

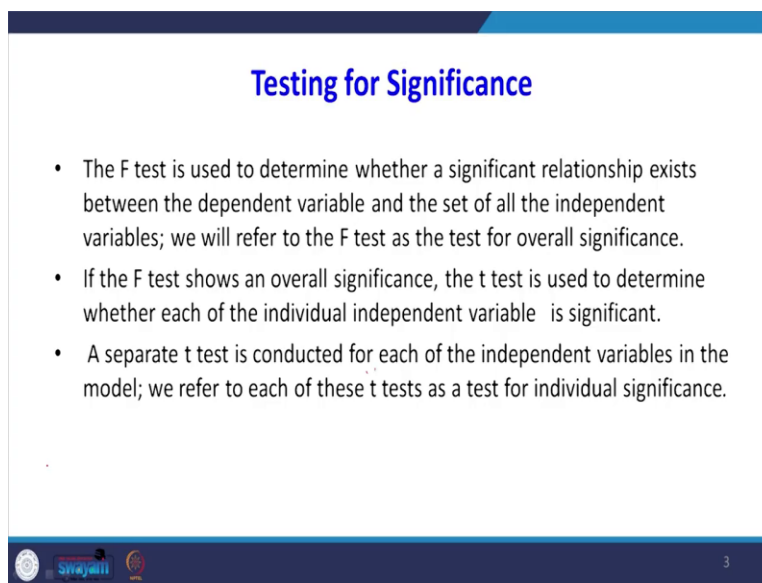
**Data Analytics with Python**  
**Prof. Ramesh Anbanandam**  
**Department of Management Studies**  
**Indian Institute of Technology – Roorkee**

**Lecture – 34**  
**Multiple Regression Model - II**

In the previous lecture we started multiple regression models. In the multiple regression model I have explain how to do a multiple regression model. What is the meaning of beta 0 beta 1 beta 2 and also explain what R squared and adjusted R square. In this lecture you are going to see how to do the significance test that means here also, like simple regression we are going to have some hypothesis about beta coefficient and beta 2 coefficients so on.

And we are going to test whether the beta 1 is equal to 0 are not equal to 0. So what you are going to do in this lecture is we are going to test the significance of regression model with the help of F test and t test and I am going to do a Python demo for your multiple regression.

**(Refer Slide Time: 01:13)**



**Testing for Significance**

- The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; we will refer to the F test as the test for overall significance.
- If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variable is significant.
- A separate t test is conducted for each of the independent variables in the model; we refer to each of these t tests as a test for individual significance.

Swayam

F test is used to determine whether a significant relationship exists between the dependent variable and the set of independent variables we will refer F test is the test of overall significance. I will show you where this F test is appearing in our Python output. If the F test shows and overall significance the t test is used to determine whether each of the individual independent variable is significant or not. A separate t test is conducted for each of the dependent

variable in the model. So we refer each of these, t-test as a test of individual significance. So, F test is used for testing the overall model of regression equation that t test is used to test for individual independent variables, whether they are significant are not.

**(Refer Slide Time: 02:08)**

**F Test**

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + \epsilon$$

The hypotheses for the  $F$  test involve the parameters of the multiple regression model.

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_a: \text{One or more of the parameters is not equal to zero}$$

4

The F test what is the null hypothesis, here null hypothesis beta 1 equal to beta 2 up to beta p equal to 0 when I accept null hypothesis, what is the meaning is for example for accept beta 1 equal to 0 there is no relation between x1 coefficient and the different variable y. If I accept to beta 2 equal to 0 there is no relation between x 2 and y variable obviously alternative hypothesis is one or more of the parameter is not equal to 0.

**(Refer Slide Time: 02:43)**

**F test significance**

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_a: \text{One or more of the parameters is not equal to zero}$$

TEST STATISTIC

$$F = \frac{MSR}{MSE}$$

REJECTION RULE

$p$ -value approach: Reject  $H_0$  if  $p\text{-value} \leq \alpha$

Critical value approach: Reject  $H_0$  if  $F \geq F_\alpha$

where  $F_\alpha$  is based on an  $F$  distribution with  $p$  degrees of freedom in the numerator and  $n - p - 1$  degrees of freedom in the denominator.

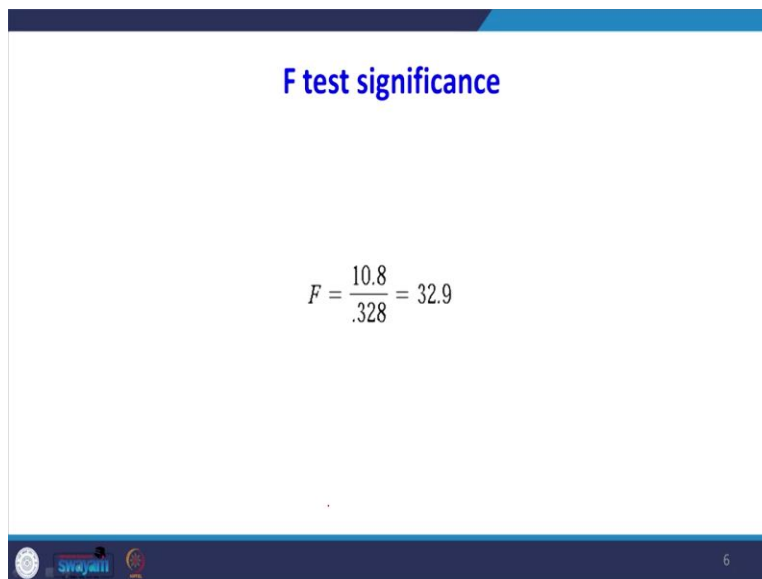
5

So how to find out the F statistic of statistics is MSR divided by MSE that is mean regression sum of square divided by mean error sum of square how we are getting this mean regression sum of square when you divide SSR divided by corresponding degrees of freedom here the degrees of freedom p divided by then will get MSR. MSE when you divide SSE divided by n - p - 1 where p is number of independent variable. Then we will get the mean error sum of square.

This hypothesis testing can be done by two way one is by p value approach and release by critical value approach. In the p value approach reject H 0 if the p value is less than or equal to alpha what will happen this is F test. This way it is a right skewed data. This is your alpha value will get F alpha corresponding this value can get it in table. What we have to do you have to find out the p-value. The p-value is lying on the rejection side you have to reject it.

Otherwise, if the F value is beyond the alpha value we have to reject it where F alpha is based on the F distribution with p degrees of freedom in the numerator and n - p - 1, degrees of freedom in the denominator.

**(Refer Slide Time: 04:18)**

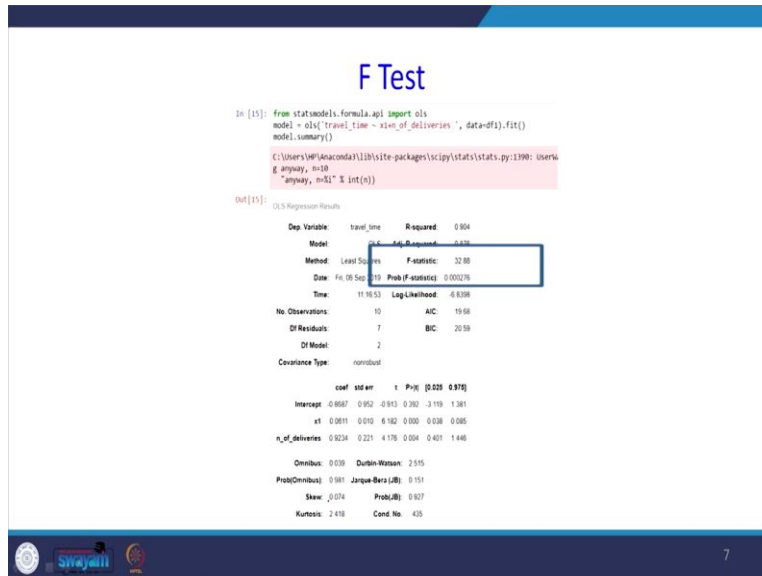


**F test significance**

$$F = \frac{10.8}{.328} = 32.9$$

So we got F equal to 10.8 divided by 0.328 equal to 32.9 I will show in another table this F value the previous slide.

**(Refer Slide Time: 04:34)**



F value is 32, so this value 32.8 that that is 32.9 approximately then we can see the easy here the p value probability of stat this p value is less than 4.05. when you say alpha is equal to 5.5 % then also we have to reject the null hypothesis.

**(Refer Slide Time: 05:00)**

### ANOVA table

| Source     | Sum of Squares | Degrees of Freedom | Mean Square                   | F                     |
|------------|----------------|--------------------|-------------------------------|-----------------------|
| Regression | SSR            | $p$                | $MSR = \frac{SSR}{p}$         | $F = \frac{MSR}{MSE}$ |
| Error      | SSE            | $n - p - 1$        | $MSE = \frac{SSE}{n - p - 1}$ |                       |
| Total      | SST            | $n - 1$            |                               |                       |

So this is anova regression table what is the sources of error? Error due to regression variable error total regression sum of square error sum of square total sum of square look at the degrees of freedom that is more important. There are 1 independent variable the degrees of freedom is 1 for regression sum of square. For TSS that is the total sum of square for that n -1, n is number of data set. So, when you subtract n - 1 - p that is why we are getting n - 1 - p is n - p - 1.

So, what is MSR is SSR divided by p MSE is SSE divided by n - p - 1. So, F equal to MSR divided by MSE this is the F value which you got anova output, when we introduce both variable into the model corresponding anova table for that two independent variable regression model is this one.

So, this is another table first you find out regression sum of square error of sum of square SST, p is the degrees of freedom n is the number of independent variable. For SST the degree of freedom is n -1. If you want to know the degrees of freedom for n -1 - p that is n - p - 1 MSR =SSR divided by p. MSE = SSE divided by n - p - 1 finally we are getting F value. So, this F value is nothing but your 32.88.

**(Refer Slide Time: 06:30)**

**t Test for individual significance**

For any parameter  $\beta_i$

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

TEST STATISTIC

$$t = \frac{b_i}{s_{b_i}}$$

REJECTION RULE

*p*-value approach: Reject  $H_0$  if *p*-value  $\leq \alpha$

Critical value approach: Reject  $H_0$  if  $t \leq -t_{\alpha/2}$  or if  $t \geq t_{\alpha/2}$

where  $t_{\alpha/2}$  is based on a *t* distribution with  $n - p - 1$  degrees of freedom.

Now will go for individual for each variable we will see that significant weather that individual variable is significant are not, for any parameter beta i H 0 equal to beta i equal to 0 H a beta i 0 equal to 0. The test statistics is bi divided by Sbi actually they should be this way bi - beta I divided by Sbi because we are assuming beta equal to 0 so the remaining is only bi divided by S bi this Sbi for ith independent variable we can see that is standard error.

When I go back you see this the standard error for x1 is this 0.010 the standard error for second independent variables 0.0221 that value we can get it this one. So for p value approach reject H 0 the p value is less than or equal to Alpha as usual. Critical value of approach reject H 0 t value is

below the lower tail or above the upper tail. Here the  $t$  alpha by 2 is based on the  $t$ -distribution with  $n - p - 1$  degrees of freedom ok we will do that one.

**(Refer Slide Time: 07:50)**

**t Test for individual significance**

$$b_1 = .061135 \quad s_{b_1} = .009888$$
$$b_2 = .9234 \quad s_{b_2} = .2211$$
$$t = .061135 / .009888 = 6.18$$
$$t = .9234 / .2211 = 4.18$$

10

So, beta 1 is 0.0611 where we got this one and going back,  $t$  test for individual significance from output of python model we get  $b_1$  equal to 0.061135  $b_2$  equal to 0.9234  $s_{b_1}$  equal to 0.009888  $s_{b_2}$  equal to 0.2211 where we are getting just see that  $b_1$ . So,  $b_1$  is this value 0.0611 the  $S_{b_1}$  is this 0.011 I am going back, see that is  $b_1$  so 0.0098 so it can be 0.01 because after rounding ok. The  $b_2$  is 0.92  $S_{b_2}$  is 0.221. So, what is the  $t$  formula of  $b_1$  divided by  $S_{b_1}$   $b_1$  is 0.061135 divided by 0.009888, 6.18 here  $t$  is 4.18 this also you can verify.

Where this for first variable it is 6.18 see that 6.18, for second variable this is 4.16. Look at the  $p$  value,  $p$  value for first variables 0.000 2nd variable is 0.00 by looking at the  $p$ -value itself without to reject the null hypothesis. When we reject null hypothesis beta 1 not equal to 0 that mean there is a relation between  $x_1$  and dependent variable  $y$  at the population level. Similarly for the second independent variable, also, we have to reject null hypothesis.

So, beta 2 not equal to 0 that means that there is a relation between  $x_2$  that is the second variable number of deliveries and dependent variable. This is the way to interpret the Python output of this multiple regression model. And going to give a demo for the multiple regression model. Ok students we have seen the theory behind the multiple regression model. We have

come to Python background. I have already prepared the code as shown the output suppose. You want to do it demo in our class are you want to someone in this course.

**(Video Start time: 10:15)**

Go to this kernel option restart and clear output so what will get it there is a restart and clear all output when you do this way to see that only the quotes will be there, there would not be any output. Suppose you want to show to others what is going to be the output of this quote you can do that way. So first we will import the necessary libraries `import Pandas as pd` from `stats models dot formula dot api` `import ols` so this library is used for doing regression analysis.

As you know that the pandas is used for reading a loading the files. From `stat model dot stat anova` `import anova underscore lm`, so this library is used to see the output of anova table for regression model `import matplotlib pyplot as plt` so this is used for plotting the figure. First I have stored my data set the file name is called tracking. I have loaded this data set into the object called `bf1` first will run the library then will see the what is the data set.

In this data set a when you look at this. The first one is the index column second one is the driving assignment third one is our independent variable that is the travel time. The next one is number of delivery third one is not the travel time. It is the distance travelled the `x1` means distance travelled the next independent variable is number of deliveries. The travel time is our dependent variable here.

Here what were going to do? First, we are going to see what is the relation between `x 1` and our dependent variable then we will see `x 2` the number of deliveries versus travel time. Then we will see both independent variable together. Then will see what is the effect on the dependent variable? First will do the scatter plot then we can understand the relationship that trend between is independent variable and the dependent variable. This one is `X1` is taken as a independent variable and `y` is taken as the different variable.

It seems to be there is some positive trend is there. Why this scatter plot is required if there is no relationship at all suppose the lines are in a horizontal manner that you need not do any regression analysis because there is no relation between this `x` and different variable. Now we

will take 2 variable then you got it. Now it is happening the green dot shows one variable and red dots shows another variable. Now, this is only second independent variable. Now will get the regression model the first regression model where we are going to consider only one independent variable. So that is model name is reg1 is equal to wireless.

The formula equal to travel underscore time is the dependent variable tilde symbol actually the first you to write the dependent variable tilde x1 in double quote data equal to df1. Because this tilde symbol even if you know the R programming, in R programming also similar Syntax will use that one. So, will you do this one you are getting the output of our regression model where only one independent variable is considered. So what is the first task is we have to construct the regression equation? What is regression equation y equal to  $1.2739 + 0.0678 \times 1$ .

Second one is where to locate the R square? R square is 0.664 if the R square is more than 50. It is considerably a good model even though it is 66 this is accepted. Then look at the F statistics, F statistics to 15.81 that look at the probability that is less than 0.05 as a whole model this regression model is acceptable. Remember that we are using only one independent variable within a look at the p value when the first independent variable 0.004 less than 0.05 when you look at an integer variable also this one variable is significant variable.

The next one what you are going to do? We are going to introduce both the independent variable together. Then we are going to see the impact on R square. So here I go to say that model is regression 2 where 2 is equal to wireless formula equal to travel time tilde x1 plus I am adding second independent. If there is third independent variable plus you have to add the third independent variable. Then fit 2 equal to reg2 fit.

So print fit 2 summary, look at this, first will come regression equation. Regression equation is y equal to  $-0.8687 + 0.0611 \times 1 + 0.9234 \times 2$  otherwise number of deliveries. You compare the R square with the previously say R square is 0.664 now when you use to independent variable, the R square is increased. So, the goodness of it the model is increased when we introduce more independent variable. There is another term adjusted R square this is adjusted for number of independent variables.



When you are in keep on introducing more number of independent variables you have to monitor the value of R squared and adjusted R square what will happen when you introduce more independent variables R square will always will increase but adjust R square it will initially start increase after certain point it start decreasing that point You should stop adding more independent variable that means the R square value when is decreasing that the new variable which have introduced into the model is not helping to explain the regression model instead of it is going to disturb the existing model that is that a new variable it is very noise variable.

Look at the F value F is 32.8 look at the probability value it is less than 0.05 as a whole model so we are going to reject null hypothesis. So, what this F statistics says F statistics is used to test overall significance of the regression model that both  $x_1$   $x_2$  by considering in this regression equation the model is valid. Then we will go for significance of each individual independent variable. There is  $x_1$  when you look at the p-value it is 0.000 that means we have to reject the null hypothesis of  $\beta_1$  equal to 0.

When you rejected it that means  $\beta_1$  is not equal to 0 then you say  $\beta_1$  is not equal to 0, even if the population over there is a relation between  $x_1$  and the y variable. Similarly look at the p value for the our second independent variable that is also less than 0.05 that means that the second variable also significant variable in our regression model. Sometime what will happen they may be different independent variables for some independent variable p value more than 0.05.

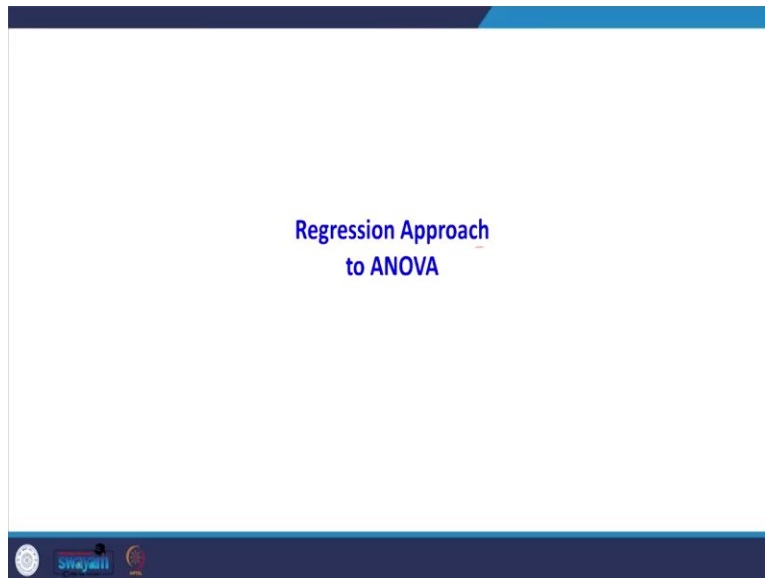
If it is more than 0.05 then you are writing regression model that corresponding independent variable has to be dropped. So meaning of the dropping is that with the help of sample data we can try regression equation by considering all independent variable but that cannot be generalized at the population level because certain variable cannot be significant at the population-level how to know that the variable is not significant we have to look at the p-value.

The p value is more than 0.05 we to accept null hypothesis, when we accept null hypothesis. that means  $\beta_1$  equal to 0 then there is no relation between that independent variable at the population. Other goodness of fit is we will see the what is the meaning of Durbin Watson in the

coming classes. Similarly there is one more measure to check the goodness of model AIC and BIC and we go for our Logistic regression, then I will explain what is the meaning and significance of AIC.

**(Video End Time: 19:02)**

**(Refer Slide Time: 19:03)**



So, far we have studied the regression analysis? Now with the help of regression I am going to tell you how to solve an anova problem. I have taken one sample problem that problem first time going to solve with the help of anova then with help of excel and good to solve it after that the same problem. I go to explain how to solve that anova problem with the help of regression analysis.

**(Refer Slide Time: 19:34)**

### Regression Approach to ANOVA

- Three different assembly methods, referred to as methods A, B, and C, have been proposed.
- Managers at Chemitech want to determine which assembly method can produce the greatest number of filtration systems per week

| A  | B  | C  |
|----|----|----|
| 58 | 58 | 48 |
| 64 | 69 | 57 |
| 55 | 71 | 59 |
| 66 | 64 | 47 |
| 67 | 68 | 49 |

The problem is like this what is happening in the three column is there A, B and C. A Represent one type of assembly method B represents another type of assembly method C represents the third type of assembly methods. If you follow method A; 58, 64, 55, 66, 67 represents number of product which are assembled per week; similarly for B under the column B 58 69 71 64 68 represents number of product assemble to per week. Here the product is filtration system. As a manager I want to know which method is producing the better result.

That means if I follow method A or B or C which one will produce are will help me to assemble more number of products. This is a typical anova problem. So in anova what we used to do generally the null hypothesis, The null hypothesis  $\mu A = \mu B = \mu C$  that means the mean of the product assemble through method A equal to mean of the product assembled by method B equal to the product obtained by method C.

Obviously the alternative hypothesis  $\mu A \neq \mu B \neq \mu C$ , the purpose of doing this is to identify which assembly method is more productive. In case if I accept null hypothesis all the three assembly methods are giving the same result I am not able to identify which method is better. In case if I reject my null hypothesis, I can clearly say which method is the better method which will give you more number of products assembled.

**(Refer Slide Time: 21:44)**

### ANOVA

Anova: Single Factor

| SUMMARY |       |     |         |          |  |
|---------|-------|-----|---------|----------|--|
| Groups  | Count | Sum | Average | Variance |  |
| A       | 5     | 310 | 62      | 27.5     |  |
| B       | 5     | 330 | 66      | 26.5     |  |
| C       | 5     | 260 | 52      | 31       |  |

| ANOVA               |     |    |          |          |          |          |
|---------------------|-----|----|----------|----------|----------|----------|
| Source of Variation | SS  | df | MS       | F        | P-value  | F crit   |
| Between Groups      | 520 | 2  | 260      | 9.176471 | 0.003818 | 3.885294 |
| Within Groups       | 340 | 12 | 28.33333 |          |          |          |
| Total               | 860 | 14 |          |          |          |          |

This I am going to solve with help of Excel enter the data in three column A column B column C. So go for data go for data analysis go for anova. Anova is a single factor because one way anova so the input range is I am selecting all this values, so labels in the First row yes, so when I say ok I am getting this output, what this mean is, when you look at the p-value here, it is 0.00382 it is less than 0.05. So I will be rejected my null hypothesis.

When I reject null hypothesis all the three assembly method not producing equal result, this was the output got it. How to interpret this output when you look at this p-value the p-value is less than 0.05 so I am rejected my null hypothesis.

**(Refer Slide Time: 22:48)**

### Dummy variables for the chemitech experiment

| A | B |  |
|---|---|--|
| 1 | 0 | Observation is associated with assembly method A |
| 0 | 1 | Observation is associated with assembly method B |
| 0 | 0 | Observation is associated with assembly method C |

Now, this is what I have got in the previous slide I am going to get with the help of regression analysis. The regression analysis I am going to use the concept of dummy variable because there are three assembly methods is there. Generally 3 - 1 number of dummy variables is required. How I am creating dummy variable I am taking say A this is A is one dummy variable for example B is another variable.

If I save 10 that represents the presence of 1 represents variable If you say 01 the presence of 1 represents on column B the represents the B variable see that the absence of 1 in both columns represent assembly method say that is why it is written here, see the 10 observation is associated with assembly line method A 01 represents the observation is associated with assembly method B 00 represents the observation is associated with assembly method C. So, I am going to do this modification then I am going to do regression analysis. So, with the help of regression I am going to explain here, but after sometime I go to explain to you. How to use Python. So, first now I will explain with help of Excel.

**(Refer Slide Time: 24:05)**

| SUMMARY OUTPUT |                       |             |                |         |         |                |           |
|----------------|-----------------------|-------------|----------------|---------|---------|----------------|-----------|
| 1              | SUMMARY OUTPUT        |             |                |         |         |                |           |
| 2              |                       |             |                |         |         |                |           |
| 3              | Regression Statistics |             |                |         |         |                |           |
| 4              | Multiple R            | 0.77759     |                |         |         |                |           |
| 5              | R Square              | 0.60465     |                |         |         |                |           |
| 6              | Adjusted R            | 0.53876     |                |         |         |                |           |
| 7              | Standard Error        | 5.32291     |                |         |         |                |           |
| 8              | Observations          | 15          |                |         |         |                |           |
| 9              |                       |             |                |         |         |                |           |
| 10             | ANOVA                 |             |                |         |         |                |           |
| 11             |                       | df          | SS             | MS      | F       | Significance F |           |
| 12             | Regression            | 2           | 520            | 260     | 9.17647 | 0.00382        |           |
| 13             | Residual              | 12          | 340            | 28.3333 |         |                |           |
| 14             | Total                 | 14          | 860            |         |         |                |           |
| 15             |                       |             |                |         |         |                |           |
| 16             |                       | Coefficient | Standard Error | t Stat  | P-value | Lower 95%      | Upper 95% |
| 17             | Intercept             | 52          | 2.38048        | 21.8444 | 5E-11   | 46.8134        | 57.2866   |
| 18             | A                     | 10          | 3.3665         | 2.97044 | 0.01169 | 2.68202        | 17.335    |
| 19             | B                     | 14          | 3.3665         | 4.15862 | 0.00133 | 6.66502        | 21.335    |

This is the given data set. So what I have done this is my after coding. For example see that upto 58 to 67 this column up to this much it represents A so I have written 10101010 see that here up to 1010 the presence of 1 represents assembly method A the absence of 1 represents a assembly method B. Similarly I have to type the remaining B values see that here 01010101. So this portion represents look at this the presence of 1 represents method B.

The last one is this portions I have taken 00 on both columns that means absence of one and both the method represents the Assembly method C. Now this one I going to do the regression analysis go for data analysis go for regression. Here the y value is this one x values here there are 2 dummy variable this one, when you run it you are getting this output you see that. When you look at the p-value here here also, you are getting 0.0038 that means here also you are rejecting the null hypothesis. I will explain how to interpret the coefficient of 52 10 14 in coming slides.

**(Refer Slide Time: 25:52)**

**Dummy variables for the chemitech experiment**

$E(y) = \text{Expected value of the number of units produced per week}$   
 $= \beta_0 + \beta_1 A + \beta_2 B$

- If we are interested in the expected value of the number of units assembled per week for an employee who uses method C, our procedure for assigning numerical values to the dummy variables would result in setting  $A = B = 0$ .
- The multiple regression equation then reduces to

$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$

Q

Swayam

Expected value of y is equal to expected value of number of units produced per week. This is the regression equation  $\beta_0 + \beta_1 A + \beta_2 B$  if you are interested in the expected value of the number of units assembled per week for an employee who uses method C our procedure for assigning numerical value to the dummy variable result in setting A equal to B equal to 0, suppose if you want to know the answer for assembly method C you to substitute A equal to 0 B equal to 0. When you substitute A equal to 0 B equal to 0 the expected value of y is nothing but your  $\beta_0$ .

**(Refer Slide Time: 26:35)**

## Dummy variables for the chemitech experiment

- For method A the values of the dummy variables are A = 1 and B = 0, and

$$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

- For method B we set A = 0 and B = 1, and

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

In case for method the value of the dummy variable is equal to 1 b equal to 0 when you substitute in the regression equation beta 0 + beta 1 because A values is 1 B value 0 so beta 0 + beta 1. If I want to know the expected value of assembly method B we have to set A equal to 0 and B equal to 1 when you substituted regression equation You are getting beta 0 + beta 2.

**(Refer Slide Time: 27:07)**

| SUMMARY OUTPUT        |              |                |          |          |                |             |             |             |
|-----------------------|--------------|----------------|----------|----------|----------------|-------------|-------------|-------------|
| Regression Statistics |              |                |          |          |                |             |             |             |
| Multiple R            | 0.777593186  |                |          |          |                |             |             |             |
| R Square              | 0.604651163  |                |          |          |                |             |             |             |
| Adjusted R Square     | 0.53875969   |                |          |          |                |             |             |             |
| Standard Error        | 5.322906474  |                |          |          |                |             |             |             |
| Observations          | 15           |                |          |          |                |             |             |             |
| ANOVA                 |              |                |          |          |                |             |             |             |
|                       | df           | SS             | MS       | F        | Significance F |             |             |             |
| Regression            | 2            | 520            | 260      | 9.176471 | 0.003818412    |             |             |             |
| Residual              | 12           | 340            | 28.33333 |          |                |             |             |             |
| Total                 | 14           | 860            |          |          |                |             |             |             |
|                       | Coefficients | Standard Error | t Stat   | P-value  | Lower 95%      | Upper 95%   | Lower 95.0% | Upper 95.0% |
| Intercept             | 52           | 2.380476143    | 21.84437 | 4.97E-11 | 46.81338804    | 57.18661196 | 46.81338804 | 57.18661196 |
| A                     | 10           | 3.366501646    | 2.970443 | 0.011692 | 2.665023022    | 17.33497698 | 2.665023022 | 17.33497698 |
| B                     | 14           | 3.366501646    | 4.15862  | 0.001326 | 6.665023022    | 21.33497698 | 6.665023022 | 21.33497698 |

Now what we got it, when you look at this coefficients the intercepts is 52 that is your beta 0. This is beta 1 for coefficient of A this is beta 2 this is coefficient of B.

**(Refer Slide Time: 27:23)**

### Estimation of E(y)

- $b_0 = 52$
- $b_1 = 10$
- $b_2 = 14$

| Assembly Method | Estimation of E(y)         |
|-----------------|----------------------------|
| A               | $b_0 + b_1 = 52 + 10 = 62$ |
| B               | $b_0 + b_2 = 52 + 14 = 66$ |
| C               | 52                         |

What will happen see beta 0 is 52 beta 1 equal to 10 B2 equal to 40 if you want know the estimated value of y for assembly method A you have to refer b 0 + b 1 how we got b 0 b 1 look at this beta 0 + beta 1 now beta 0 52, b 1 is 10 so totally 62. If you want to know the estimated value of y for assembly with B it is 52 + 14 how we got 52 + 14 using this equation beta 0 is 52 Beta 2 is 14, what is beta 2 look at this Beta 2 is 14. So when you substitute here we are getting 66. If you want to know the expected mean of methods C you have substitute A equal to 0 B equal to 0 you get only beta 0, beta 0 is estimate with help of b 0 that values 52 .

**(Refer Slide Time: 28:24)**

### Testing the significance

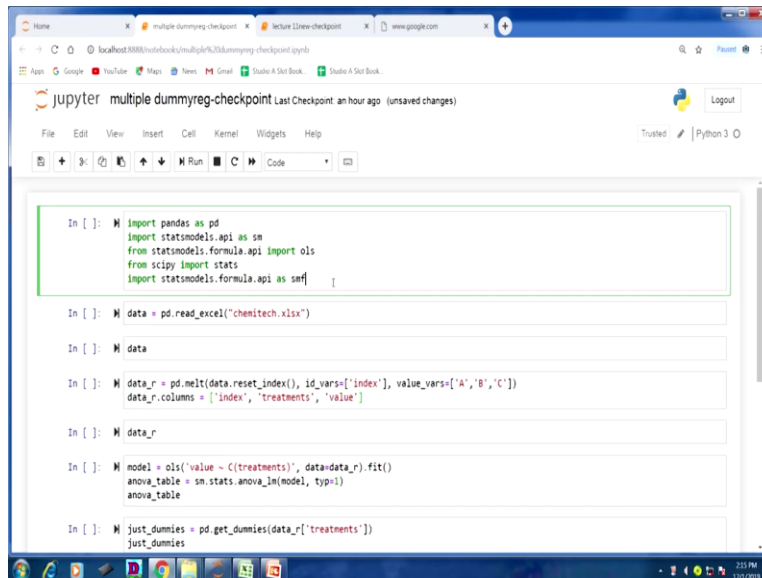
$$H_0: \beta_1 = \beta_2 = 0$$

Then we can go for significance test beta 1 equal to beta 2 equal to 0 this we have seen already we can go for t test or F test. In the F test if the value is less than 0.05 then we can say both the



variables are significant. What happened here when you do with the help of Excel you see that the p value of F see the p-value is less than 0.05 we can see the regression coefficient A and B is significant. So far I have done with the help of Excel. Now I am going to do the same problem python.

**(Refer Slide Time: 29:01)**

A screenshot of a Jupyter Notebook interface. The browser address bar shows 'localhost:8888/notebooks/multiple-dummyreg-checkpoint'. The notebook title is 'multiple dummyreg-checkpoint'. The code in the cells is as follows:

```
In [ ]: import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy import stats
import statsmodels.formula.api as smf

In [ ]: data = pd.read_excel("chemitech.xlsx")

In [ ]: data

In [ ]: data_r = pd.melt(data.reset_index(), id_vars=['index'], value_vars=['A','B','C'])
data_r.columns = ['index', 'treatments', 'value']

In [ ]: data_r

In [ ]: model = ols('value ~ C(treatments)', data=data_r).fit()
anova_table = sm.stats.anova_lm(model, typ=1)
anova_table

In [ ]: just_dummies = pd.get_dummies(data_r['treatments'])
just_dummies
```

First will import the necessary files like import Pandas as pd import stats models dot api dot sm from stat model dot com dot api import wireless form scipy import stats import stats model dot formula dot api as smf the data is stored in a file called chemitech. So this is my data set so far this data set and going to do an anova after doing anova I go to check the result the same problem I run it with help of regression analysis.

For doing regression analysis I go to use the concept called dummy variable. So, first I will run this given data set anova so for that purpose. I am converting this data set into this form. What is that form? That I will show in the data underscore you see that all the treatments are in one column all the values are in one column model equal to wireless. Value is our dependent variable Tilde see the treatment is my dependent variable.

The command for regression analysis from the regression analysis I am going to get the anova table the anova table is this one? You see that the p-value 0.0038 when I am doing the **the** same problem with the help of Excel also because the same result. So, what we are concluding here for

all the means or not equal so we are rejecting our null hypothesis. Now this problem we are going to do with the help of regression analysis by using the concept called dummy variables that one treatment that is assembly method with 3 variables ABC so that I am going to convert into dummy variables.

You look at this ABC is the presence of 1 represents A the presence of 1 here represents B the C column the presence of 1 represents C because in the treatment of three levels, we need only two dummy variables. So we are going to drop the column C then we are going to use the two column A and B. In excel also when I am solving meet you seen this kind of data. So what I am going to do I am going to drop this column C then I am going to add this dummy variables into the filename called step underscore 1.

This one see that now the file is changed. Now the value is taken as it is only the column A and B is maintained. So for this dataset going to do the regression analysis, so the results is equal to smf dot ols step underscore 1 the value is my different variable sm dot add underscore constant step underscore 1 A, B is my independent variable then I go to get the regression output. Now look at this regression output. Look at this probability that the p-value 0.00382 here also we are rejecting our null hypothesis.

Then look at the constant value constant is 52 that is b 0, b 1 is 10 and b 2 is 14 this value is taken for interpreting the output. Now look at the variable A and B both are the p values less than 0.05 both variables are significant the value of b0 b1 b2 can be used for interpreting as I explained in my slide for interpreting the output. In this class we have seen how to do the significance test for multiple regression model.

That significance test we have done with the help of two test one is F test and t test. Ftest is used to test the overall significance of the regression model, t test is used to check the individual significance of each independent variable. After that I have taken a sample problem. Then I explain the given a demo how to do the multiple regression, then we have interpreter the output of multiple regression model.

In the next class going to do another regression model that is where the independent variable that is categorical independent variable. So far what we have done that one dependent variable and independent variable both are continuous. There may be situation where that independent variable is categorical variable. For example gender is a categorical variable that can have only two option male or female in that case you to do some adjustment in existing; our regression model. How to do that one that you will see in the next class, thank you very much for listening.