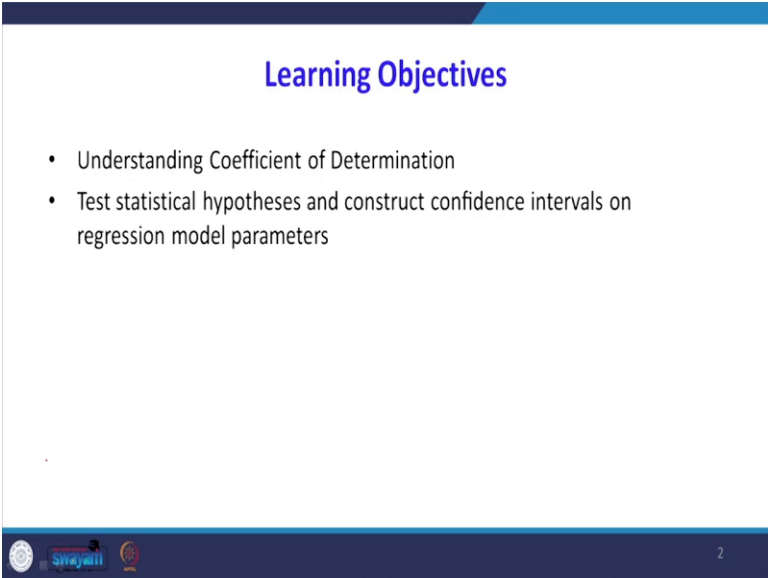


**Data Analytics with Python**  
**Prof. Ramesh Anbanandam**  
**Department of Management Studies**  
**Indian Institute of Technology – Roorkee**

**Lecture – 30**  
**Linear Regression – III**

Dear students in the previous class I would take in a sample example I have explained to construct how to construct a regression equation.

**(Refer Slide Time: 00:37)**



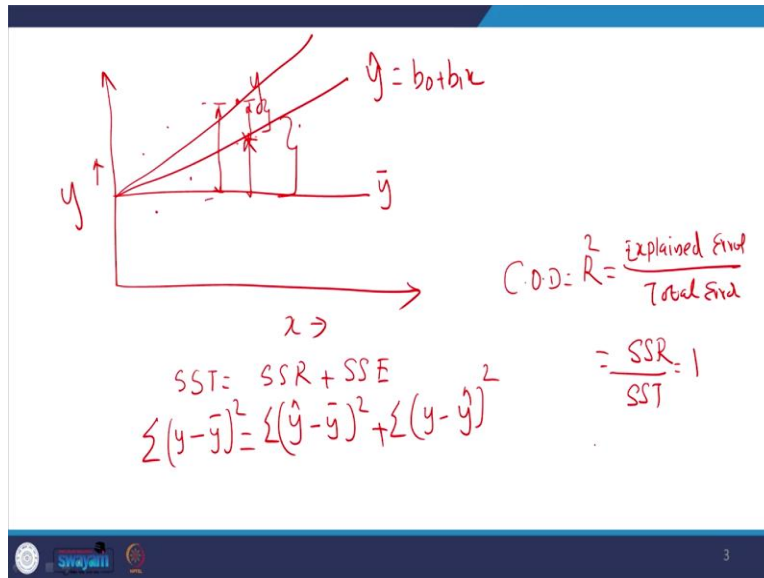
**Learning Objectives**

- Understanding Coefficient of Determination
- Test statistical hypotheses and construct confidence intervals on regression model parameters

2

In this class I will explain what is the meaning of coefficient of determination and test statistical hypothesis and construct confidence interval and regression model parameters, there are one parameter is the 'b' the coefficient of x is one parameter.

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



Now I will explain what is the goodness of fit that is coefficient of determination  $r^2$  assume that why value assume that there is no independent variable the easiest way for prediction is mean of the  $y$  what is the meaning of there is no independent variable. Suppose I have demand for say first eleven month I want to predict in 12th month what is the demand of a particular product, the easiest way is you find the mean of the previous data.

So that will be used as the mean for the next data so without considering any independent variable suppose if there is no independent variable so the actual point is say this is the  $y$  is the actual value. So, the one way to predict without any independent variable is say  $\bar{y}$  but I know there is a one independent variable what is this much this is actual this is predicted, so this much is my error what is this error this point to this point this is error.

What is the error total error we can say total error actual - predictor. Suppose if I know one independent variable that I have as I am assuming that is affecting my dependent variable then that regression equation is like this, so this I am writing  $\hat{y} = b_0 + b_1 x$ , now what has happened now this much distance because this point so this much distance is see total error is this much, so this much error is you could draw it this way.

So here what is yes this much portion this much portion I am able to explain with the help of regression independent variable. So, total error is this point to this point so this much error with

the help of independent variable  $x$  I am able to explain. So, the remaining error this one's this much distance is unexplained the error. So, what I am saying this is only one point there is no linear relationship like that there are different  $y$  values may be why here  $y$  may be here  $y$  may be here maybe here maybe here.

So if I find the total sum of square there is a total error then nothing but  $y$  SST . So, SST equal to  $SST + SSE$  what is the SST total sum of square what is the SSR a regression sum of square what is SSE error sum of square. So, now what is the logic behind this is the total error is this point to this point total error I am splitting that error due to how much error we are able to explain with the help of this independent variable  $x$  that is SSR, so the remaining portions that is the which is in the bracket that is unexplained error.

You when you look at this there will be a connection with on over in ANOVA what do I written the same thing you know what you written SST equal to SS treatment + SSE , your treatment is nothing but your independent variable . Now we will find out what is the formula this so what is SST so this point is  $y$  the total error  $\sum (y - \bar{y})^2$  , that is SST equal to so this much portion what is the error is  $\sum (\hat{y} - \bar{y})^2 + SSE$  unexplained error.

What is unexplained error  $\sum (y - \hat{y})^2$  ,so what has happened the total error is the regression sum of square + error sum of square. Here I want to predict the coefficient of determination, the coefficient of determination referred as  $r$  square is nothing but explain the error divided by total error what is explained error explained is nothing but SSR that means that much error we are able to explain because that much error is due to this independent variable  $x$  what is a total error this is SST total sum of square.

There is a two possibility of this  $r$  square is it cannot be more than 1, if it is 1 what is the meaning the total SST the numerator also SST denominator also SST so what we are saying this point this line pass through that point. If it is less than 1 so what is happening SSE error is smaller SST is bigger, if equal to 1 both are same. So, the upper limit of the  $r$  square is 1 the lower limit is 0 to 1 so 0 to 1 is the interval for  $r$  square.

**(Refer Slide Time: 07:32)**

### Coefficient of Determination

- Relationship Among SST, SSR, SSE

$$SST = SSR + SSE$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SS_{yy} = \left( \frac{SS_{xy}^2}{SS_{xx}} \right) + \left( SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}} \right)$$

where:

SST = total sum of squares  
 SSR = sum of squares due to regression  
 SSE = sum of squares due to error

We will see this one yes, see the relationship among SST, SSR and SSE. So, SST you see  $\sum (y - \bar{y})^2$  equal to SSR  $\sum (\hat{y} - \bar{y})^2$  for SSE  $\sum (y - \hat{y})^2$  remember the previous also I was showing this  $SS_{yy}$  that that  $SS_{yy}$  is nothing but  $\sum (y - \bar{y})^2$  so, this is the a very handy formula if you are using calculator and a very short cut very quickly you can get the answer for what is SST, SSR and SSE from that you can easily find out r square nothing but SSR divided by SST.

(Refer Slide Time: 07:58)

### Coefficient of Determination

- The coefficient of determination is:

$$r^2 = SSR/SST$$

where:

SSR = sum of squares due to regression  
 SST = total sum of squares

R square is  $SSR / SST$ , SSR is sum of square due to regression SST is total sum of square.

(Refer Slide Time: 08:07)

## Jupyter code

```
In [5]: t= tbl['TV Ads']
        c= tbl['car Sold']
```

```
In [8]: import statsmodels.api as sm
        t = sm.add_constant(t)
        model1 = sm.OLS(c,t)
        result1 = model1.fit()
        print(result1.summary())
```

```

=====
OLS Regression Results
=====
Dep. Variable:      car Sold      R-squared:      0.877
Model:              OLS          Adj. R-squared:  0.876
Method:             Least Squares
Date:               Fri, 30 Aug 2019    Prob (F-statistic): 0.0190
Time:               08:31:20          Log-Likelihood:  -9.6687
No. Observations:    5              AIC:              23.34
DF Residuals:        3              BIC:              22.58
DF Model:            1
Covariance Type:     nonrobust
=====
                    coef    std err          t      P>|t|      [0.025    0.975]
-----
const            10.0000     2.366     4.226     0.024     2.469    17.531
TV Ads             5.0000     1.000     4.629     0.019     1.563     8.437
=====
Omnibus:          nan    Durbin-Watson:      1.214
Prob(Omnibus):    nan    Jarque-Bera (JB):    0.674
Skew:             0.256    Prob(JB):           0.714
Kurtosis:         1.276    Cond. No.           6.33
=====
```

You see that the previous also I was saying what is the meaning of this r square. So, we are getting in our problem we are getting 0.87 I will show you how it has come yes the coefficient of determination formula is  $r^2 = \text{SSR} / \text{SST}$ , so SSR in our problem is 100 SST is 114 so I can show you this how this SSR 100 you see that SSR.

(Refer Slide Time: 08:42)

## Coefficient of Determination

$$r^2 = \text{SSR} / \text{SST} = 100 / 114 = .8772$$

The regression relationship is very strong; 88% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

So how what is the formula for SSR we have seen previously how we are getting SSR you see that SSR, is nothing but see  $\sum (\hat{y} - \bar{y})^2$  square first you have to find out the regression equation in that when you substitute first value of x you will get  $\hat{y}_1$ ,  $\hat{y}_1$  hat is when substitute the value of x into that then  $y - \bar{y}$  whole square then when you substitute x equal to 2 they will get  $\hat{y}_2$  so then  $y - \bar{y}$  whole square when you sum that one that is nothing but your SSR.

SST is  $y - y$  - it is a numerator of that variance of  $y - y$  bar whole square so from that you can find out SST. So, in this our problem it is 100 order by 114 now what is the meaning of this r square so the meaning of r square is as we know that it is 0 to 1 the regression relationship is very strong what is the meaning is 88% of the variability in the number of cars sold can be explained by the linear relation between the number of TV ads and the number of cars sold.

So what is meaning that 87 it is 87.7, 88% of the variability of  $y$  can be explained by the help of this independent variable there is a remaining 13% that we are not able to explain that may be due to two reasons one is we might all miss you that some other independent variable there may be some other variable that affects the car sales. Another reason is that we have fixed a linear regression but the actual data may follow non linear regressions so that is why we are not getting exactly 1.

In Python output you see that when you see the r square is the 0.77 this is r square is 0.77 that is the meaning of that is 87.7% of the variability of car sold can be explained with the help of number of TV ads that is our independent variable.

**(Refer Slide Time: 10:56)**

### Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1) \sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$

$$\hat{y} = b_0 + b_1x$$

where:

$b_1$  = the slope of the estimated regression equation

From r square we have to find out the r that is a correlation coefficient. So, the sample correlation coefficient  $r_{xy}$  equal to sign of  $b_1$  in our problem it is the sign of  $b_1$  is positive root

of coefficient of determination  $r^2$  so sign of  $b_1$  into  $r^2$ . So, what is this  $b_1$  is that the slope of the regression equation.

(Refer Slide Time: 11:23)

### Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$

The sign of  $b_1$  in the equation  $\hat{y} = 10 + 5x$  is "+".

$$r_{xy} = +\sqrt{.8772}$$
$$r_{xy} = +.9366$$

In our problem it is  $y$  equal to  $10 + 5x$  so the sign is + the root of 0.8772 is this is your correlation coefficient and remember that the range of correlation coefficient is  $-1$  to  $+1$  but the range of  $r^2$  is 0 to 1. Here in the correlation coefficient if it is a  $-1$  it is a perfectly negative correlation if it is a  $+1$  it is perfectly positive correlation if it is 0 there is no correlation. In the context of  $r^2$  if it is 1 it is a perfect model that means all variability of  $y$  can be explained with the help of independent variable. If it is 0 there is no relationship between  $x$  and  $y$ .

(Refer Slide Time: 12:13)

### Assumptions About the Error Term $e$

1. The error  $e$  is a random variable with mean of zero.
2. The variance of  $e$ , denoted by  $e^2$ , is the same for all values of the independent variable.
3. The values of  $e$  are independent.
4. The error  $e$  is a normally distributed random variable.

Another important point is assumptions about the error term  $e$ , here the two tests the goodness of the model not only  $r$  square is important you have to plot the error term. When you look at the error term we have to look at the behavior of that error is nothing but actual minus predicted value, so what are the assumption of the error term the error ' $e$ ' is he a random variable with mean equal to 0, so the error has to be appear in a random manner where the sum of positive error should be equal to sum of negative errors so that sum will be 0.

The variance of  $e$  denoted by  $e^2$  is the same for all values of the independent variable, so that a concept called a homoscedasticity what is the meaning of that one is if there are many  $x_1$  say  $x_2$   $x_3$  independent variable these variance of  $x_1$  variance of  $x_2$  variance of  $x_3$  should be same then only there is a meaningful comparison otherwise the variance of the error should be the same then only there is a meaningful comparison.

And the value of  $e$  is independent another important there should not be any pattern in the error term sometime what will happen when you plot the error term sometimes there is an increasing trend sometimes there may be in decreasing trend this kind of, this kind of pattern is not allowed the error term has to be distributed randomly. And the another point is the error  $e$  is normally distributed random variable now testing for significance.

**(Refer Slide Time: 14:03)**

**Testing for Significance**

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of  $\beta_1$  is zero.
- Two tests are commonly used:

**t Test** and **F Test**

- Both the  $t$  test and  $F$  test require an estimate of  $s^2$ , the variance of  $e$  in the regression model.

12

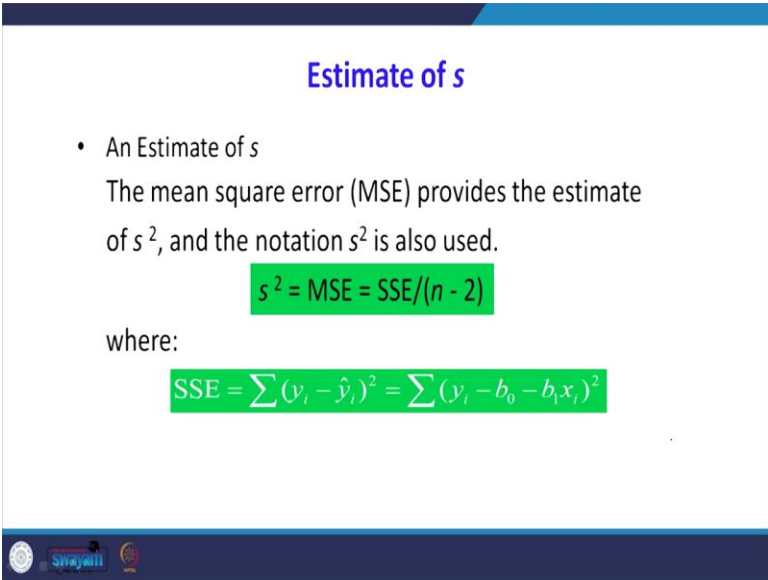


So far as I told you in the beginning of the class whatever regression equations and the goodness of model which you have tested only for the sample data what is the sample data  $y$  equal to  $b_0 + b_1 x$  so this capital  $Y$  equal to  $\beta_0 + \beta_1 X$  whatever we have know done is only for the sample. Now we are going to see whether this model is valid even at the population level for that purpose we are going to do some assumption we will see what is that assumption that is a hypothesis?

To test for a significant regression relationship we must conduct a hypothesis test to determine whether the value of  $\beta_1$  is 0. What will happen if the  $\beta_1$  is 0 there is no relation between  $x$  and  $y$  at the population level. But there is a possibility that there may be a relation between  $x$  and  $y$  at the sample data it is not necessary that even at the population level there will be a relation between  $x$  and  $y$ . so, that testing can be done by two methods one is a t-test another one is F test.

Both t-test and F tests require an estimation of  $S$  square.  $S$  square is called the variance of the error otherwise if you say  $S$  it is the standard error the variance of  $e$  in the regression model.

**(Refer Slide Time: 15:22)**



**Estimate of  $s$**

- An Estimate of  $s$   
The mean square error (MSE) provides the estimate of  $s^2$ , and the notation  $s^2$  is also used.

$$s^2 = \text{MSE} = \text{SSE} / (n - 2)$$

where:

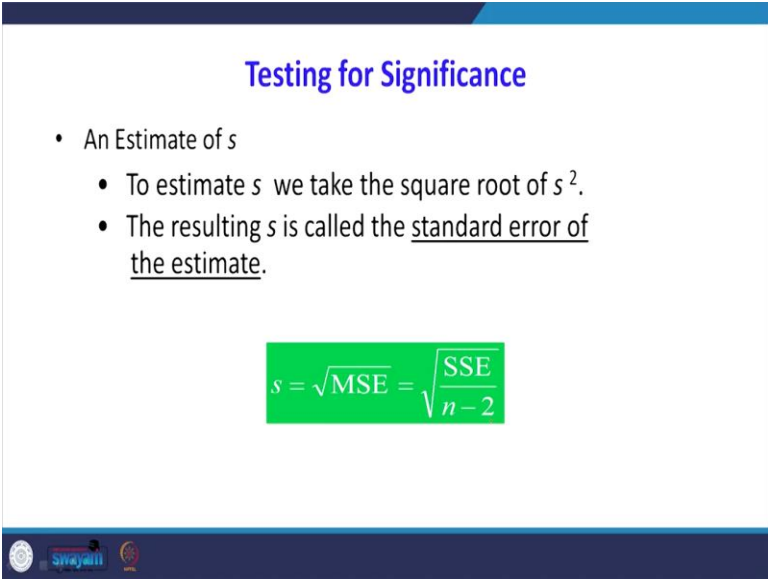
$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

What is the estimation of the standard error suppose in years normal data set if there are two data set data set 1 and 2, 1 is having lesser variance than the other one so the first data it is more homogeneous in the same way when you do in regression model suppose there are two model is there model 1 model 2 for the model in which there is a lesser standard error that is a more

suitable model. So, the mean square error provides the estimate of  $S^2$  and the notation  $S^2$  is used so  $s^2$  is nothing but MSE mean squared error.

We know that how we got to MSE, MSE is SSE divided by  $n - 2$  here what the degrees of freedom  $n - 2$  so the logic is  $n - 1 - K$  there is a logic of degrees of freedom.  $K$  is number of independent variable in this we are having only one independent variable we know that already the degrees of freedom is  $n - 1$ , so  $n - 1 - K$  will be  $n - 2$  and SSE also you see that we can find out that formula which I in the beginning of the class which I am saying  $y_i - \hat{y}$  whole square. The  $\hat{y}$  you can substitute  $b_0$  actually it is  $b_0 + b_1 x$  when you bring - inside  $b_0$ ,  $- b_0 - b_1 x$ .

(Refer Slide Time: 16:51)



**Testing for Significance**

- An Estimate of  $s$ 
  - To estimate  $s$  we take the square root of  $s^2$ .
  - The resulting  $s$  is called the standard error of the estimate.

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n - 2}}$$

The term  $S$  we make the square root of  $s^2$  the resulting is called standard at other term. So, when you take the square root of this mean squared error so you will get the standard error see in our problem I will go back I will see what is the standard error here where it is standard error as I told you by using shortcut method you can use SSE divided by  $n - 2$  so that is MSE so  $S_{xy} - S_{xy} S_{xy} S_{xy}$  whole square by  $S_{xx}$  divided by  $n - 2$  that is the standard error of the estimate. So you have to take the square root of that then you look at the standard error.

(Refer Slide Time: 17:37)

## Testing for Significance: t Test

- Hypotheses

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- Test Statistic

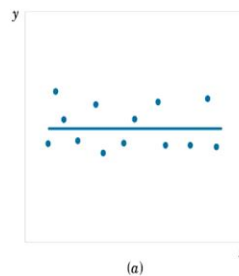
$$t = \frac{b_1}{s_{b_1}}$$

Now you go for hypothesis testing so what is the hypothesis testing beta 1 equal to 0 that means that there is no relation between x and y. Alternate hypothesis is beta 1  $\neq$  0 the test statistic is  $b_1 - \beta_1$  divided by  $s_{b_1}$   $s_{b_1}$  is the standard error for the coefficient of b. So, since beta 1 we are assuming 0 it is simply  $b_1$  divided by  $s_{b_1}$ .

(Refer Slide Time: 18:06)

### Case 1

$$H_0: \beta_1 = 0$$



In this case hypothesis is not rejected

What is the meaning of  $\beta_1$  equal to 0 that means there is no relation between x and y. If it is when you plot the data in this case hypothesis is accepted because  $\beta_1$  equal to 0 now what is happening the  $\beta_1$  is not equal to 0 you see for this kind of data set so there is some relation between x and y see in this case the hypothesis is rejected so we are saying  $\beta_1 \neq 0$ .

(Refer Slide Time: 18:30)

## The Standard Deviation of the Regression Slope

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

$$s_{b_1} = \frac{s_\epsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\epsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

$s_{b_1}$  = Estimate of the standard error of the least squares slope

$$s_\epsilon = \sqrt{\frac{SSE}{n - 2}} = \text{Sample standard error of the estimate}$$

So, this is very important how to find out the standard error of the coefficient of x that is the  $b_1$ , so  $s_{b_1}$  is  $s_\epsilon$  that is a standard error will be divided by root of  $\sum (x - \bar{x})^2$ , you see intuitively the total error that is  $s_\epsilon$  then we are dividing how much error is due to this independent variable, so total error divided away portions of error from independent variable x. So, that will give you  $s_{b_1}$ .

**(Refer Slide Time: 19:04)**

## Testing for Significance: t Test

### ■ Rejection Rule

Reject  $H_0$  if  $p\text{-value} \leq \alpha$   
or  $t \leq -t_{\alpha/2}$  or  $t \geq t_{\alpha/2}$

where:

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

What is the rejection rule reject  $H_0$  if the p-value is less than or equal to alpha we have seen many times this one so what will happen this one if the p value, the p value is less than alpha the p-value is less than that you are to reject it otherwise accept it, where  $t_{\alpha/2}$  is the two-tailed test because we are writing  $\beta_0 = 0$ ,  $\beta_1 = 0$ ,  $\beta_0 = 0$  and when you look at the t table you were to see  $n - 2$  degrees of freedom.

(Refer Slide Time: 19:42)

### Testing for Significance: t Test

1. Determine the hypotheses. 
$$\begin{array}{l} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{array}$$
2. Specify the level of significance.  $\alpha = .05$
3. Select the test statistic. 
$$t = \frac{b_1}{s_{b_1}}$$
4. State the rejection rule. Reject  $H_0$  if  $p\text{-value} \leq .05$   
or  $|t| > 3.182$  (with 3 degrees of freedom)

So, first determine the hypothesis beta 1 equal to 0 beta 1 not equal to 0 specify the significant level alpha equal to 5% select the test statistics b 1 by Sb1 state the rejection role reject H 0 if the p-value is less than or equal to 0.05 otherwise the t is greater than 3.182 when n - 2 degrees of freedom there are 5 data so 5 - 2 is the 3 degrees of freedom.

(Refer Slide Time: 20:16)

### Testing for Significance: t Test

5. Compute the value of the test statistic.
$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{5}{1.08} = 4.63$$
6. Determine whether to reject  $H_0$ .  
 $t = 4.541$  provides an area of .01 in the upper tail. Hence, the  $p\text{-value}$  is less than .02. (Also,  $t = 4.63 > 3.182$ .) We can reject  $H_0$ .

So, this was the compute the value of the test statistics so 5 data by the Sb 1 is 1.08 we get 4.63 determine whether to reject H 0 because t equal to 4.5 form provides in the area of 0.01 in the upper tail here is the p-value is less then we will see how the got the p-value less than 0.02 so P

is greater than 4.63 we can reject null hypothesis and we reject null hypothesis what we are conclude there is a relation between x and y.

(Refer Slide Time: 20:51)

### Hypothesis Tests for the Slope of the Regression Model

$H_0: \beta_1 = 0$	$t = \frac{b_1 - \beta_1}{S_b}$	
$H_1: \beta_1 \neq 0$	where: $S_b = \frac{S_e}{\sqrt{SS_{XX}}}$	
$H_0: \beta_1 \leq 0$	$S_e = \sqrt{\frac{SSE}{n-2}}$	
$H_1: \beta_1 > 0$	$SS_{XX} = \sum X^2 - \frac{(\sum X)^2}{n}$	
$H_0: \beta_1 \geq 0$	$\beta_1$ = the hypothesized slope	
$H_1: \beta_1 < 0$	$df = n - 2$	

This was the, here also  $H_0$ : beta 1 equal to 0 ,  $H_1$ : beta 1  $\neq$  0, 2 tail test, this is the right tailed test this is a left tailed test this was the formula for finding  $t = (b_1 - \beta_1) / S_b$  to find  $S_b = S_e$  is divided by root of  $SS_{XX}$  that is this is a form of  $S_e = \text{root of } (SSE \text{ divided by } (n - 2))$ ,  $SS_{XX}$  is  $\sum (X)^2 - ((\sum X)^2 / n)$  remember it is  $n - 2$  degrees of freedom.

(Refer Slide Time: 21:19)

### Confidence Interval for $\beta_1$

- We can use a 95% confidence interval for  $\beta_1$  to test the hypotheses just used in the  $t$  test.
- $H_0$  is rejected if the hypothesized value of  $\beta_1$  is not included in the confidence interval for  $\beta_1$ .

Next we can use the 95% confidence interval for beta 1 to test the hypothesis just used in the test. So, now with the help of conference interval also we can decide whether null hypothesis should

be accepted or rejected it is not as rejected if the hypothesis value of  $b_1$  is not included in the confidence interval. our  $b_1$  value what we are assuming to 0 so in that confidence interval if the 0 is appearing we have to accept the null hypothesis otherwise we have to reject null hypothesis.

(Refer Slide Time: 21:55)

### Confidence Interval for $\beta_1$

- The form of a confidence interval for  $\beta_1$  is:

$b_1$  is the point estimator

$b_1 \pm t_{\alpha/2} s_{b_1}$

$t_{\alpha/2} s_{b_1}$  is the margin of error

Where  $t_{\alpha/2}$  is the  $t$  value providing an area of  $\alpha/2$  in the upper tail of a  $t$  distribution with  $n - 2$  degrees of freedom

So, confidence interval for beta 1 is the form of controlled  $b_1 + \text{or} - t_{\alpha/2} S_{b_1}$ ,  $b_1$  which you got from our regression equation that is a coefficient of  $x_1$ ,  $S_{b_1}$  previously we are getting out. I told you what is the formula for getting  $b_1$ , so when you substitute it here.

(Refer Slide Time: 22:22)

### Confidence Interval for $\beta_1$

- Rejection Rule  
Reject  $H_0$  if 0 is not included in the confidence interval for  $\beta_1$ .
- 95% Confidence Interval for  $\beta_1$   

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.182(1.08) = 5 \pm 3.44$$

or 1.56 to 8.44
- Conclusion  
0 is not included in the confidence interval.  
Reject  $H_0$

So, what is getting  $b_1$  is 5 + or -  $t_{\alpha/2}$  is 3 point 18 -  $S_{b_1}$  is 1.08 so 5 + or - 3.44 so lower limit is 1.56 upper limit is 8.44 you see in that there is no 0's there so we have to reject our null

hypothesis, conclusion 0 is not included in the confidence interval so we are rejecting null hypothesis.

**(Refer Slide Time: 22:43)**

The slide is titled "Testing for Significance: F Test" in blue text. It contains two bullet points: "Hypotheses" and "Test Statistic". Under "Hypotheses", there are two green boxes containing the equations  $H_0: \beta_1 = 0$  and  $H_a: \beta_1 \neq 0$ . Under "Test Statistic", there is a green box containing the equation  $F = MSR/MSE$ . The slide has a dark blue header and footer with some logos.

The previous way we have used the t-test some time what will happen if the number of independent variable is more than 2 we have to do the t-test to 2 times. If there are say 5 independent variable you have to do file individually as I told you whenever you are comparing more than two we should go for Anova that is the F-test so here also whenever there is a number of independent variables more otherwise a generic method for testing the beta 1 equal to 0 hypothesis is going for F test.

So here you have F test is a MSR divided by MSE even in Anova also you know anova what we write is we write MS treatment divided by MSE, MS treatment is nothing but our regression sum of square mean regression sum of square.

**(Refer Slide Time: 23:43)**



### F-Test for Significance

- F Test statistic: 
$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{k}$$

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

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

So, F equal to MSR divided by MSE you see that MSR how we are getting MSR SSR divided by K and K is number of degrees of freedom that is nothing but number of independent variable n - K - 1 is degrees of freedom for the error term.

**(Refer Slide Time: 24:02)**

### Testing for Significance: F Test

- Rejection Rule

$$\text{Reject } H_0 \text{ if}$$

$$p\text{-value} \leq \alpha$$

$$\text{or } F \geq F_{\alpha}$$

where:

$F_{\alpha}$  is based on an F distribution with  
1 degree of freedom in the numerator and  
n - 2 degrees of freedom in the denominator

So, what is the rejection rule reject  $H_0$  if the p-value is less than equal to alpha otherwise if the calculated F value is greater than the value which we got from the table. So, if alpha is based on the F distribution we have to look at the what is the degrees of freedom as you look at you see enumerated degrees of freedom in this problem we have only one independent variables so one degrees of freedom numerator so n - 2 is 5 - 2 = 3, degrees of freedom for denominator.

**(Refer Slide Time: 24:33)**

## Testing for Significance: F Test

1. Determine the hypotheses.  $H_0: \beta_1 = 0$   
 $H_a: \beta_1 \neq 0$
2. Specify the level of significance.  $\alpha = .05$
3. Select the test statistic.  $F = \text{MSR}/\text{MSE}$
4. State the rejection rule. Reject  $H_0$  if  $p\text{-value} \leq .05$   
or  $F \geq 10.13$  (with 1 d.f.  
in numerator and  
3 d.f. in denominator)

So, beta 1 equal to 0 beta 0 equal to 0 alpha equal to 0.05 F equal to MSR divided by MSE, so p value we got to find out so numerator degrees of freedom is 1 t nominated is a freedom is 3 so that p value is 10.13.

(Refer Slide Time: 24:51)

## Jupyter Code

```
In [2]: import numpy as np
import matplotlib.pyplot as plt

In [3]: import seaborn as sns

In [4]: import pandas as pd
import matplotlib as mpl
import statsmodels.formula.api as sm
from sklearn.linear_model import LinearRegression
from scipy import stats

In [5]: tbl = pd.read_excel('C:/Users/Soni/Documents/regr.xlsx')
```

So, now we will use Python code rules that will do that import numpy as np import matplotlib.pyplot as plt, import seaborn ssn, import pandas as pd input matplotlib as mpl, import stats.models.formula.api as sm, from sklearn.linear\_model import LinearRegression, from scipy import stats. So, tbl we are going to that regression data we are going to save in the object called tbl and we are reading that.

Now what is happening you see that the p-value for the TV ad we say alpha equal to 5% it is less than 0.01 so TV ad is insignificant variable if it is more than 5% say, if it is a 0.06 the regression equation we will not include this independent variable TV ads you have statistics 21.43 I will go back will verify this answer that we can find out MSR, how we do MSR and that is what I am saying that time the first you have to find out SSR regression sum of square regression sum of square is see  $\sum (\hat{y} - \bar{y})^2$  divided by k, k is number of independent variable will get MSR.

So the p-value sorry the F value is 21.43 and see the probability it is less than 0.01 so that is less than 0.05 so we are saying that the model as a whole there are two things is there as a whole model the F value is less than 0.05 the model is valid and if you want to check individual independent variable also. So, see here it is less than 0.05 so this variable is significant see the lower limit, upper limit there is no 0 here 1.563, 8.43. So we can we cannot accept where to reject null hypothesis.

**(Refer Slide Time: 27:02)**

### Testing for Significance: F Test

5. Compute the value of the test statistic.
 
$$F = MSR/MSE = 100/4.667 = 21.43$$
6. Determine whether to reject  $H_0$ .
 

$F = 17.44$  provides an area of .025 in the upper tail. Thus, the  $p$ -value corresponding to  $F = 21.43$  is less than  $2(.025) = .05$ . Hence, we reject  $H_0$ .

The statistical evidence is sufficient to conclude that we have a significant relationship between the number of TV ads aired and the number of cars sold.

You see MSR is 21.43 we are here 21.43 so we can verify that our Python result then what we have done it with the help of manually. Some cautions about the interpretation of significant test is so it is very important this one rejecting  $H_0$ :  $b_1$  or  $\beta_1$  equal to 0 and concluding that the relationship between x and y significant does not enable us to conclude that there is a cause and effect relationship in present between x and y. Just because of there is a correlation we cannot say there is a cause-and-effect relationship.

So, just because of we are able to reject  $H_0: \beta_1 \text{ equal to } 0$  and demonstrate statistical significant does not enable us to conclude that there is a linear relation between  $x$  and  $y$ . Dear students in this class what we have seen we have taken one sample problem then we have fitted a regression equation in the regression equation we gone for hypothesis testing we have tested the significance of that independent variables.

There are two way to do the significance test one is by using  $t$  method  $t$  statistic method another one is  $F$  test method in both method we all got the same answer. Then I have explained what is the meaning of coefficient of determination, that is  $r^2$  from the  $r$  square I have explained how we can get the  $r$ . The next class will go for multiple regression equation where we will consider more than one independent variable and we will also ill explain some important assumptions in the regression equations. Thank you very much