autocorrelation present. So, there is a Durbin-Watson statistics which is simple to interpret you have e_i values for each particular say $X_i Y_i$ you have e_i . And you have $e_i - 1$ that is the previous value and you have e_i square.

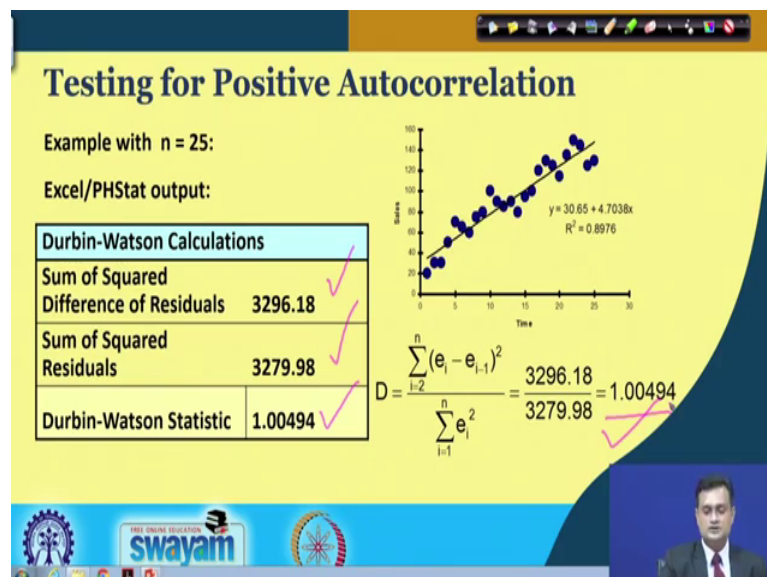
So, the possible range for this is $0 \leq D \leq 4$ and D should be close to 2 if H_0 is true. So, these are the standard guidelines develop and if D is less than two may signal positive autocorrelation. And D greater than 2 may signal negative autocorrelation. So, we accept this standard guideline to infer about the autocorrelation present in my error data. And I would like to statistically infer that whether there is an autocorrelation present or not.

(Refer Slide Time: 26:39)



So, this way I can just move ahead. Now let us say you have a regression model like this; and this is typically say express like Y is equal to 30.65 that is the intercept plus 4.7038 x and my R square is 0.8976. So, is there autocorrelation or this is just my say doubt suspicion I want to check it statistically.

(Refer Slide Time: 27:12)



Now, just see I computed the Durbin-Watson statistics using excel you can do it in Minitab also. So, I have or I can computed manually also Durbin-Watson statistics is

1.00494 this is the sum of squared difference of residual sum of squared residuals and this is 1.00494. So, if you refer our guidelines then what does it tell?

(Refer Slide Time: 27:41)

Testing for Positive Autocorrelation

- Here, $n = 25$ and there is $k = 1$ one independent variable
- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$
- $D = 1.00494 < d_L = 1.29$, so reject H_0 and conclude that significant positive autocorrelation exists

Decision: reject H_0 since
 $D = 1.00494 < d_L$

Reject H_0 | Inconclusive | Do not reject H_0

$d_L = 1.29$ | $d_U = 1.45$

The slide includes a horizontal axis with three regions: 'Reject H_0 ' (left), 'Inconclusive' (middle), and 'Do not reject H_0 ' (right). The boundary between 'Reject H_0 ' and 'Inconclusive' is marked at $d_L = 1.29$, and the boundary between 'Inconclusive' and 'Do not reject H_0 ' is marked at $d_U = 1.45$. A box above the axis shows the decision logic: 'Decision: reject H_0 since $D = 1.00494 < d_L$ '. The slide also features a 'swayam' logo and a small video inset of a presenter in the bottom right corner.

So, I have the situation like this n is equal to 25 k is equal to 1 Durbin-Watson table. So, like z table t table you have the Durbin-Watson table that you will find in the suggested text books. You can find d_L and d_U and you can set the rule like this. That if your computed value of Durbin-Watson statistics is less than d_L then you reject null hypothesis.

If it is in between you are inconclusive you cannot comment rather autocorrelation is present or not. And here you do not reject null hypothesis so here the kind of data set I have assume I fall into the region which is less than d_L and I will say reject null hypothesis it means there is some autocorrelation present.

(Refer Slide Time: 28:38)

Inferences About the Slope

The standard error of the regression slope coefficient (b_1) is estimated by

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

where:

S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

The slide includes a Swamyam logo and a small video inset of a presenter in the bottom right corner.

Now, you can also make statistical inference about the slope. So, you have you have the S_{b_1} , SSX divided by square root of SSX . And this particular expression can help us to compute the S_{b_1} . So, S_{b_1} is estimate of the standard error of the slope and S_{YX} sum of square that is square root of SSE sum of square of error divided by n minus 2.

(Refer Slide Time: 29:20)

Inferences About the Slope: t Test

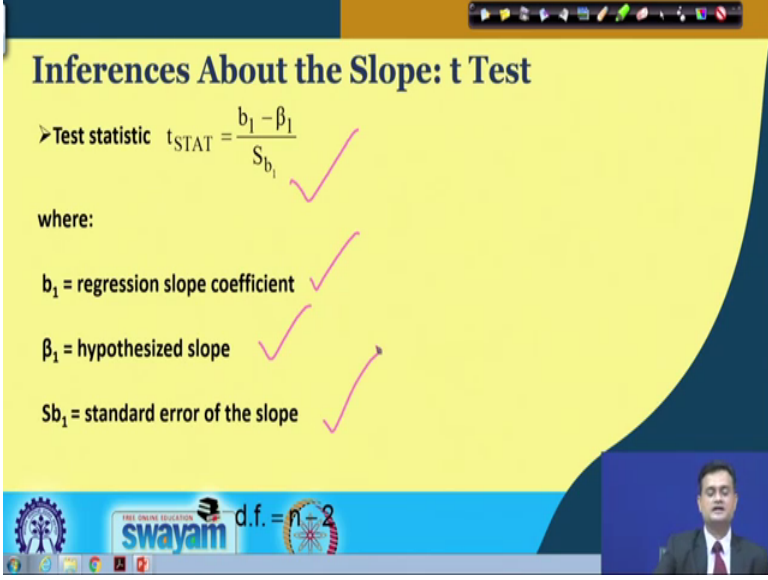
- t test for a population slope
- Is there a linear relationship between X and Y?
- Null and alternative hypotheses
 - $H_0 : \beta_1 = 0$ (no linear relationship)
 - $H_1 : \beta_1 \neq 0$ (linear relationship does exist)

The slide includes a Swamyam logo, the text "d.f. = n - 2", and a small video inset of a presenter in the bottom right corner.

Now, when you do this you can compute your inferential statistics which is the t test for checking the relevance of the slope that is beta 1 for your regression model. So, please remember that we want to compute beta 0 we want to check the validity of my regression

model which is dependent on the beta 0, beta 1, beta 2, beta 3. I want to see that whether my slope is statistically significant or not. So, I would set the null hypothesis beta 1 is equal to 0, beta 1 is not equal to 0. So, no linear relationship exist and I would say linear relationship does exist.

(Refer Slide Time: 30:13)



Inferences About the Slope: t Test

➤ Test statistic $t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$

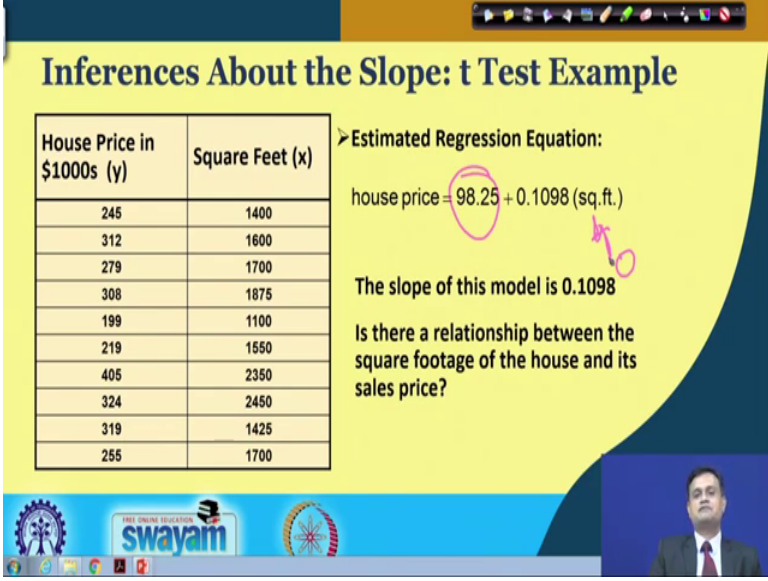
where:

- b_1 = regression slope coefficient
- β_1 = hypothesized slope
- S_{b_1} = standard error of the slope

d.f. = n - 2

So, this is my premise of testing the beta 1. And I can set the test t STAT b 1 minus beta 1 divided by S b 1. So, b 1 I would refer as regression slope coefficient beta 1 hypothesized slope and S b 1 is the standard error of the slope.

(Refer Slide Time: 30:41)



Inferences About the Slope: t Test Example

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

➤ Estimated Regression Equation:

house price = 98.25 + 0.1098 (sq.ft.)

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

So, with this particular test statistics t I can verify my assumption or my understanding about the β_1 let us try to take an example. So, the example goes like this you have the house prices in dollar maybe 1000 and you have this square feet. I am only considering two one is the independent variable that is my square feet and the other is the dependent variable that is the house price.

Now my equation regression equation comes out to be house price 98.25 plus 0.1098 square feet. This says that these two are related like this and I can predict the house price based on the square feet. There could be many factors which I have not considered here just to explain the pitfalls or some of the say errors that you may cause in conducting regression analysis.

Now you just here just by observation I can say that if I put square feet is equal to 0 even though I will have to pay some 28.25 1000 rupees for not purchasing the house. Square foot 0 means I am not purchasing the house; who will accept such a model? I am paying some lakhs of rupees to my builder some builder for not purchasing the price. So, β_0 my intercept has no value here and this can even be validated tested though inferential statistics.

(Refer Slide Time: 32:18)

Inferences About the Slope: t Test Example

$H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$

From Excel output:

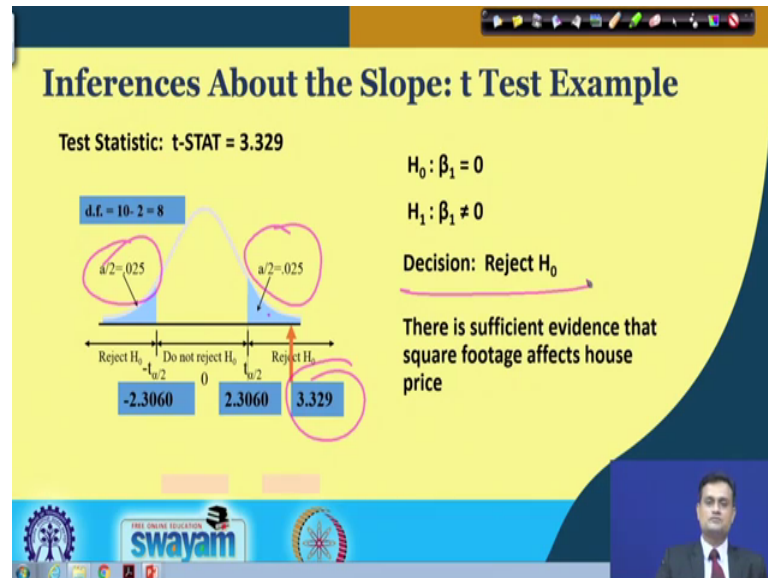
	Coefficients	Standard Error	t Stat	P-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

b_1 S_{b_1} $t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$

So, just try to see that I can compute this β_1 . So, this is the excel output you can do it in minitab also this is my S_{b_1} and I can compute the t statistics β_1 on my b_1 minus

beta 1 divided by S b 1 this comes out to be 3.329 and I can make the inferences about my hypothesis beta 1 is equal to 0 beta 1 is not equal to 0.

(Refer Slide Time: 32:43)



So, what I would say that the alpha by 2 alpha by 2 is 0.025. And I have the computed value 3.329 which falls within the rejection region. So, decision is reject null hypothesis. This says that there is sufficient evidence that square foot effects the house price. It means my particular assumption about the relationship between square feet and the house price is true. And the inclusion of this independent variable will help me to explain the total variability present in predicting the dependent variable that is Y.

(Refer Slide Time: 33:40)

Inferences About the Slope: t Test Example

From Excel output:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

	Coefficients	Standard Error	t Stat	P-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

➤ **Decision: Reject H_0 , since p-value < α**

There is sufficient evidence that square footage affects house price.

p-value

So, now let us move head and I can also check the p value. So, again you can see that p value is less than alpha and I can reach to the same conclusion.

(Refer Slide Time: 33:49)

F Test for Significance

➤ F Test statistic:

$$F_{STAT} = \frac{MSR}{MSE}$$

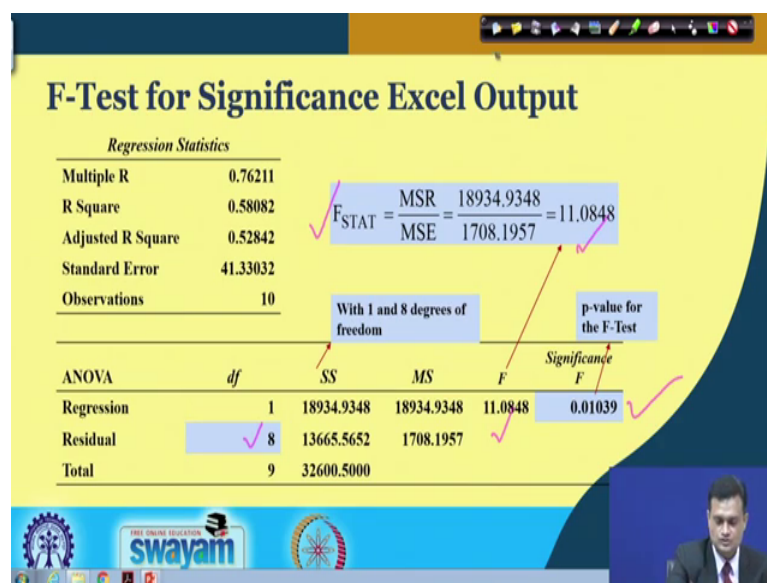
where

$$MSR = \frac{SSR}{k}$$
$$MSE = \frac{SSE}{n - k - 1}$$

- ✓ Where F-STAT follows an F distribution with k numerator and (n - k - 1) denominator degrees of freedom
- ✓ (k = the number of independent variables in the regression model)

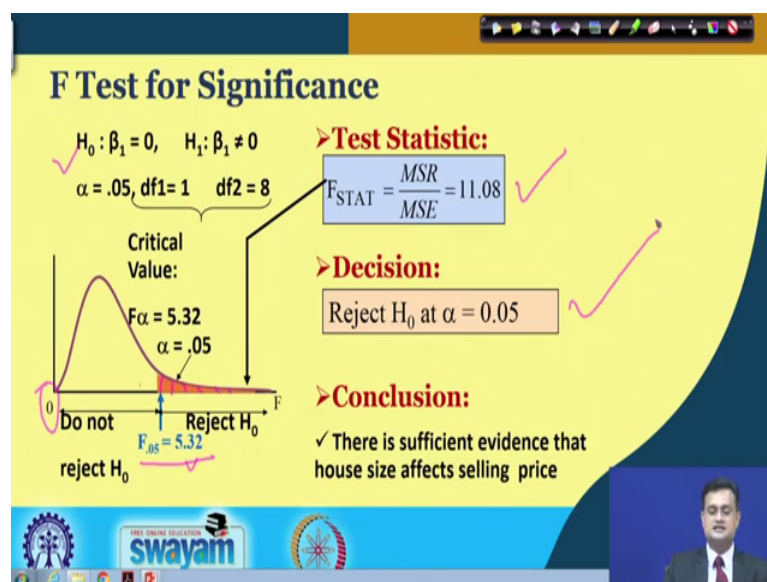
Now, you can also conduct the F test for significance. And as you know that F test is basically MSR divided by MSE. So, mean square of regression mean square of error and when I compute this SSR divided by degree of freedom k SSE divided by n minus k minus 1. So, f stat follows an F distribution we have studied in case the number of independent variables in the regression model.

(Refer Slide Time: 34:25)



So, I can also do the inferential analysis about the quality of my regression model. And what I am trying to do here I have completed the F STAT. And I have let us say residual degree of freedom 8, I have p value computed for f test is point 0.01039 and I do have the value F 11.0848.

(Refer Slide Time: 35:00)



So, with this particular result what I am trying to investigate is that whether my assumption about relevance of the beta 1 is true or not. So, H_0 says beta one is equal to 0 alternate says beta 1 is not equal to 0 for the computed example. My $F_{0.05}$ is 5.32

which is the critical value. And this is my region of rejection and my computed value F STAT is 11.08.

So, I am counting from 0 onwards. So, this much higher and it falls much higher than 5.32 and it falls in the rejection region. So, your conclusion is that reject null hypothesis at alpha is equal to 0.05 and there is a sufficient evidence that house size effects the selling price.

(Refer Slide Time: 35:57)

Confidence Interval Estimate for the Slope

➤ **Confidence Interval Estimate of the Slope:**

$$b_1 \pm t_{\alpha/2} S_{b_1} \quad \text{d.f.} = n - 2$$

Excel Printout for House Prices:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.5772	232.0738
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

The slide includes a Swayam logo and a small video inset of a presenter in the bottom right corner.

So, you can check it through t test you can check it through f test and you can better make the inferences. You can also said the confidence interval and we have three different approaches through test hypothesis you can also apply the confidence interval approach you can compute at lower 95 upper 95. And this is the confidence interval you create and you can check that whether you are beta one or b one is falling in this range or not.

(Refer Slide Time: 36:25)

Confidence Interval Estimate for the Slope

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is \$1000s, we are 95% confident that the average impact on sales price is between \$33.74 and \$185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

So, this is what is about confidence interval we have already done.

(Refer Slide Time: 36:28)

t Test for a Correlation Coefficient

➤ **Hypotheses**

$H_0 : \rho = 0$ (no correlation between X and Y)

$H_1 : \rho \neq 0$ (correlation exists)

➤ **Test statistic** (with $n - 2$ degrees of freedom)

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

where

$r = +\sqrt{r^2}$ if $b_1 > 0$

$r = -\sqrt{r^2}$ if $b_1 < 0$

Now, you can have t test for correlation coefficient. So, you have rho is equal to 0 as null hypothesis rho 1 is equal to 0. And you can set the t statistic as r minus rho. So, r is basically related to your sample rho is related to your population and this is the value of standard deviation for conducting this test. So, if you have let us say nth minus 2 degree of freedom. And r is basically plus square root of r square if b 1 is greater than 0 it would be minus square root of r square if b 1 is less than 0.

(Refer Slide Time: 37:21)

t-test For a Correlation Coefficient

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

✓ $H_0 : \rho = 0$ (No correlation)
✓ $H_1 : \rho \neq 0$ (correlation exists)
✓ $\alpha = .05$, $df = 10 - 2 = 8$ ✓

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$

The slide includes a Swayam logo and a small video inset of a presenter in the bottom right corner.

So, for this particular kind of approach that to check the correlation coefficient I can have the hypothesis rho is equal to 0 rho is not equal to 0. I would be interested to check it at alpha is equal to 0.05 degree of freedom 10 minus 2, 8 my computed statistics is 3.329.

(Refer Slide Time: 37:38)

t-test For A Correlation Coefficient

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$

d.f. = 10 - 2 = 8

$\alpha/2 = .025$

Reject H_0 | Do not reject H_0 | Reject H_0

-2.3060 | 2.3060

3.329

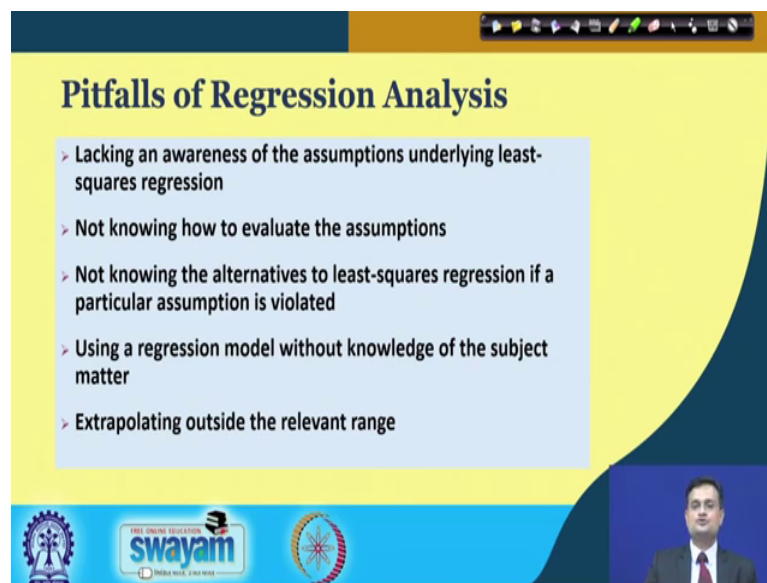
Decision: Reject H_0

Conclusion: There is evidence of a linear association at the 5% level of significance

The slide includes a Swayam logo and a small video inset of a presenter in the bottom right corner.

Now, if I just go further than this 3.29 329 it basically falls in the reject region this is my t alpha by 2 which is the critical value from the table. So, my decision is that reject null hypothesis there is an evidence of linear association at 5 percent level of significant.

(Refer Slide Time: 38:02)



Pitfalls of Regression Analysis

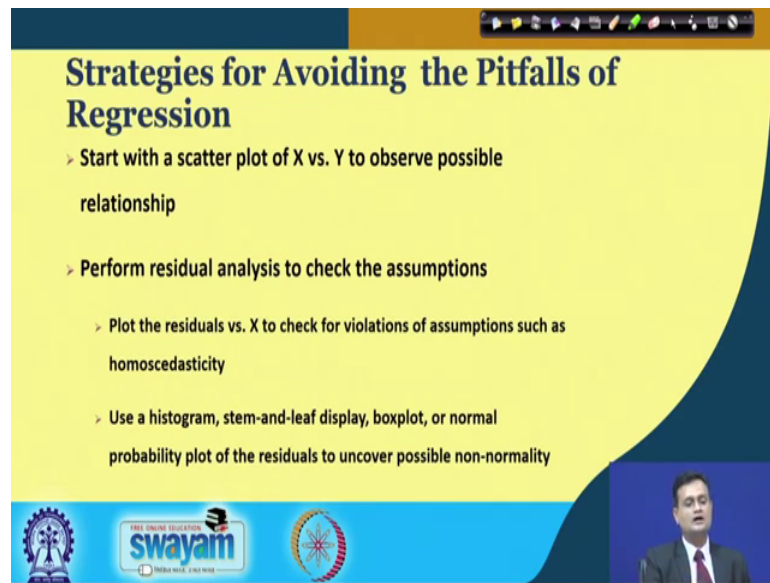
- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range

The slide is part of a presentation, as evidenced by the navigation icons at the top and the presenter's video feed in the bottom right corner. The bottom of the slide features logos for 'swayam' and other educational institutions.

So, this is what I can do I would just like to sum up with couple of pitfalls in regression analysis. So, looking and awareness of the assumption underlying least square regression. So, this is something lacking that we are trying to do it under least square say method. So, this awareness is one of the part that I we are trying to minimize the sum of square of error. Not knowing how to evaluate the assumptions many a times you just feel happy with the outcome of your software. But you do not really go into the detail of checking the line assumption and this is where the prediction ability quality of your model is at stake.

Not knowing the alternatives to least square variegation if a particular assumption is violated. And we have discussed that what are the various ways you can better make yourself clear about the assumptions. Using a regression model without the knowledge of the subject matter. So, like length of the most rack and the increase in sales; you do not exactly know that what are the factors that can affect the sales or most rack. And you just try to manipulate with the situation and remember you cannot set the regression model for a universal range. So, exploding be beyond a particular range will again have a negative effect on your prediction ability.

(Refer Slide Time: 39:34)



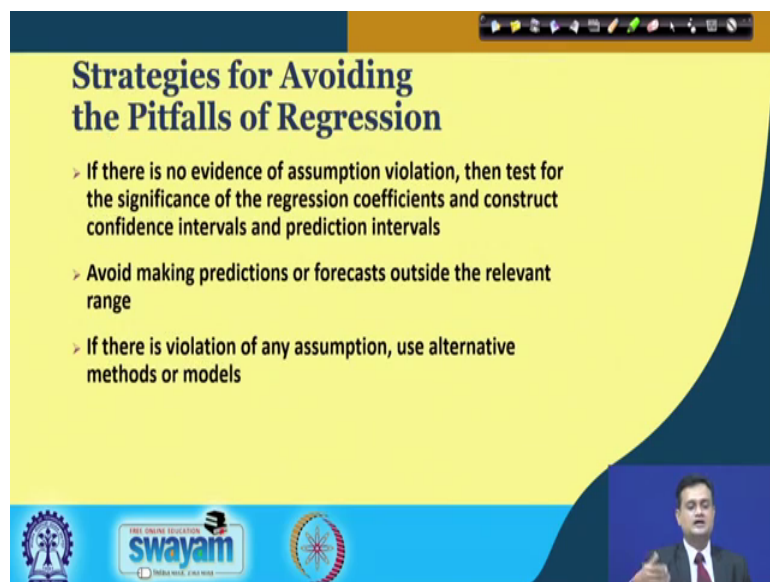
Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X vs. Y to observe possible relationship
- Perform residual analysis to check the assumptions
 - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity
 - Use a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to uncover possible non-normality

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there are logos for 'swayam' and 'INDIA WISE, LEAD WISE' along with a small video inset of a man in a suit.

So, there are certain strategies for avoiding pitfalls; start with a scatter plot; try to see whether there is some relationship exist or not, perform residual analysis, conduct the various test for autocorrelation and the checking the significance of beta 0 and beta 1.

(Refer Slide Time: 39:54)



Strategies for Avoiding the Pitfalls of Regression

- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range
- If there is violation of any assumption, use alternative methods or models

This slide continues the same theme as the previous one, with a yellow background and dark blue curved border. It includes the same logos at the bottom and a video inset of the same man.

And then you can better feel confident about the quality of your regression model.

(Refer Slide Time: 39:59)



Just to share quickly the mini tab application of linear regression.

(Refer Slide Time: 40:05)

A presentation slide with a yellow background and a dark blue curved border on the right. The title "Example" is in bold black. Below it, there is a list of six bullet points. At the bottom, there are logos for "swayam" and "MOOCs". A small video inset in the bottom right corner shows a man in a suit speaking.

Example

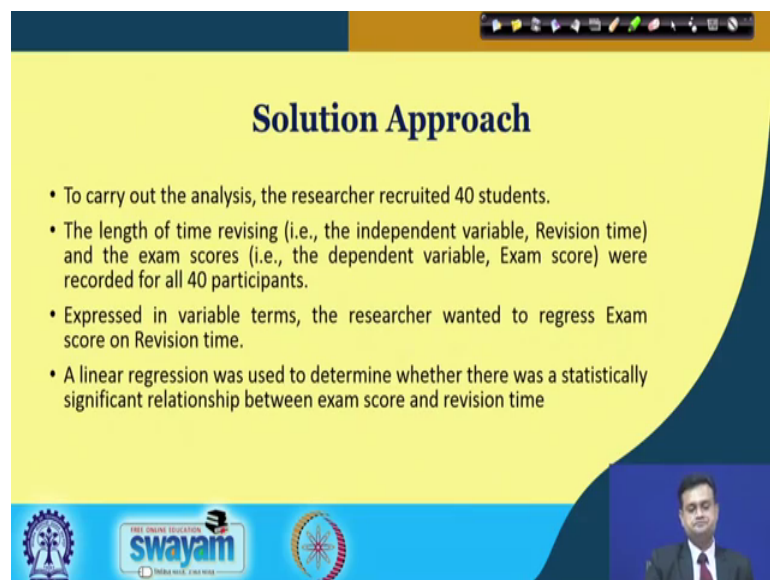
- An educator wants to determine whether students' exam scores were related to revision time.
- For example, as students spent more time revising, did their exam score also increase (a positive relationship); or did the opposite happen?
- The educator also wanted to know the proportion of exam score that revision time could explain, as well as being able to predict the exam score.
- The educator could then determine whether, for example, students that spent just 10 hours revising could still pass their exam.
- Therefore, the dependent variable was "exam score", measured on a scale from 0 to 100, and the independent variable was "revision time", measured in hours.

We have an example that an educator wants to determine whether student's exam score were related to revision time. So, you will also have an exam for this six sigma and we will go through many many lectures there would be more than 60 lectures. And you will find it difficult to revise. So, I would advise that right at this stage when you are going through the lecture please make your summary note which you can revise very quickly at the time of exam.

So, here in this example teacher educator is interested to see that whether there is some relationship between the revision time and the student score in exam. So, here suppose a students spend more time revising did their exam score increase or did the opposite happen. Sometimes opposite may happen the educator also wanted to know the proportion of exam score that revision time could explain.


So, total variability to be explained by the independent variable in the dependent variable and he basically collected the called the dependent variable was the exam score measure on a scale of 0 to 100 and independent variable is the revision time.


(Refer Slide Time: 41:22)



Solution Approach

- To carry out the analysis, the researcher recruited 40 students.
- The length of time revising (i.e., the independent variable, Revision time) and the exam scores (i.e., the dependent variable, Exam score) were recorded for all 40 participants.
- Expressed in variable terms, the researcher wanted to regress Exam score on Revision time.
- A linear regression was used to determine whether there was a statistically significant relationship between exam score and revision time



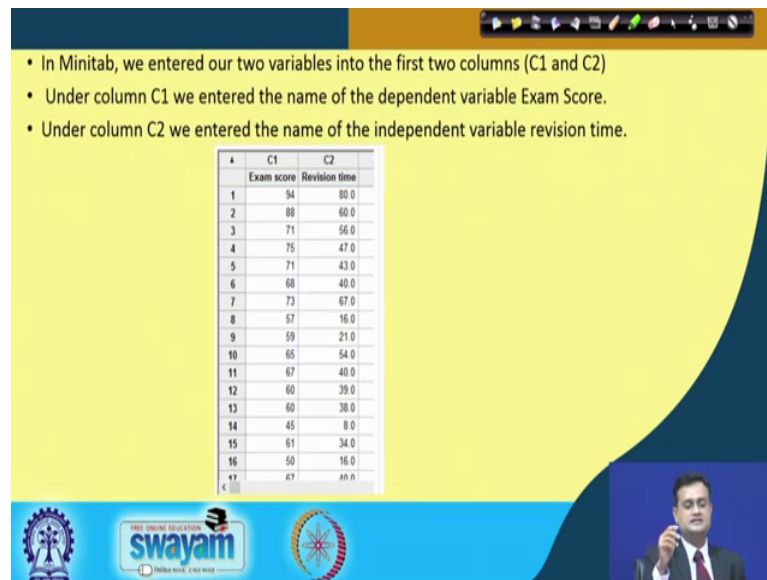


So, we want to carry out this analysis and recorded the revision time exam score of 40 participants.

(Refer Slide Time: 41:32)

- In Minitab, we entered our two variables into the first two columns (C1 and C2)
- Under column C1 we entered the name of the dependent variable Exam Score.
- Under column C2 we entered the name of the independent variable revision time.

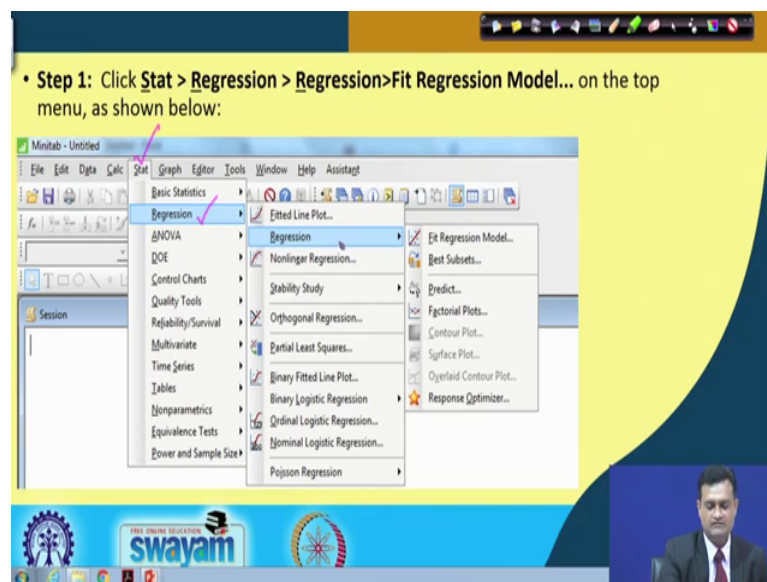
	C1	C2
	Exam score	Revision time
1	94	80.0
2	88	60.0
3	71	56.0
4	75	47.0
5	71	43.0
6	68	40.0
7	73	67.0
8	57	16.0
9	59	21.0
10	65	54.0
11	67	40.0
12	60	39.0
13	60	38.0
14	45	8.0
15	61	34.0
16	50	16.0
17	47	11.0



And just try to tabulate this in the Minitab for excel sheet and then borrow it. So, C1 we entered the dependent variable that is the exam score, C2 is the independent variable revision time.

(Refer Slide Time: 41:46)

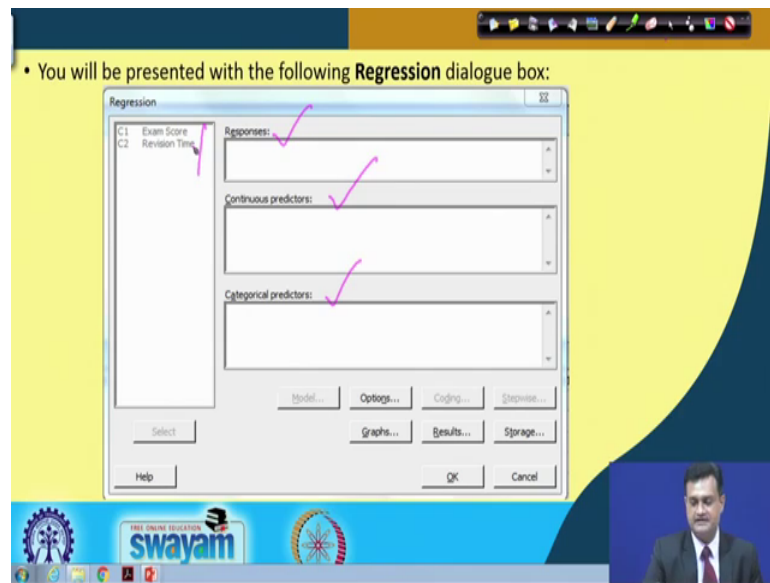
- **Step 1:** Click **Stat > Regression > Regression>Fit Regression Model...** on the top menu, as shown below:



The screenshot shows the Minitab software interface with the 'Stat' menu open, and the 'Regression' submenu selected, highlighting the 'Fit Regression Model...' option. The presenter is visible in the bottom right corner.

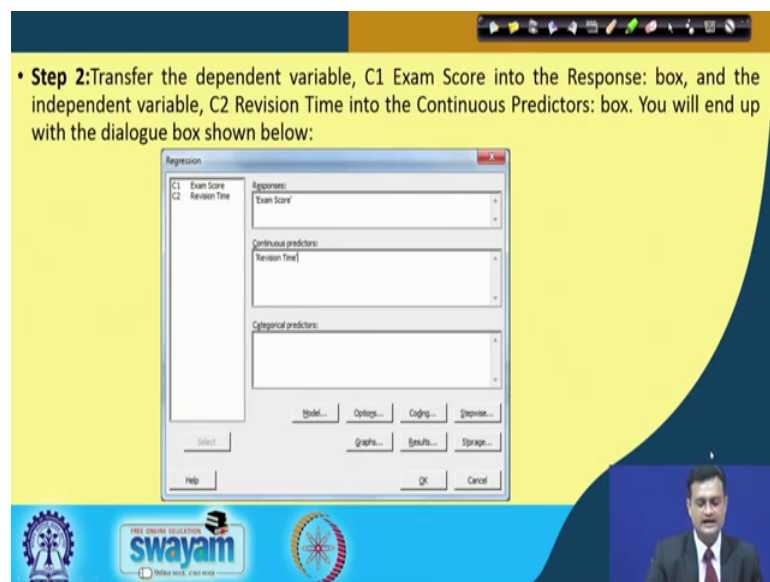
Now, just see how it is easy to perform regression analysis in Minitab. Go to stretch see the regression then you will have a drop down window regression fit regression model just doing this after doing this.

(Refer Slide Time: 42:06)

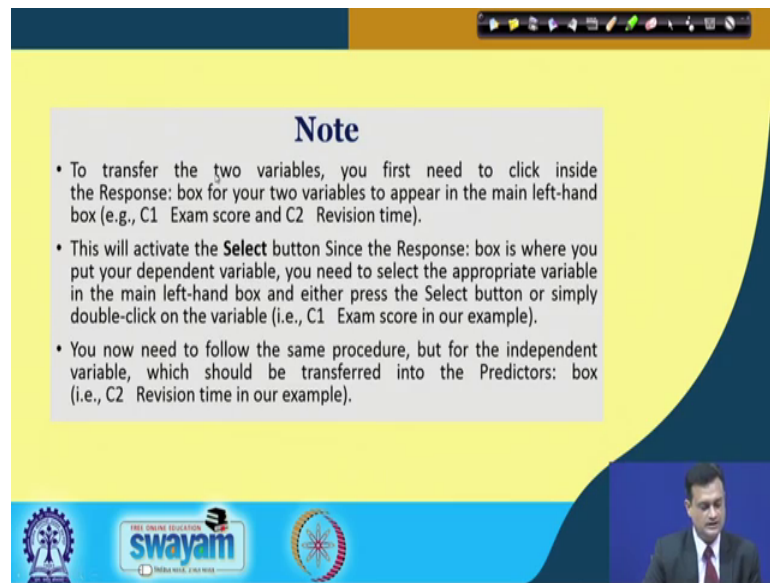


You will have this kind of window where you will see the response variable to be entered continuous production categorical predictor. You have basically two exam score has the dependent variable revision time as independent variable.

(Refer Slide Time: 42:27)



(Refer Slide Time: 42:29)



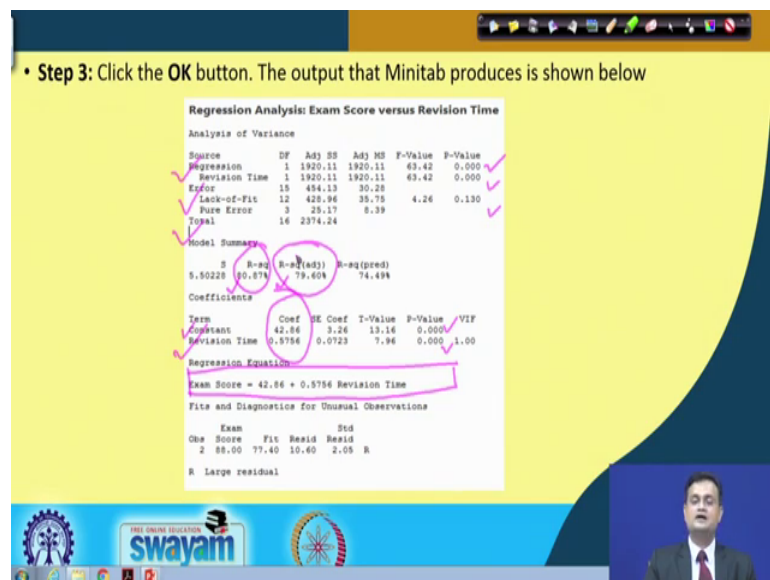
Note

- To transfer the two variables, you first need to click inside the Response: box for your two variables to appear in the main left-hand box (e.g., C1 Exam score and C2 Revision time).
- This will activate the **Select** button Since the Response: box is where you put your dependent variable, you need to select the appropriate variable in the main left-hand box and either press the Select button or simply double-click on the variable (i.e., C1 Exam score in our example).
- You now need to follow the same procedure, but for the independent variable, which should be transferred into the Predictors: box (i.e., C2 Revision time in our example).

swayam

So, once you have this window then, you can choose this from this particular say window and put it here exam score revision time.

(Refer Slide Time: 42:38)



• **Step 3: Click the OK button. The output that Minitab produces is shown below**

Regression Analysis: Exam Score versus Revision Time

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1920.11	1920.11	63.42	0.000
Revision Time	1	1920.11	1920.11	63.42	0.000
Error	15	494.13	32.94		
Lack-of-Fit	12	428.96	35.75	4.26	0.130
Pure Error	3	25.17	8.39		
Total	16	2374.24			

Model Summary

S	R-sq	R-sq(Adj)	R-sq(Pred)
5.50228	80.87%	79.60%	74.49%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	42.86	3.26	13.16	0.000	
Revision Time	0.5756	0.0723	7.96	0.000	1.00

Regression Equation

Exam Score = 42.86 + 0.5756 Revision Time

Fits and Diagnostics for Unusual Observations

Exam	Score	Fit	Resid	Std
Obs	85.00	77.40	10.60	2.05
2				

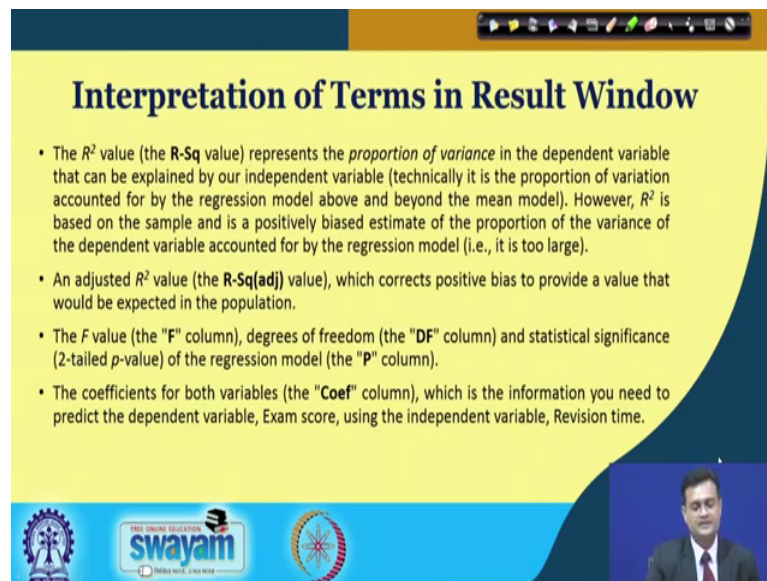
R Large residual

swayam

Then you click ok and you will have the output generated. So, you can see here that you have the source of variation degree of freedom adjusted sum of square. Adjusted mean sum of square F value P value. So, regression error total you have the P value you can check whether your beta 0 beta 1; they are significant or not here you have the coefficient and corresponding p value.

So, constant and revision time and you have the regression equation sat that is exam score is equal to 42.86 plus 0.5756 into revision time and you have also the values of R square and adjusted r square. So, you can see here that R square and adjusted R square is very close. So, whatever independent variable I have that is revision time it is really able to explain the variability in the dependent variable which is the exam score.

(Refer Slide Time: 43:46)



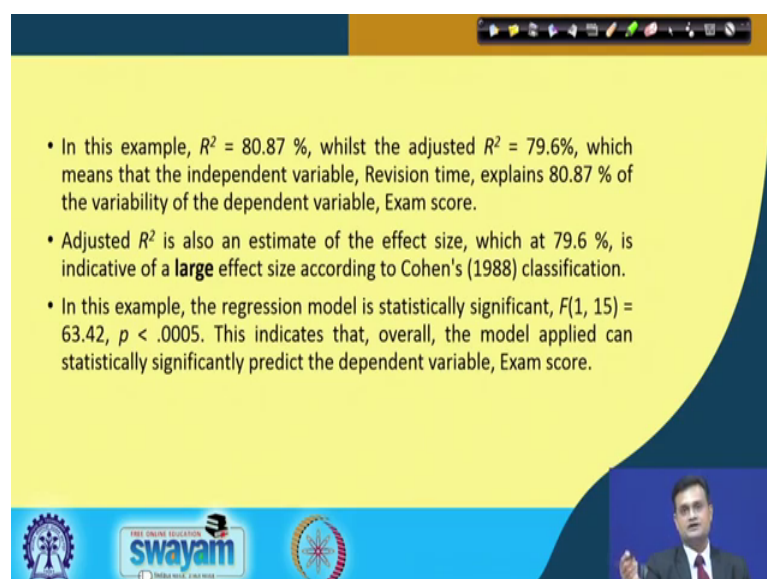
Interpretation of Terms in Result Window

- The R^2 value (the **R-Sq** value) represents the *proportion of variance* in the dependent variable that can be explained by our independent variable (technically it is the proportion of variation accounted for by the regression model above and beyond the mean model). However, R^2 is based on the sample and is a positively biased estimate of the proportion of the variance of the dependent variable accounted for by the regression model (i.e., it is too large).
- An adjusted R^2 value (the **R-Sq(adj)** value), which corrects positive bias to provide a value that would be expected in the population.
- The F value (the "**F**" column), degrees of freedom (the "**DF**" column) and statistical significance (2-tailed p -value) of the regression model (the "**P**" column).
- The coefficients for both variables (the "**Coef**" column), which is the information you need to predict the dependent variable, Exam score, using the independent variable, Revision time.

swayam

So, this analysis you can do very easily. And you can interpret as I explained; what is adjusted R square? What is R square? What is coefficient column?

(Refer Slide Time: 43:54)

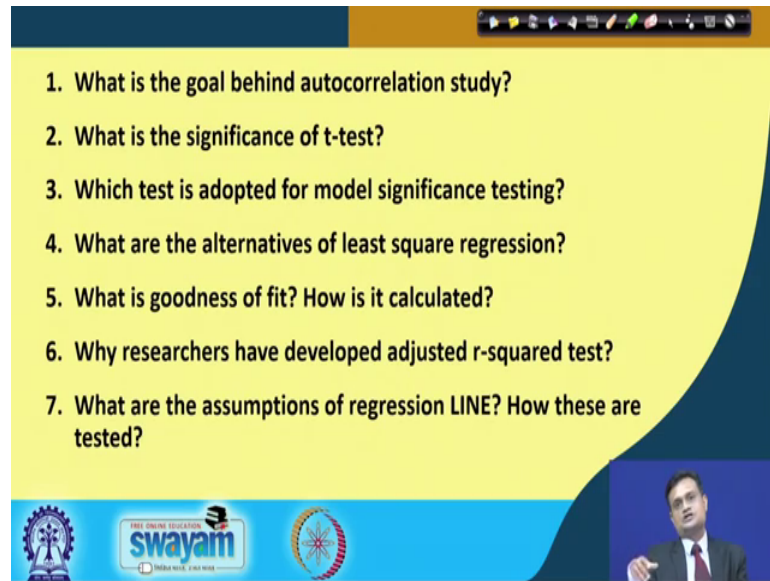


- In this example, $R^2 = 80.87\%$, whilst the adjusted $R^2 = 79.6\%$, which means that the independent variable, Revision time, explains 80.87 % of the variability of the dependent variable, Exam score.
- Adjusted R^2 is also an estimate of the effect size, which at 79.6 %, is indicative of a **large** effect size according to Cohen's (1988) classification.
- In this example, the regression model is statistically significant, $F(1, 15) = 63.42$, $p < .0005$. This indicates that, overall, the model applied can statistically significantly predict the dependent variable, Exam score.

swayam

And you can interpret the results like this that R square is equal to 80.87 percent adjusted R square is 79.6 no issue with the quality of the model. R square estimate 79.6. So, its a indicative or a large effect size and you have the p value 0.0005. So, this indicates that overall the model applied can statistically significantly predict the dependent variable that is the exam score.

(Refer Slide Time: 44:29)



1. What is the goal behind autocorrelation study?
2. What is the significance of t-test?
3. Which test is adopted for model significance testing?
4. What are the alternatives of least square regression?
5. What is goodness of fit? How is it calculated?
6. Why researchers have developed adjusted r-squared test?
7. What are the assumptions of regression LINE? How these are tested?

The slide also features a 'swayam' logo and a small video inset of a man in a suit in the bottom right corner.

So, think it before we end; what is the goal behind autocorrelation study? What is the significance of t test? What test is adapted for model significance testing? What is the goodness of fit in regression model?

How it is calculated? Why researchers have develop adjusted R square? And what are the assumptions of regression line how these are tested?

(Refer Slide Time: 44:51)

References:

- ❑ Aczel, A., Sounderpandian, J. and Saravanan, P. , Complete Business Statistics, McGraw Hill Publication.
- ❑ David M. Levine, Timothy C. Krehbiel, Mark L. Berenson and P. K. Vishwanathan, Business Statistics, Pearson Publication.
- ❑ T. M. Kubiak, Donald W. Benbow, The Certified Six Sigma Black Belt Handbook, Pearson Publication.
- ❑ Forrest W. Breyfogle III, Implementing Six Sigma, John Wiley & Sons, INC.
- ❑ <https://statistics.laerd.com/minitab-tutorials/linear-regression-using-minitab.php>

The slide features a dark blue background on the left with the word 'References' in a yellow script font. The right side is a light yellow trapezoidal shape containing the reference list. At the bottom, there are logos for a university, 'swayam' (Free Online Education), and a presenter's video feed.

So, use this references for your better understanding and revising the concept.

(Refer Slide Time: 44:58)

Conclusion:

- ❖ Correlation is used to test if residuals in one time period are related to residuals in another period
- ❖ T-test tests if there is any linear relationship between dependent variable and independent variables
- ❖ F-test checks the validity of null or alternate hypothesis

The slide features a dark blue background on the left with the word 'Conclusion' in a yellow script font. The right side is a light yellow trapezoidal shape containing the conclusion list. At the bottom, there are logos for a university, 'swayam' (Free Online Education), and a circular emblem.

Correlation is used to test if residual in one time period are related to residual in other period. T-test if there is any linear relationship between dependent and independent variable, F-test checks the validity of null or alternate hypothesis for my regression model. And with this I would say that this lecture you should revise appropriately if necessary try to revise it twice strengthen your understanding.

And with this thank you very much do well in your day to day studies professional. And try to now visualize the application of; six sigma in some of the facets. So, here now you should try to define some of the problems in your day to day life in your organization siding where you would like to apply DMAIC cycle for the improvement.

So, with this thank you very much be with me enjoy.