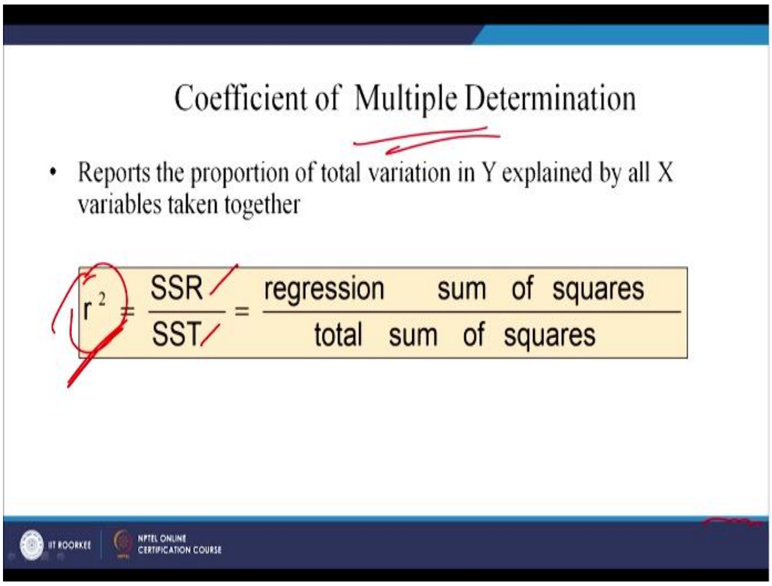
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 59
Multiple Regression

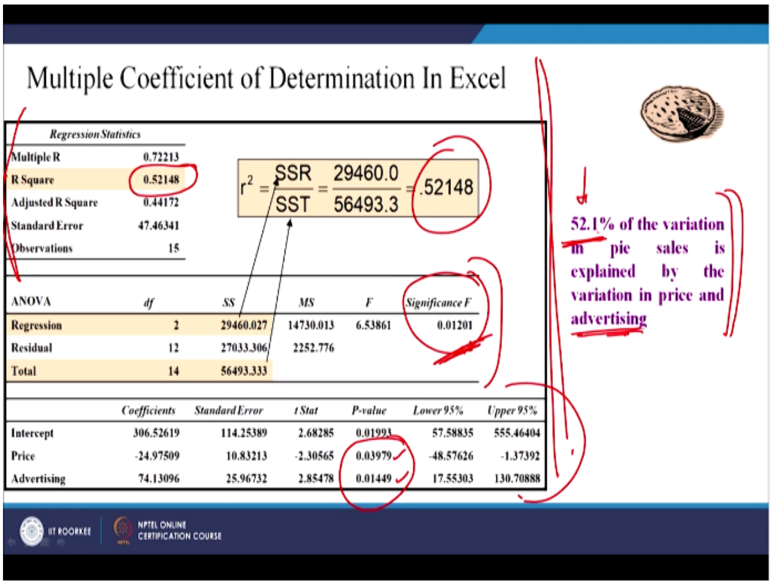
Hello friends. I welcome you all in this session. As you are aware in previous session, we were discussing about multiple linear regression and we have seen how to work out a question having more than one independent variables. Today, will see some more examples on multiple regression including one of the variables, one of the independent variables as categorical variable.

(Refer Slide Time: 00:59)



So let us look at what is r squares coefficient of determination similar to simple linear regression. So there itself we have seen that it is regression sum of square to total sum of square. So same formula will apply in case of multiple regression as well. So this is how you can calculate r square value for a given question.

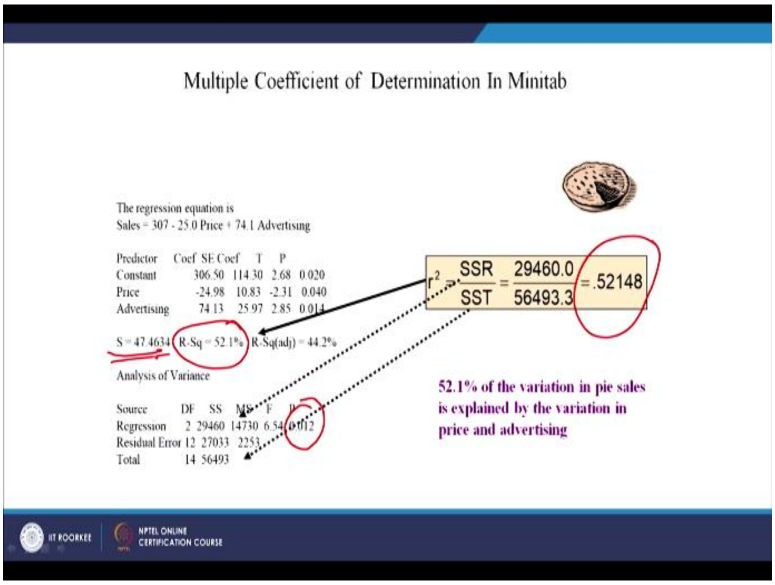
(Refer Slide Time: 01:21)



In fact, if you look at the same example which we worked out in previous session when the dependent variable was sales, the independent variables were price and the second independent variable was advertising expenditure right. So this is the output of multiple regression question. So r square is 52.148% right. This can be seen our here in this part of the table. It has got 3 parts in this table. First part, second is ANOVA table and this is third part right and this is nothing but the P value of F right.

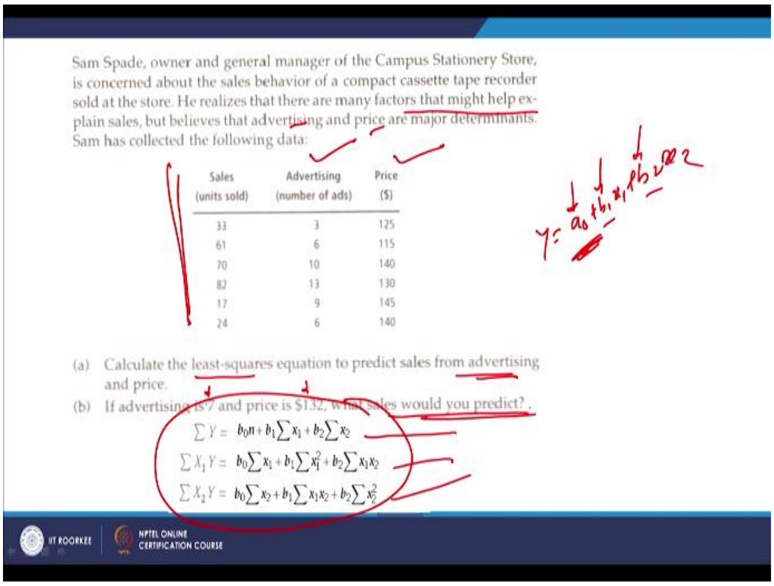
Whether the overall significant model is where the overall model is significant or not and these are P values of individual independent variables. So these two are also significant independent variables and these two independent variables explain 52.1% variation in dependent variable.

(Refer Slide Time: 02:31)



This output from Minitab for the same question will get the same information, this r square, this P value, you have got r square here, standard error 57.46 and so on right. So this is the output from Minitab.

(Refer Slide Time: 02:50)

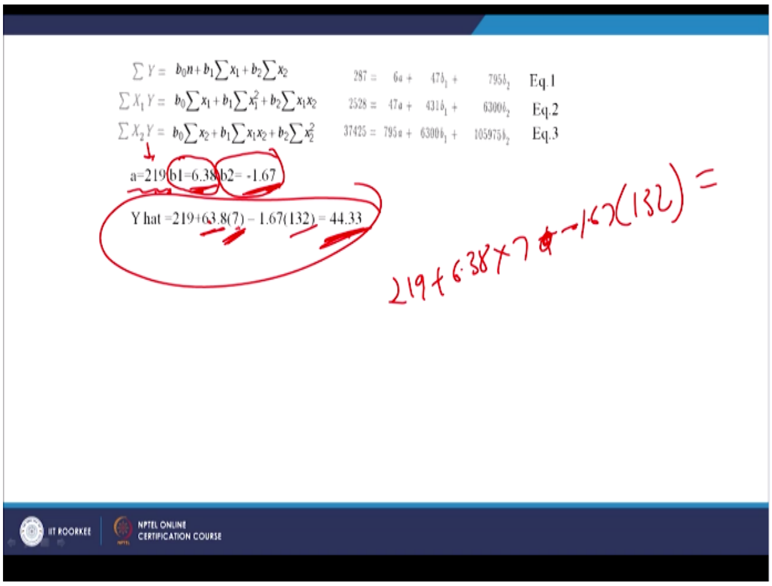


So let us look at one more example on multiple regression. So Sam Spade owner and general manager of Campus Stationary Store is concerned about the sales behaviour of compact cassette tape recorder sold at the store. He realizes that there are many factors that might help explain sales but believe that advertising and price are more important ones compared to other factors right.

So she has collected following data right, so number of units sold, advertising expenditure and price. So she wants to know is there any relationship between dependent and these two independent variables. So calculate least square equation to predict the sales from advertising and price right. If advertising expenditure is 7 dollars and price is 132 dollars, what would be the sales?

So as I have told you that you can work out this question because this is a question y is = a_0 let us say $b_1x_1+b_2x_2$ right. So this is one unknown variable, second and third right so 3 unknown variables, you have got 3 equations right; first, second and third. So you can solve these equations and you can get a0, b1, b2 right.

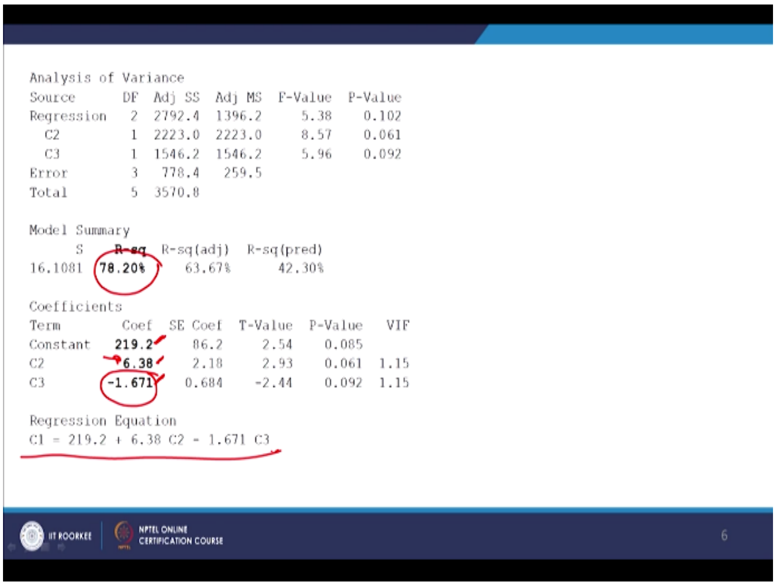
(Refer Slide Time: 04:25)



So this is how you would be getting it, so a is=219, b1 is the intercept of first independent variable is the coefficient not intercept, this is intercept, this is coefficient of second one right which is price right. So if you increase price, sales will come down by this much amount and if you increase advertising expenditure, sales will increase by this much percentage right and this much value not percentage.

So when advertising expenditure is 7, just look at this. Yeah when advertising is 7 and price is 132, 7 and 132 this would be in fact I think this is 63.8 is what we have written here 63.8, no it is not 63 it is 6.38, so this equation is let me rewrite this one. It is 219+6.38*7+ not plus this is minus 1.67*132. So you will get this value as 44 right. So this is 6.38 right, just you can make this correction.

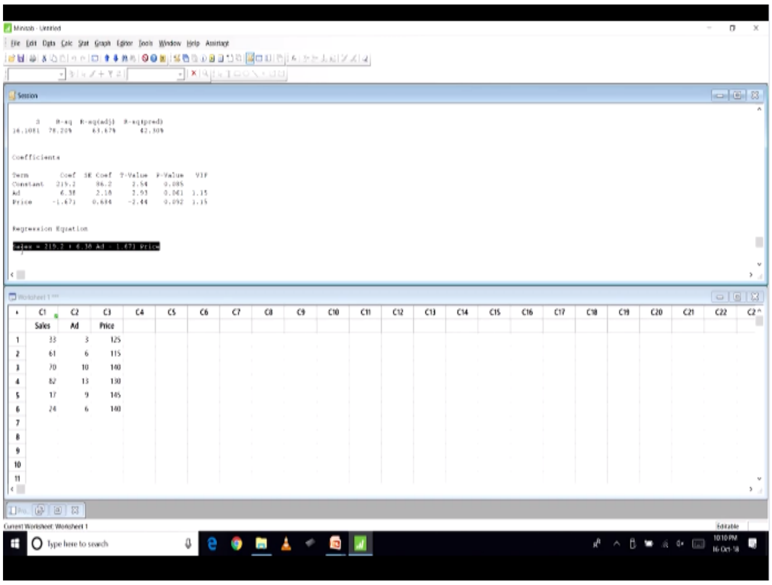
(Refer Slide Time: 06:07)



Now this is the output from Minitab software for this question. So r square is 78.2 it means the variance in dependent variable which is being explained by two independent variables is 78.20%. This is your constant value or intercept, coefficient of first independent variable and coefficient of second independent variable right, so this is price and this is advertising expenditure. So this is your regression equation okay.

So let us work out this question using Minitab. So this is the question we have got. Let us solve this question using Minitab.

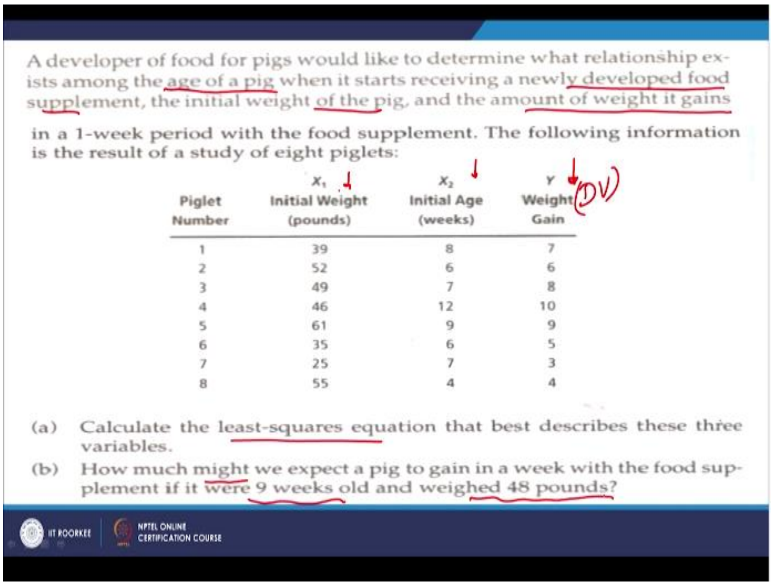
(Refer Slide Time: 06:57)



So first of all you need to write the dependent variable, sales, advertisement and price is 33, 61, 70, 82, 17 so just 6 is the sample size so advertising expenditure is 36, 10, 13, 9 and 6. Then, you need to enter price as well 125, 115 150, 130, 155 and 140. So will go to stat, regression, fit regression model, so response is sales right, so this is your dependent variable and these two are continuous predicted variables or independent variables.

So we click at okay so this is what the answer you are getting right. So if you look at this these are ANOVA table, P value 0.09 and 0.06. So these two are significant at let us say 10% significance level. If you look at r square value this is 78.20 that's what I told you and this is your regression equation. So if you want to calculate sales for a given price and for advertisement expenditure you can easily calculate okay. So this is how you can solve a question on multiple regression using Minitab right.

(Refer Slide Time: 08:58)



A developer of food for pigs would like to determine what relationship exists among the age of a pig when it starts receiving a newly developed food supplement, the initial weight of the pig, and the amount of weight it gains in a 1-week period with the food supplement. The following information is the result of a study of eight piglets:

Piglet Number	x_1 Initial Weight (pounds)	x_2 Initial Age (weeks)	Y Weight Gain (DV)
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

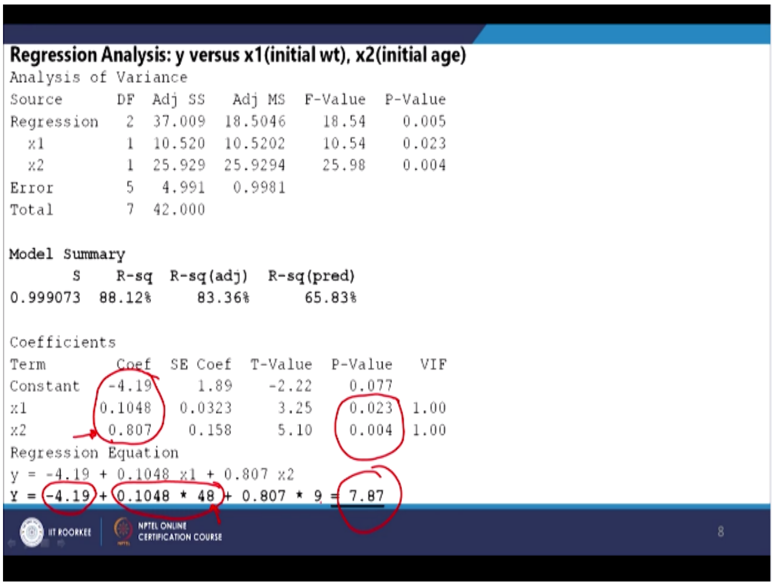
(a) Calculate the least-squares equation that best describes these three variables.

(b) How much might we expect a pig to gain in a week with the food supplement if it were 9 weeks old and weighed 48 pounds?

So let us look at one more example on this. A developer of food for pigs would like to determine what relationship exists among the age of a pig when it starts receiving a newly developed food supplement. The initial weight of the pig and the amount of weight it gains. So the weight it gains in a week period with the food supplement right. The following information is available right.

So weight of the pig, initial age and initial weight, so this is your dependent variable and these two are independent variables. So first of all you need to calculate the least square equation and then what would be the weight right, weight if the age is let us say 9 weeks and the initial weight was 48 pounds.

(Refer Slide Time: 10:07)

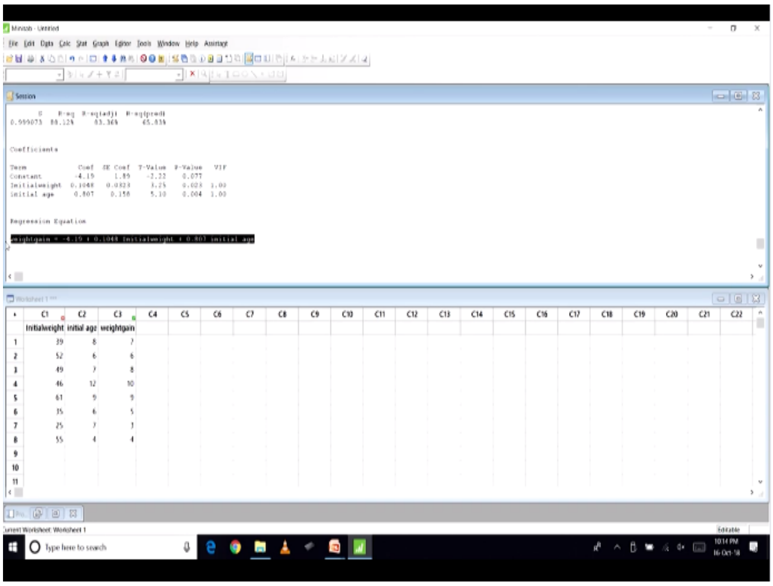


Regression Analysis: y versus x1(initial wt), x2(initial age)

```
Analysis of Variance
Source      DF  Adj SS   Adj MS  F-Value  P-Value
Regression   2  37.009  18.5046    18.54    0.005
  x1         1  10.520  10.5202    10.54    0.023
  x2         1  25.929  25.9294    25.98    0.004
Error        5   4.991   0.9981
Total        7  42.000

Model Summary
       S    R-sq  R-sq(adj)  R-sq(pred)
0.999073  88.12%     83.36%      65.83%

Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   -4.19     1.89    -2.22    0.077
x1        0.1048   0.0323     3.25    0.023  1.00
x2         0.807    0.158     5.10    0.004  1.00

Regression Equation
y = -4.19 + 0.1048 x1 + 0.807 x2
Y = -4.19 + 0.1048 * 48 + 0.807 * 9 = 7.87
```

So when you solve this question using Minitab this would be the output you would be getting. So if you look at this at 5% significance level, both are important right. If you look at this coefficients right just see this so this -4.19 is your intercept, coefficient of X1 which is 0.1048, coefficient of X2 independent variable which is initial age right okay. So finally at 48 pounds an initial age of 9 and 48 pounds which is initial weight, the weight gain would be 7.87 pounds right.

(Refer Slide Time: 11:17)

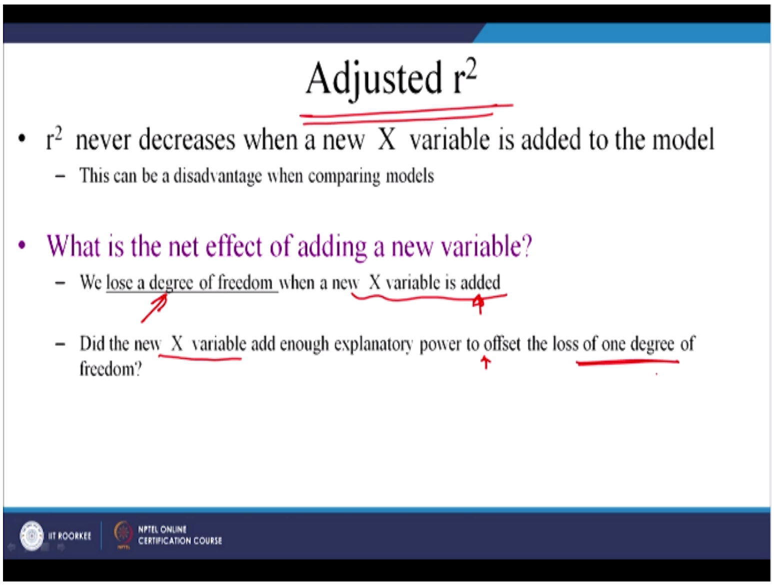


So you can work out this example again using Minitab, so you will have to enter data for dependent and independent variables. So first of all your dependent variable which is let it be initial weight for independent variable not a problem, initial age and then weight gain right. Now just go for entering data for initial weights which is 39, 52, 49, 46, 61, 35, 25 and finally 55. So initial age 8, 6, 7, 12, 6, 7 and 4 so initial age in terms of weeks right.

And the weight gain which is the dependent variable in terms pounds. So will go to stat, regression, regression, fit regression model, so here c3 is the yeah this is your response variable or dependent variable while c1 and c2 are independent variables, so click okay so you will get the same answer right. In fact, in slide I have taken this output only right. So just see this is your regression line.

This -4.19+0.10 initial weight +0.807 initial age right. So you can work out questions on multiple regression using Minitab.

(Refer Slide Time: 13:20)



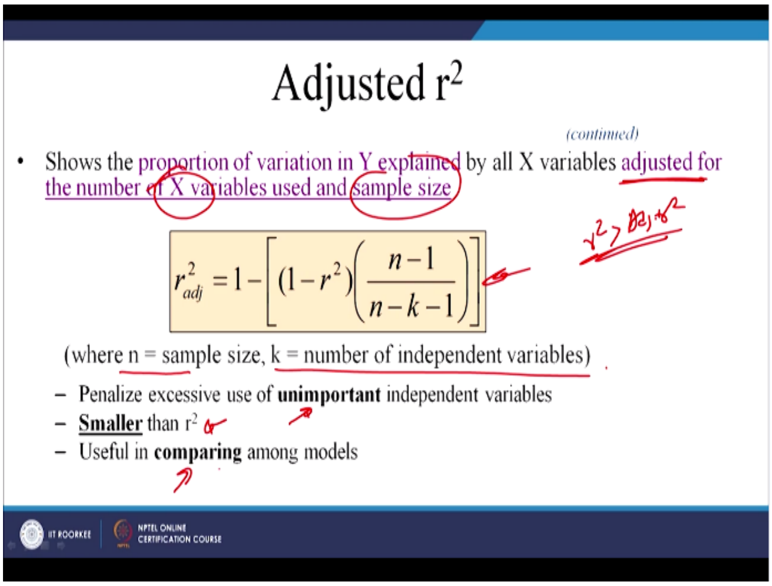
So far we have checked the goodness of model by looking at r square value. Now you may have a situation where there are two models, both have got different sample size and both of them have got different number of independent variables and the situation like this you need to compare you need to use adjusted r square for comparison purpose. So there is something called adjusted r square.

Now what happens to the r square value whenever you add an independent variable the r square value increases. So let us say r square is 90% with two independent variables, if you add one more independent variable r square will go up. If you add one more, r square will again go up. So the characteristics of r square is that it always increases with addition of an independent variable irrespective whether that independent variable is really adding to dependent variable or not.

So that is the drawback of r square, r square does not taken into account sample size, it does not take into account an independent variables as well right. So there is one more measure for looking goodness of fit of the model is adjusted r square. So whenever we add an independent variable what happens, is that we lose one degree of freedom and the moment you lose one degree of freedom and the value of r square will increase.

So you need to have a balance whether you are losing a degree of freedom and the variable which you are adding, is it really significant one? So there has to be a balance between these two. So you need to always ask a question did the new variable add enough explanatory power to offset the loss of one degree of freedom right.

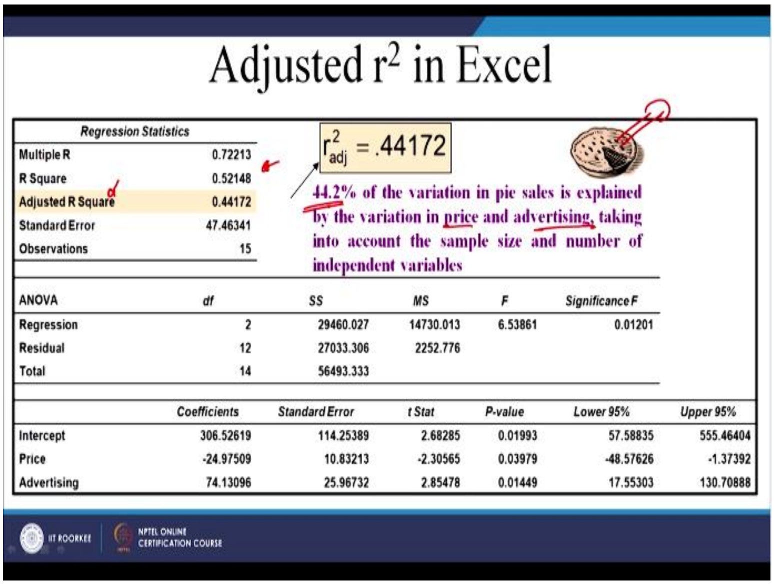
(Refer Slide Time: 15:45)



So there is something called adjusted r square which is adjusted for two things, sample size as well as for number of independent variables and it will always be, its value will always be smaller than so now r square will always be greater than adjusted r square right. So in other words it always be smaller than r square. So this is how you can calculate adjusted r square for a given question.

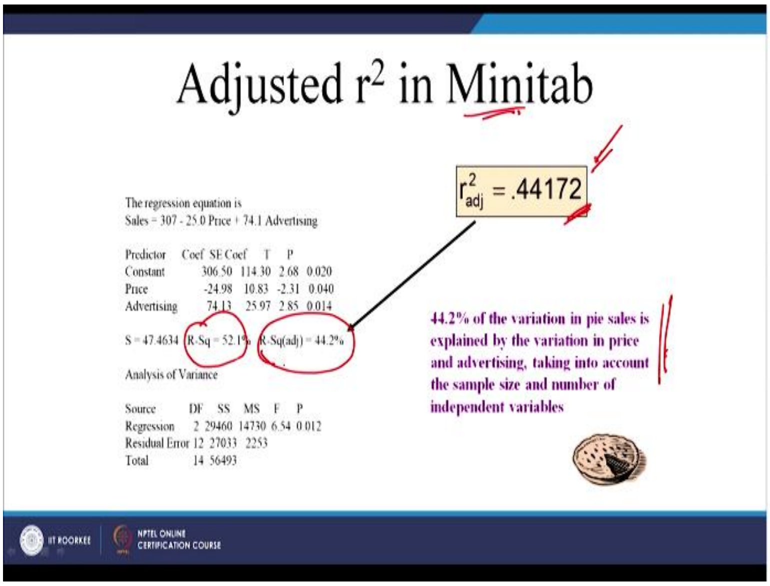
So n is sample size, k is number of independent variables. So what adjusted r square is the moment you add an independent variable if it is a significant one if it is really helping the dependent variable then only the adjusted r square will increase, otherwise it will not increase. So in penalizes the excessive use of unimportant independent variables. It will always be smaller than r square and it is useful for comparing different models.

(Refer Slide Time: 16:49)



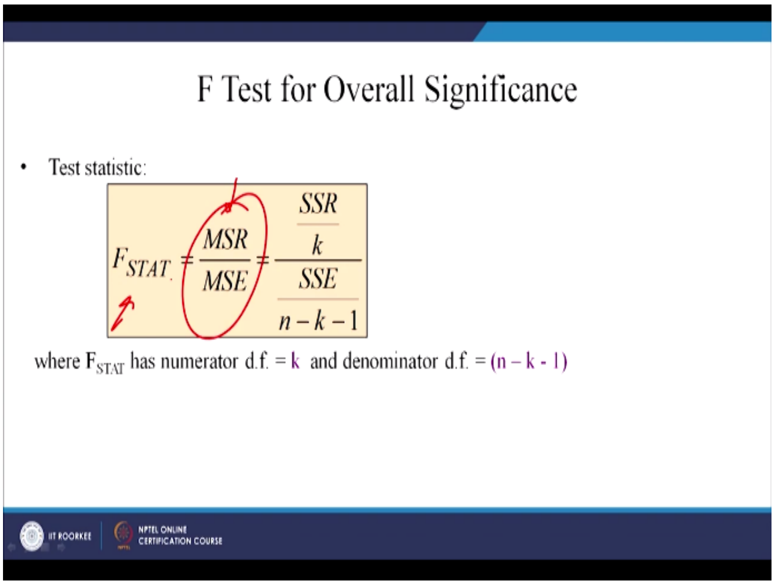
So this is the adjusted r square value for a question which we solved in previous class on multiple regression. So r square was 52.148% but the adjusted r square was 44.2% so now onwards do not look at r square value in any regression model, always look at adjusted r square because it gives you a better picture of the model. So we will say that 44.2% of variance in data is explained by independent variables namely the price and advertising expenditure right. So this is how you can find out adjusted r square.

(Refer Slide Time: 17:35)



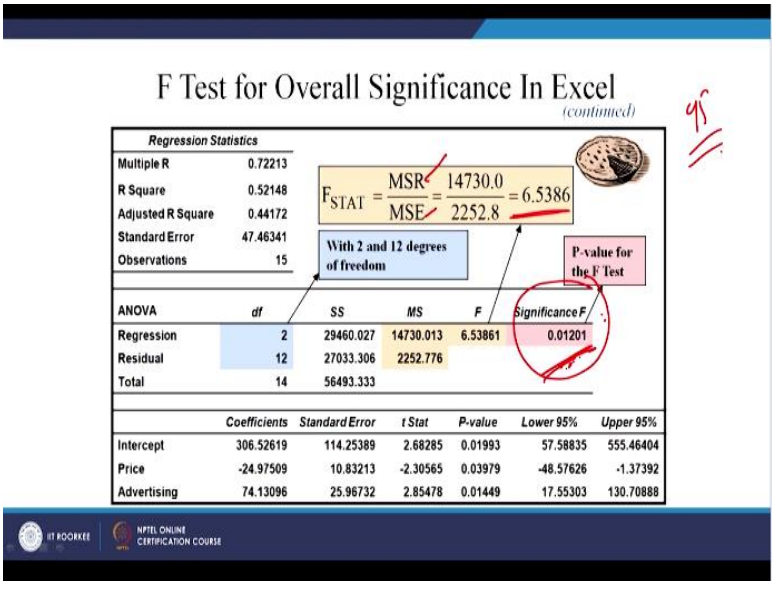
Now this is output form Minitab same value and same interpretation right. So here you are getting r square and adjusted r square which is always smaller than r square.

(Refer Slide Time: 17:54)



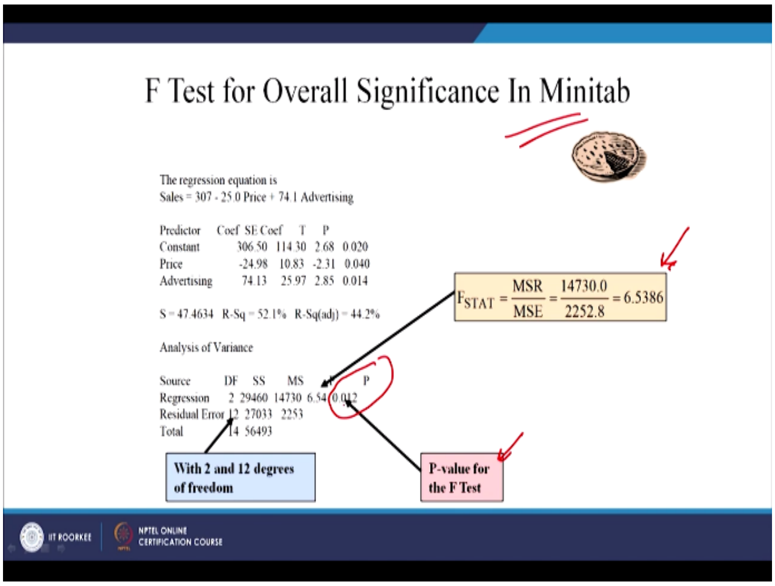
Now let us test the model for overall significance. So whether the entire model is significant or not, though we have seen the P value of F right, so that will give you whether the model is significant or not. So F statistics is this or ratio of these two.

(Refer Slide Time: 18:19)



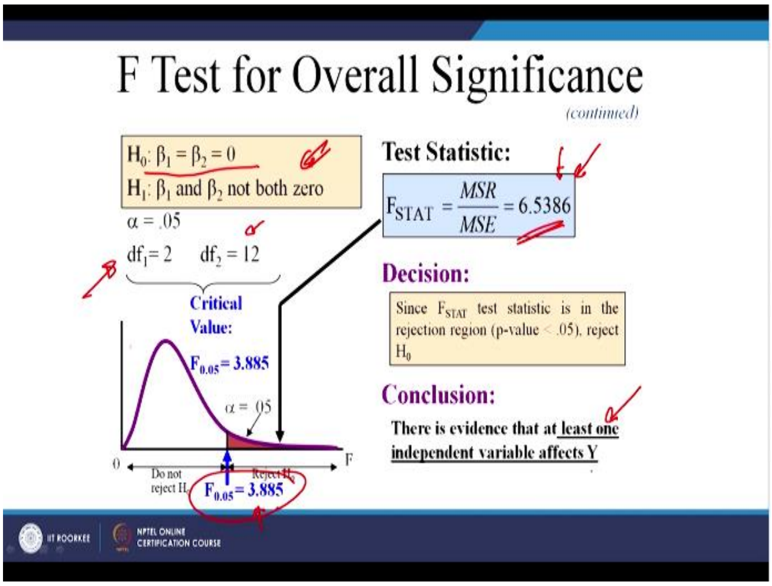
It is regression sum of square to error sum of square right. F statistics is 6.53 in this question and this is what is P value for our test. So will say that the model is significant at 95% significance level right.

(Refer Slide Time: 18:39)



Of course, you will have to find out this output from Minitab, same values for F and for P as well and this is P value right 0.012.

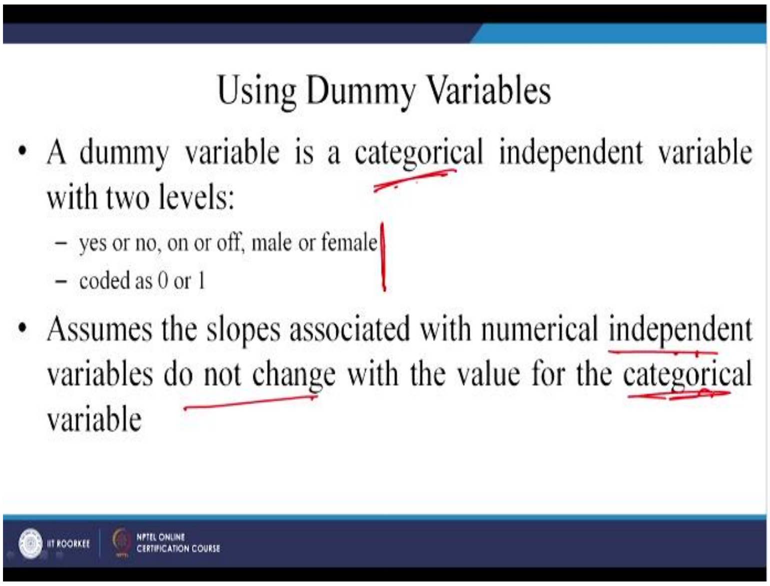
(Refer Slide Time: 18:53)



So here for testing overall significance of multiple regression model you are framing a null hypothesis that both these coefficients are equal and they are 0. Beta 1 and beta 2 are 0 and alternative is that they are not 0. So calculated F statistics is this, the table value at appropriate degrees of freedom at 2 numerator and 12 denominator. So you will calculate F value like this and your table value is 3.885 so will reject the null hypothesis.

You will reject the null hypothesis it means these coefficients are not same, so there is evidence that at least one independent variable affects dependent variable Y. Now let us look at something called dummy variables. So far we have taken cases wherein the independent variables were metric in nature. Now let us take an example where dummy variable is a nonmetric.

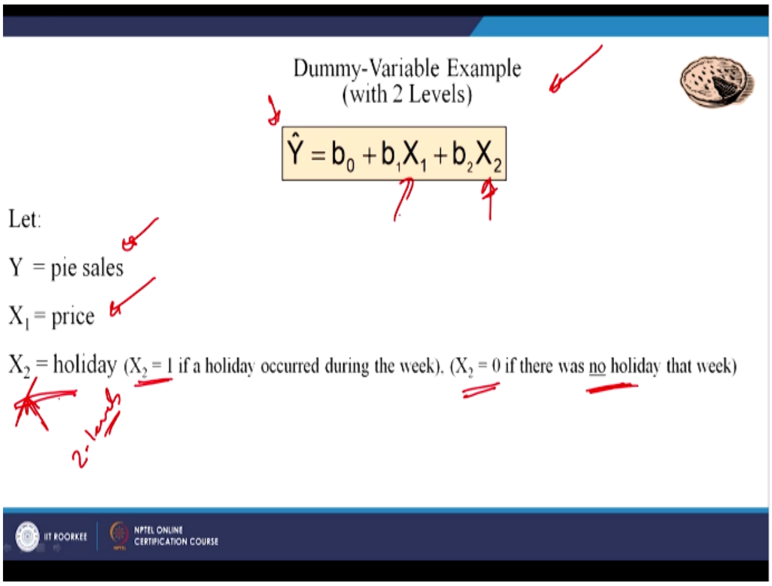
(Refer Slide Time: 20:07)



So nonmetric variables are you can think of again several examples, let us say male, female, group 1, group 2, heavy coffee drinker, medium coffee drinker right, 0, 1 right, married, bachelor is not it, male, female and so on. So you can have different categorical variables and you can have different levels as well right but when I say whether a person is married or bachelor just two levels of a categorical variable right.

So it assumes the slopes associated with numerical independent variables do not change with value of categorical variable right.

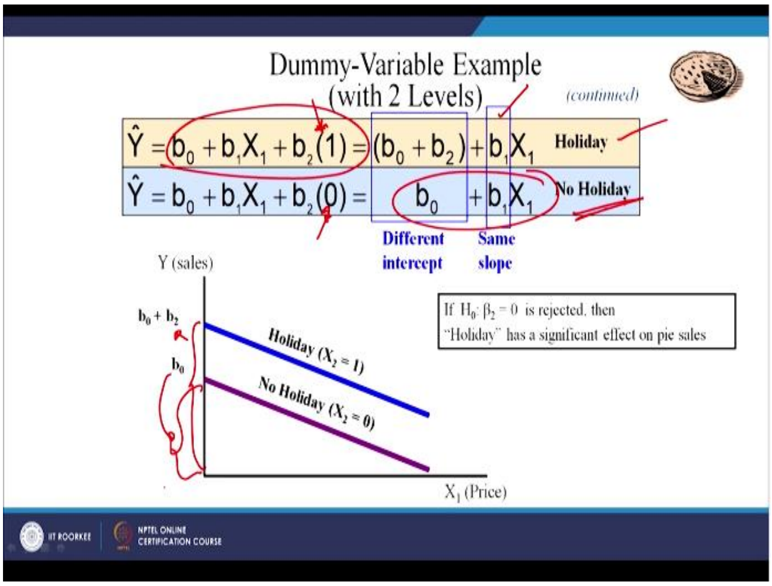
(Refer Slide Time: 20:55)



So let us look at this example, so will take the example which we have already solved but here in this question this second independent variable is categorical in nature. So Y is sales, X1 is price of the product and X2 is a day of the week, now it may be a holiday or it may be a working day. So will say X2 is a categorical variable with two levels, is not it? So holiday, X2 if it is 1, if it is 0 it was no holiday right so working day.

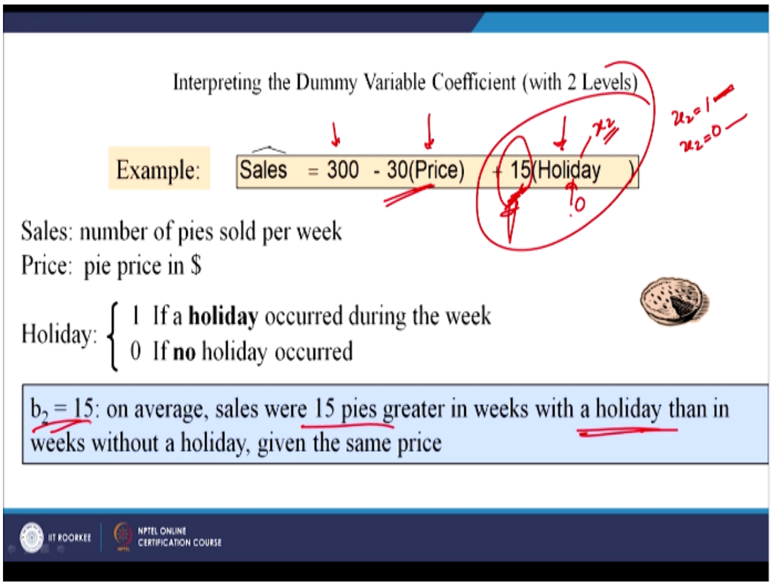
So you can find out what is the effect of holiday or a working day on sales for a given value of for a given price.

(Refer Slide Time: 21:49)



So Y hat is=b0+b1 X1+b2 is=1 it means it is a holiday. We are saying X2 is=1 means it is a holiday so you can rewrite this equation like this. So this is the slope for X1 and this is slope for X2 right. Similarly, when it is not a holiday, it means now b2 is this is 0, X2 is=0 and you can rewrite this like this. So when it is a holiday, this is the slope right, this is the slope b0+b2. If it is not a holiday then this is the slope b0. So this is how you can write equations for a categorical independent variable.

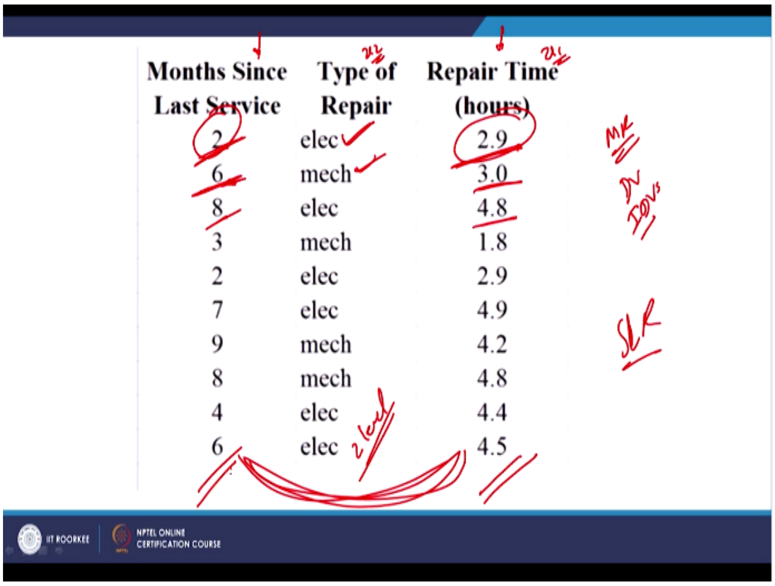
(Refer Slide Time: 22:43)



So for this question we are just looking at the output, we are not looking at input data. So will say that the sales is=300 - 30*price + 15*second independent variable which is X2 right. Now as we know that X2 is=0 and X2=1, 1 and 0 right, so when I say X2 is=1 it means it is a holiday right. So if it is a holiday then the sales will increase by 15 units. Otherwise, if it is 0 then this entire term will become 0 when you put 0 over here.

So b2 is 15 on an average sales where 15 pies greater than in weeks with a holiday. So if there is a holiday in a week, then the sales will increase by 15 units, otherwise it will not increase by 15 units. So this is how you can take an example having independent variable which is a categorical one right. So let us look at one more question.

(Refer Slide Time: 24:21)



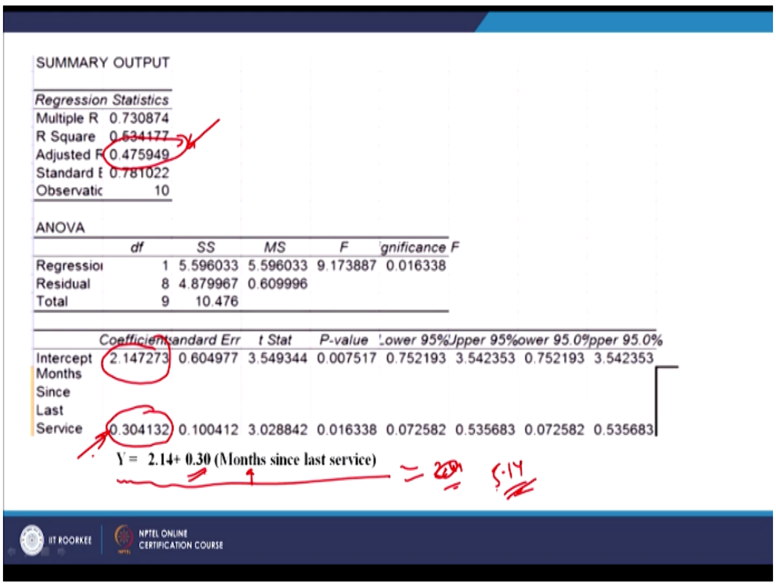
This is a question on multiple regression, you have got dependent variable and you have got independent variables right. So the question goes like this let us say there is a building and whenever there is some repair work is to be done in the building, you just call repair crew members and get the work done right get the repairing done. So you have got two types of information available with you.

So month since last service and repair time, so we know that if the maintenance is frequently happening then repair time would be less. If the time between two different two maintenance is more then definitely repair time would be higher right. So we want to know is there any relationship between the repair time and month since last service. So let us say month since last service is 2 months, repair time was 2.9 hours.

Months since last service when from previous service to next service if the duration is 6 months, then this is the repair time. When you put this 8 months, then this is the repair time. So we want to know is there any relationship between these two right. So this is your independent variable X1 right, this second independent variable is this X2 which is categorical in nature which is nonmetric in nature.

And we know that the repair time is also depends on whether the repair type is an electrical type or mechanical type, is not it? So we will have two levels for this particular independent variable. So first we will solve this question as just simple linear regression model with one dependent variable and one independent variable right.

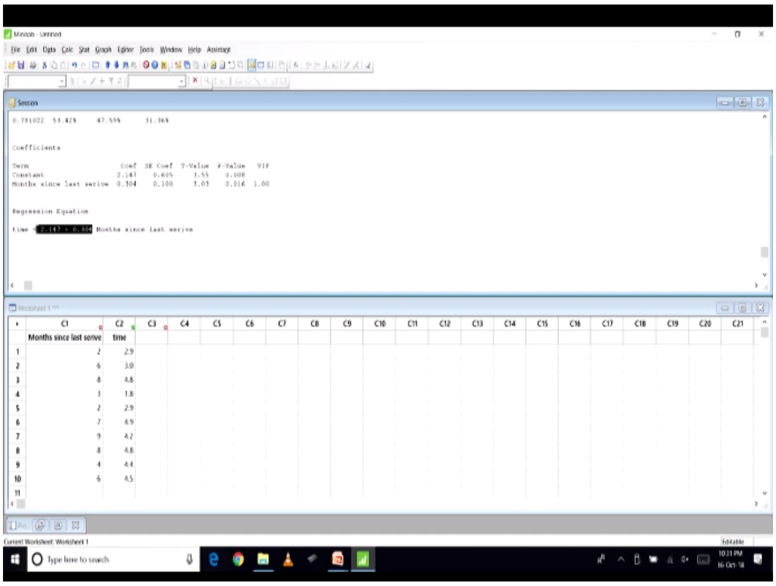
(Refer Slide Time: 26:41)



Let us solve this question. So when you solve this question, this is your adjusted r square. Now onwards do not look at r square value right, so adjusted r square is 47.59% and the intercept is this one. This is intercept and this is coefficient of independent variable months since last service right. So question is Y is=2.14+0.30 months since last service, so let us say if months since last service is 10, so just write 10 over here this will become 3, so Y value would be 2.44 is not it?

No, this is 3 right, so this would be 5.14 right, so that will be the value of Y. Now let us work out this example using Minitab and let us see whether you are getting this output or not right.

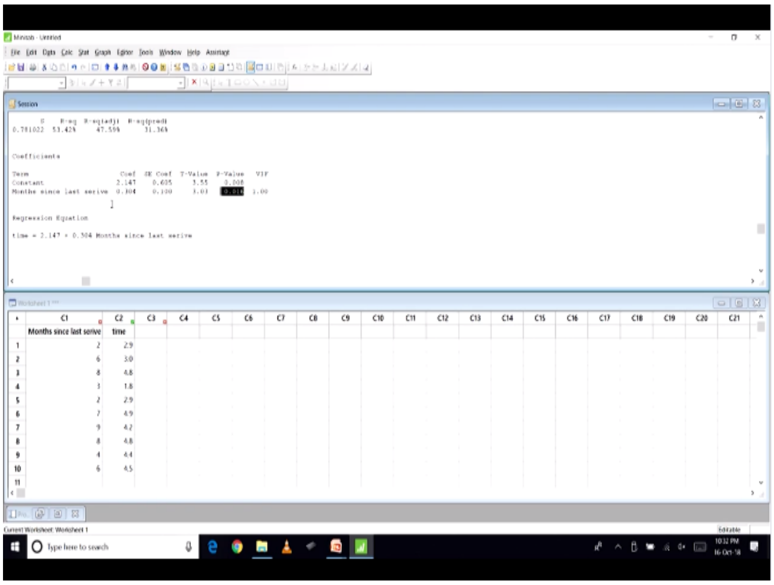
(Refer Slide Time: 27:48)



So let us look at Minitab software and try to solve this question. So we just deleted the output of previous question right. So months since last service it is okay no need to write everything. This is the repair time yeah time okay. So months since last service you just enter data, it was 2, 6, 8, 3, 2, 7, 9, 8, 4 so 10 data points. Then, 2.9 the repairing time, for the time being we are not writing type of maintenance work right.

We will take into account second independent variable as well, so this is how you should enter data for dependent and independent variable. Just go to stat, regression, fit regression model, so time is response, repair time is response and months since last service is continuous predictor variable right, so just click okay. So this is 2.14 and 0.30 right. This is what I told you which was there in this slide right, 2.14 and 0.30 right.

(Refer Slide Time: 29:44)



Let us look at whether the model is significant. In fact, we can look at P value of the F okay. So if you look at the ANOVA table anyway you just look at the P value of independent variable let us look at this. So this is your independent variable, months since last service. So this is less than 0.05 so will say that this is quite a significant independent variable right. So for the time being, let me stop here.

In the next class, we will consider the second independent variable which is categorical in nature. Thank you very much.