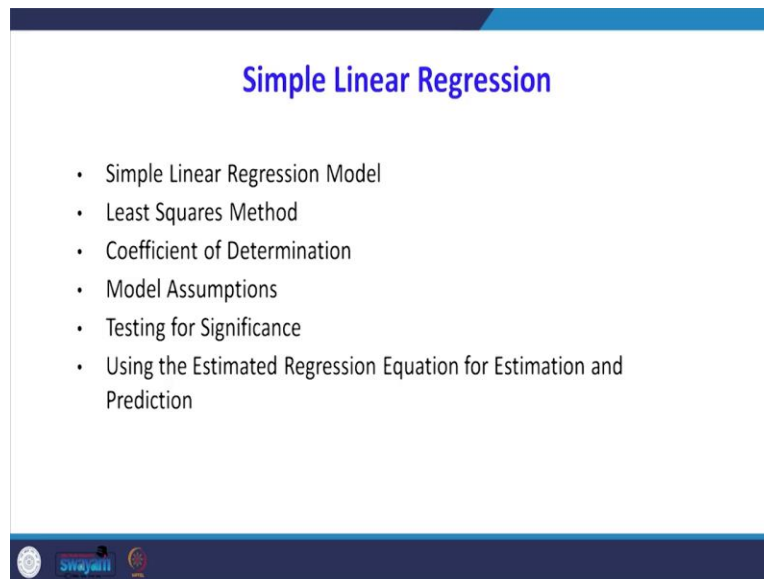


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 28
Linear Regression - I

Dear students in this class we will go to the new topic called regression analysis, the class objectives of this lecturers is;

(Refer Slide Time: 00:34)



We will study simple linear regression model when you say simple linear regression model only one independent variable will be considered then will see what is the least square method. That is the principle behind this regression model. Then I will see what is coefficient of determination goodness of regression model explained, generally with the help of this coefficient of determination called R square we will see in detail later.

What are the model assumptions then we can test for significance, even hypothesis testing also can be done with the help of our regression analysis, then using the estimated regression equation for estimation and use the prediction also.

(Refer Slide Time: 01:19)

Empirical Models

- Many problems in engineering and science involve exploring the relationships between two or more variables
- Regression analysis is a statistical technique that is very useful for these types of problems
- This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes

Many problems in Engineering and Science involve exploring the relationship between 2 or more variables. So far what you have seen the same variable we have compared with the, we have taken some sample with help of sample we are predicted the population parameter the same variable sometime we are compared the mean sometime we compared variance but this lecture they are going to take 2 different variables.

Regression analysis is a statistical technique that is very useful for these type of problems where the cause and effect has to be measured. This model can also be used for process Optimisation such as finding the level of temperature that maximizes yield or process control purposes. There are many independent variable we can say which independent variable is more important variable that affect our dependent variable.

(Refer Slide Time: 02:12)

Empirical Models Example

- As an illustration, consider the data in the table.
- In this table y is the purity of oxygen produced in a chemical distillation process, and x is the percentage of hydrocarbons that are present in the main condenser of the distillation unit.

Hydrocarbon level (X)	Purity (Y)
0.99	90.01
1.02	89.05
1.15	91.43
1.29	93.74
1.46	96.73
1.36	94.45
0.87	87.59
1.23	91.77
1.55	99.42
1.4	93.65

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

We will see one example. There is a table is given there is an X variable called hydrocarbon level Y Variable is called purity as an illustration considered data in the table. In this table Y is the purity of oxygen produced in a chemical distillation process and X percentage of hydrocarbons that are present in the main contents of the distillation unit. Now, we are going to see what is the influence of X on purity of oxygen?

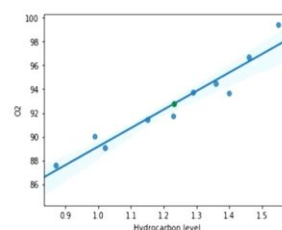
(Refer Slide Time: 02:43)

Using python for plotting the data

```
In [20]: data = pd.read_excel('C:/Users/Somi/Desktop/reg2.xlsx')

In [19]: x= data['Hydrocarbon level']
y = data['02']
plt.figure()
sns.regplot(x,y,fit_reg= True)
plt.scatter(np.mean(x), np.mean(y), color = "green")

Out[19]: <matplotlib.collections.PathCollection at 0x21ada0ab1d0>
```



So, I enter the data in excel so I saved the file name is reg2.xlsx when import data you go to pd.read_excel I have specified the path. So X = data that is a hydrocarbon level that column is my X variable, Y = in data file '02' that is my dependent variable and I use plot.figure then sns.regression plot regplot (x, y fit_regression equal to true then can I use this plt.scatter (

`np.mean(X), np.mean(Y), color = 'green')`, would be green so what I am saying and getting a scatter plot between hydrocarbon level and oxygen.

So, what is happening whenever the hydrocarbon level is increasing the oxygen level also increase. There is a positive relationship suppose if you want to make a relation between quantify the magnitude of X and how it is influencing and Y then I should go for regression equation that will do incoming slides.

(Refer Slide Time: 03:58)

Simple Linear Regression Model

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

β_0 and β_1 are called parameters of the model,
 ε is a random variable called the error term.

So, the theory behind the simple linear regression model the equation that describes how Y is related to X and an error term is called regression model. The simple linear regression model is $y = (\beta_0) \text{ beta } 0 + (\beta_1) \text{ beta } 1 x + (\varepsilon) \text{ error term}$. Here $y = \text{equal to beta } 0 + \text{beta } 1 x + \text{error term}$ where beta 0 and beta 1 are called the parameter of the model, ε is a random variable called the error term. So what we are going to do we are going to estimate the value of Y with help of independent variable X.

Because X itself will not enough to predict the Y variable there maybe some unknown variable other than X the error due to that unknown variable, otherwise unexplained variance we are going to call it is error term.

(Refer Slide Time: 04:52)

Simple Linear Regression Equation

The simple linear regression equation is:

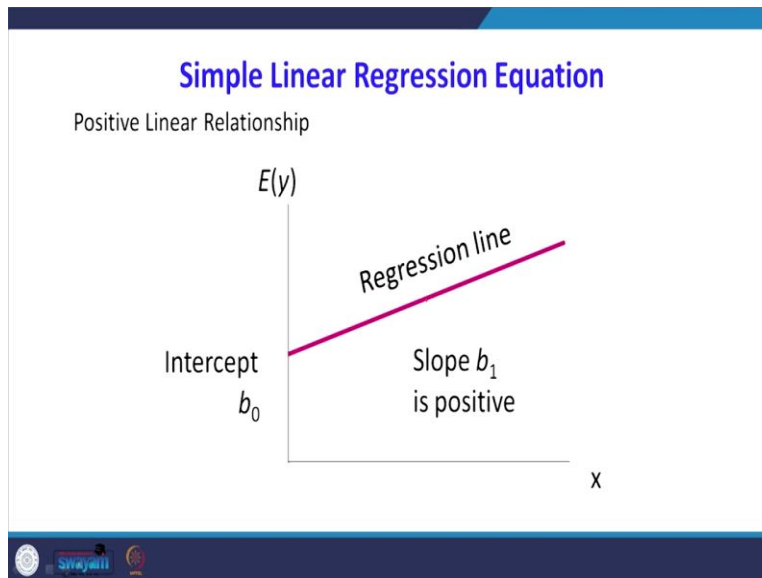
$$E(y) = \beta_0 + \beta_1 x$$

- Graph of the regression equation is a straight line.
- β_0 is the y intercept of the regression line.
- β_1 is the slope of the regression line.
- $E(y)$ is the expected value of y for a given x value.

The simple linear regression equation is expectation of Y equal to beta 0 + beta 1 X when you comparing the previous slide. That was there is no error term because while calculating the value of beta 1 when we have taken care that error is minimized not only that the previously the previous slide we are writing Y now it is expected value of Y. Now what we are predicting is the mean value of Y not the actual value of Y where the graph of the regression equation is a straight line.

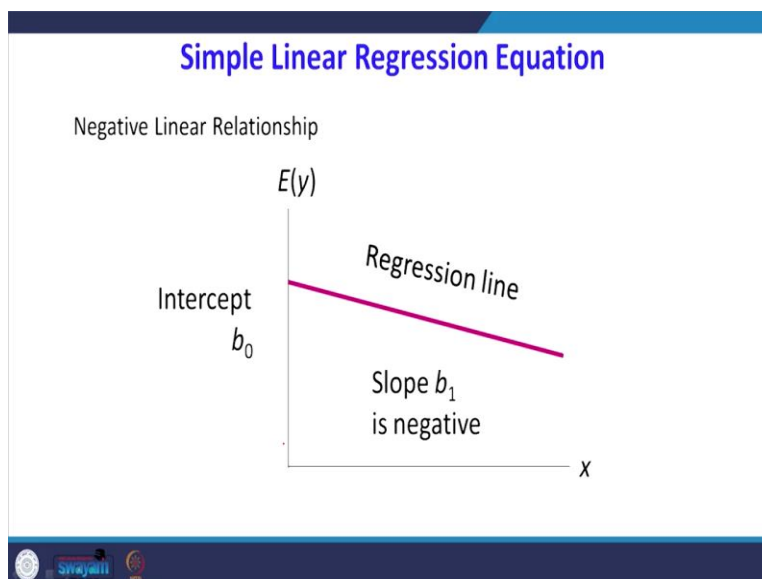
Because the power of x is 1, beta 0 is the Y intercept of the regression line beta 1 is the slope of regression line. So, the expected Y is the expected value of Y for a given X value expected values nothing but mean value.

(Refer Slide Time: 05:41)



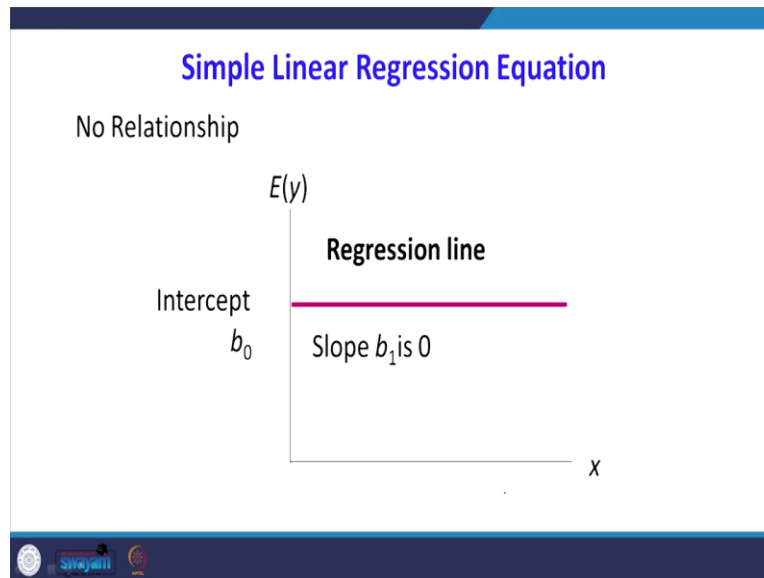
The simple linear regression equations. This is an example of a positive linear relationship in the x -axis. There is a positive relationship when the value of x is increasing, the expected value of y also increases, so the slope b_1 of this is positive. The intercept so this distance is your b_0 .

(Refer Slide Time: 06:04)



This is an example of a negative linear relationship. What is happening when x increases, the expected value of y is decreasing; here the slope is negative.

(Refer Slide Time: 06:16)



Here, there is no relationship because it is a line which is parallel to x-axis to the slope is 0. So what is the meaning of this irrespective of any value of X, the expected value of Y is same. Here we can see the value of x and y are independent.

(Refer Slide Time: 06:35)

Estimated Simple Linear Regression Equation

- The estimated simple linear regression equation

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

- The graph is called the estimated regression line.
- b_0 is the y intercept of the line.
- b_1 is the slope of the line.
- \hat{y} is the estimated value of y for a given x value.

The estimated simple linear regression equation is $\hat{y} = b_0 + b_1x$ generally write the capital Y if I write $\beta_0 + \beta_1X$ if I use capital letters that is for the population. What we write in small letter that is for the sample. So y hat is the estimated regression line b 0 is the y intercept of the line b1 is the slope of the line y hat is the estimated value of y for given x value.

(Refer Slide Time: 07:09)

Least Squares Method

- Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

where:

y_i = observed value of the dependent variable
for the i th observation

\hat{y}_i = estimated value of the dependent variable
for the i th observation

The principle behind the least square method is the sum of the square of the error has to be minimized. Suppose I have some x value y, I have some number for x I have a number for y suppose, I have drawn this way. This is x axis. This is y axis and plotted line like this. So my objective is I have to draw a line. I have to draw a line. Ideally that line has to pass through all the given points, but that is not possible.

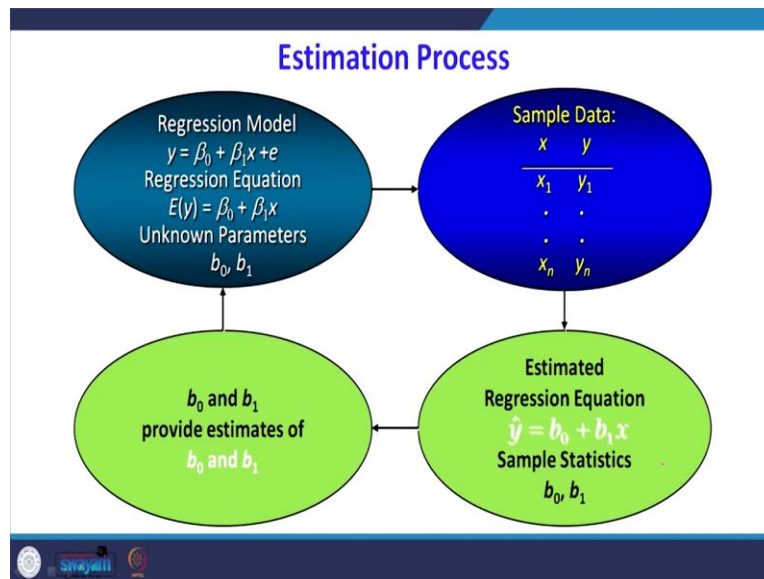
So what I am going to do I am going to draw a line so that the error is minimised not only the for, example this is e1 this is e2 this is like that be many e3, so this is positive error actual minus predicted value. So this line is \hat{y} equal to so $b_0 + b_1 x$. So now what happening this much distance is it is a vertical lines to this much distance is my error actual minus predicted value. This vertical line distances e2 so what is happening? This is error.

So what I have to do the error square and sum of the square has to be minimised like this. So, some of the error has to be minimised. I have to draw a line in such a way that sum of the square of the error has to be minimised. I can draw different line suppose. I can draw this, this way also, this way also for each line. I want to find out this sum of the square of error wherever the sum of the square of the error is minimum so that line is the best line that principle called least square method.

Why we are squaring there is logic behind this, if you are squaring the positive and negative error will become nullify we will 0 that is why we are squaring that the same logic for example the formula for variance what we are doing $\Sigma(X - \bar{X})^2 / (n - 1)$, the logical why we ask squaring the 2 purpose otherwise $\Sigma(X - \bar{X})$ equal to 0 here the sum of positive or negative or 0, the square transformation says one more implications that suppose the deviation is less for example it is 0.5 when you square it then it is 0.25.

Suppose deviation is 5 the net value is 55. What is happening? There is lesser deviation had lesser penalty, there is a larger deviation larger penalty. That is beauty of this squared the transformation.

(Refer Slide Time: 10:19)



In the estimation process what is happening initially, we will assume a regression model Y equal to $\beta_0 + \beta_1 X + e$, that regression model will predict with help of regression equation that is expected value of Y equal to $\beta_0 + \beta_1 X$ you see that. The regression equation there is no error term. Here the unknown parameters are the population regression model say Y equal to $\beta_0 + \beta_1 X + e$, the regression equation is expected value of Y equal to $\beta_0 + \beta_1 X$.

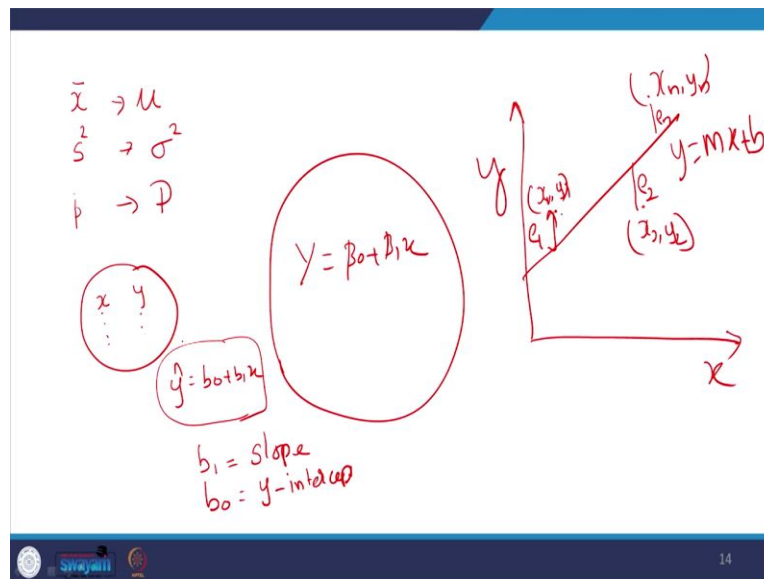
Here the unknown parameters are β_0 and β_1 . We have to estimate the value of β_0 and β_1 more importantly the value of β_1 what you are going to estimated whether the value of β_1 is 0 or other than 0 if I estimate $\beta_1 = 0$ that means there is no relation between X and

Y. So, this is our equation from that what I am going to do I going to collect the sample for my X variable and Y variable X is independent variable. Y is dependent variable. So, with the help of the sample data are going to make a regression equation that is applicable only for the sample. So $\hat{y} = b_0 + \beta_1 x$, here b_0 and $\beta_1 x$ is the sample statistics.

So this equation is valid only for the sample now I going to estimate that whether the value of β_1 and β_0 is valid event for the population level also, what is the meaning of that one is sometime this Y you could to $b_0 + \beta_1 X$ with the help of sample. You can construct a regression equation. What will you predict it when you estimate value of β_1 estimate value of β_0 , then may not be significant other population-level.

So that time what we are going to see the β_1 is equal to 0, if the β_1 is equal to 0 there is no relation between X and Y we will see in the coming slide.

(Refer Slide Time: 12:42)



How this regression is different from our previous concept which have studied. For example what happened with help of \bar{x} we have predicted population parameter mean with the help of sample variance we have predicted the population variance with the help of sample proportion. We have predicted population proportion the regression actually what is happening there is a sample smaller circle sample bigger circle is population.

I have some X and Y value from the sample with help of x and y value I hope predicted regression equation y equal to $b_0 + \beta_1 x$ now I am going to prove that whether this relationship is valid even for the population for that what I am going to do capital $y = \beta_0 + \beta_1 x$ with the help of regression equation. That means a sample model I going to predict whether this model are this relationship is valid for even for the population are not.

Sometime what happened you can construct a regression equation with help of sample data. You can say there is a relation between x and y but when you go to the population level, there were not be relation between x and y. So, what is the difference between this regression modelling? And previous our hypothesis testing, in hypothesis testing we have tested only one parameter at a time what you done we have tested, we are predicted mean are variance are population proportion.

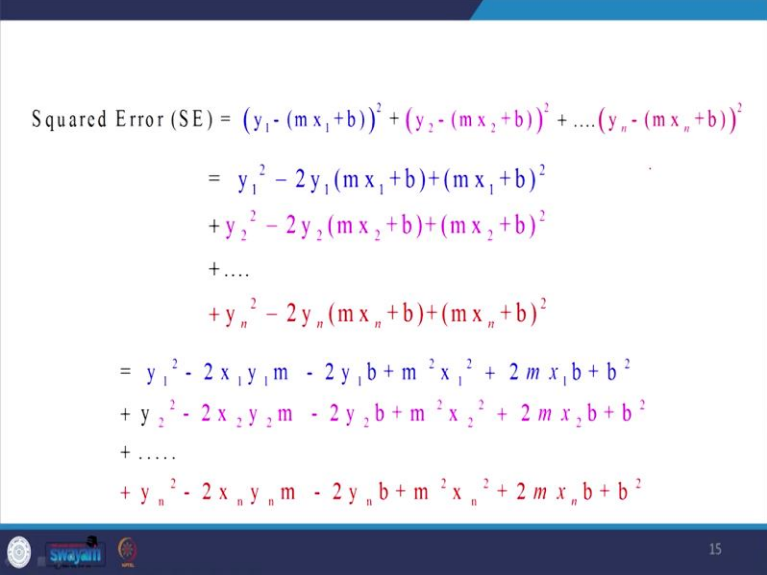
Now I have constructed a model small model with the help of sample data I am testing this model in the population level, I am simultaneously I am checking 2, 3 parameter one is my β_1 one parameter, β_0 is another parameter, like that we may have different this is simple regression like that in the multiple regression different independent variable. This is the logic of regression modelling.

What will you do suppose in the regression equation with the help of sample data? What is required is I have to find out what is the value of β_1 , b_1 this is called slope. This is my 'y' intercept what is happening a line is like this suppose there are 2 points. Suppose I am going to call this is a (x_1, y_1) , this is (x_2, y_2) like that there will many point for example this is (x_n, y_n) ok. So this is my x-axis. this is y-axis. So what I am going to do this is my error going to call to e_1 this is e_2 this is e_n so what are you going to do? First go to find out the error for each values then I going to square the error.

Then I am going to sum the error then for what value of this b_1 and b_0 the error will get minimised so that I am doing here the next slide to what is happening. So here I am going to call it is this is y equal to $mx + b$ this is traditional notation because our school in your study this

really you can use any notations. So here what is the error term actual minus predicted so my actual error is for this one my actual point is y_1 my predictable value is y .

(Refer Slide Time: 16:57)



$$\begin{aligned}
 \text{Squared Error (SE)} &= (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots (y_n - (mx_n + b))^2 \\
 &= y_1^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2 \\
 &\quad + y_2^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 \\
 &\quad + \dots \\
 &\quad + y_n^2 - 2y_n(mx_n + b) + (mx_n + b)^2 \\
 &= y_1^2 - 2x_1y_1m - 2y_1b + m^2x_1^2 + 2mx_1b + b^2 \\
 &\quad + y_2^2 - 2x_2y_2m - 2y_2b + m^2x_2^2 + 2mx_2b + b^2 \\
 &\quad + \dots \\
 &\quad + y_n^2 - 2x_ny_nm - 2y_nb + m^2x_n^2 + 2mx_nb + b^2
 \end{aligned}$$

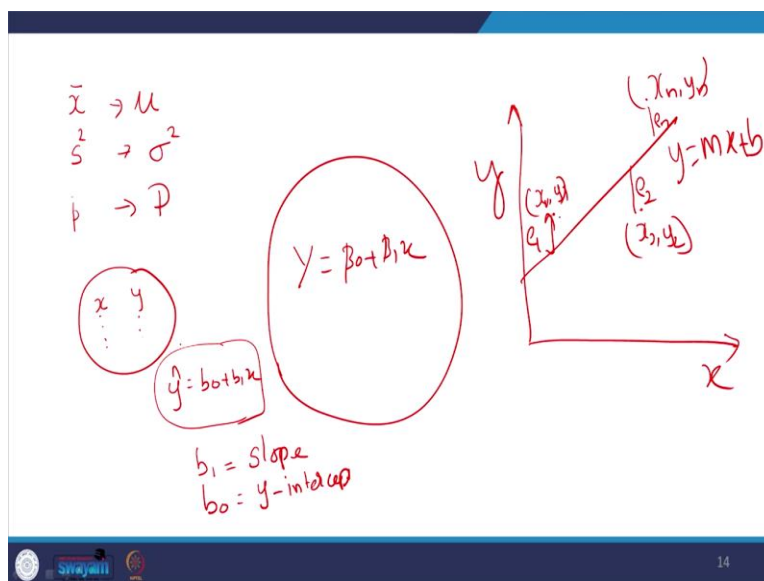
So, $y_1 - y$, y is my $mx_1 + b$ whole square for second term y_2 is actual $mx_2 + b$ because you know this y equal to $mx + b$. so y equal to y_1 when you substitute x_2 equal to x_1 you will get y predicted value actual minus predicted that is a square for finding this actual minus predicted is error and squaring error. so, for the second term $y_2 - mx_2 + b$ like this $y_n - 1 + mx_n + b$ whole square just I am going to simplify this. This is nothing but a - b whole square formula.

y_1 square - $2y_1mx_1 + b + mx_1 + b$ whole square similarly for the pink one y_2 square a - b whole square - $2y_2mx_2 + b + mx_2 + b$ whole square for nth term y_n square - $2y_nmx_n + b + mx_n + b$ whole square just i am going to simplify this so when you simplify this y_1 square will be there you take this $-2y_1$ inside it will become $-2x_1y_1m$ then $-2y_1b +$ this is a + b whole square, m square x_m square + $2x_1b + b$ square this was is the first term.

For the second term y_2 square - $2x_2y_2m - 2y_2b + m$ square x_2^2 square + $2mx_2b + b$ square for the nth term y_n square is nothing but sometime a - b sometime a + b whole square formula y_n square - $2x_ny_nm - 2y_nb + m$ square x_n square + $2x_nb + b$ square. now what i am going to do the next term are going to add all this y_1 square + y_2 square y_n square then here the $-2m$ is the constant here. The next second term so I am going to add this I go to group

it, similarly $-2b$ same so I go to add only $y_1 + y_2 + \dots + y_n$ here the m square is the constants $x_1^2 + x_2^2 + \dots + x_n^2$ here $2mb$ is a constant so I am going to add $x_1 + x_2 + \dots + x_n$ and here there are b square n times and b square that should have done this one.

(Refer Slide Time: 19:52)



So I have grouped all y_1^2 term $y_1^2 + y_2^2 + \dots + y_n^2 - 2m$ as I told you previously I am going to here the $-2m$ is constant so $x_1 y_1 + x_2 y_2 + \dots + x_n y_n$ I may go back here $-2b$ for the third term $-2b$ is the constant. $-2b$ the remainder is $y_1 + y_2 + \dots + y_n$. the 4th term m^2 is a constant $m^2 x_1^2 + x_2^2 + \dots + x_n^2 + 2x_1^2$. then next $2mb$, $2mb$ is constant in all the terms so when you bring it common $2mb$ the remaining is $x_1 + x_2 + \dots + x_n$ the last term is b^2 b^2 b^2 .

So here but I am going to do you see that. I want to know y_1^2 mean. so what time to do σ^2 of $y_1^2 + y_2^2 + \dots + y_n^2$ divided by n . so, what are you going to do the submission and go to write in terms of its average. so when I take this one so instead of $y_1^2 + y_2^2$ I can write $n \bar{y}^2$. similarly this $-2m$ $-2m$ so this one I can write $n \bar{x} \bar{y}$ here $-2b$ what I have done.

I have group the square term. the first term is $y_1^2 + y_2^2 + \dots + y_n^2$ and so on $+ y_n^2$ to this I want to write in terms of its average value so I go to write \bar{y}^2 nothing but $y_1^2 + y_2^2 + y_3^2 + \dots + y_n^2$ divided by n square. Now what is happening Y_1

square y^2 square + yn square can be replaced by multiplied by y square bar. so that is wrote it as y square bar. similarly the second term $x_1 y_1 x_2 y_2$ + up to $x_m y_n$ can be written as xy bar multiplied by n .

so the remaining term is $2mn xy$ bar the next $- 2b$ that i go to write $n y$ bar. so we will look at the third bar m square, m square will come as it is so x_1 square + x_2 square and so on going to write in terms of n multiplied by x square bar the. Next term is $2mb$ I am going to write $n x$ bar there are $n b$ square writing $n b$ square. So this is the simplified of the error term.

(Refer Slide Time: 23:00)

$$\begin{aligned}
 SE &= n \overline{y^2} - 2mn \overline{xy} - 2bn \overline{y} + m^2 n \overline{x^2} + 2mbn \overline{x} + nb^2 \\
 \frac{\partial (SE)}{\partial m} &= -2n \overline{xy} + 2m \overline{nx^2} + 2bn \overline{x} = 0 \\
 \frac{\partial (SE)}{\partial m} &= -2n \overline{xy} + 2m \overline{nx^2} + 2bn \overline{x} = 0 \\
 &= -\overline{xy} + m \overline{x^2} + b \overline{x} = 0 \\
 m \overline{x^2} + b \overline{x} &= \overline{xy} \\
 m \frac{\overline{x^2}}{\overline{x}} + b &= \frac{\overline{xy}}{\overline{x}} \quad \text{one point } \left(\frac{\overline{x^2}}{\overline{x}}, \frac{\overline{xy}}{\overline{x}} \right)
 \end{aligned}$$

Actually it is a squared error. So what is happening $n y$ square - n square - $2 m n xy$ bar - $2b ny$ bar + m square $n x$ square bar + $2 m b n x$ bar b square. Now what is happening here which is variable there are 2 variable one is here comes the slope the slope is the variable because why I am saying slope is variable I can draw different slope different line the slope is variable. So I have to find out for what value of this slope the square of the error will be minimised.

So it is we say it is Maxima minima principal. So what will do for generally what will you do for maxima minima principal dy by dx equal to 0, might have studied in schooling so $d^2 y$ by dx^2 is negative less than 0 so that the; it is this way. This way what happening dy by dx equal to 0 is this point if it is negative. That means you are, it is a maximum point if it is a positive. So that is the minimum point.

So, what is happening even to the both for both the conditions. The first one is dy by dx is equal to 0. Here x is variable for example here the m the slope is variable and the b , y intercept is the variable. so, first we will; because there are 2 variable is there you partially differentiate this squared error first with respect to m when you, partially differentiate with respect to m . so there is no n term there will be $-2m \sum xy$ plus here also there is no m term it is 0 so $2mn \sum x^2$ square bar here there is x term $2b \sum nx$ bar this will become 0 then equate to 0.

So when you simplify this $2n$ is the constant $2n$ here $2n$, $2n$ remove this the remaining $-\sum xy$ square + $m \sum x^2$ square bar + $b \sum x$. so i am going to write in this one, y equal to $mx + b$ + format. so what will happen $m \sum x^2$ + $b \sum x$ take right and side $\sum xy$ bar, so i am going to divide by $\sum x^2$ bar so $\sum xy$ bar by $\sum x^2$ bar + b . this is $\sum xy$ bar $-\sum xy$ bar divided by $\sum x^2$ bar now what happening this is $mx + b$ equal to y format which one this equation.

So, m is m so the x coordinate is $\sum x^2$ bar divided by $\sum x$ bar so the y coordinate is $\sum xy$ bar divided by $\sum x$ bar. so, what this implies is if you want to draw a best line that line has to pass through this point this is x coordinate this one is this is first one x coordinate and the second one y coordinate. so if you want to draw a line which should minimise the sum of the squared error that has to pass through this point.

(Refer Slide Time: 26:27)

$$\begin{aligned}
 SE &= n \bar{y}^2 - 2mn \bar{x} \bar{y} - 2bn \bar{y} + m^2 n \bar{x}^2 + 2mbn \bar{x} + nb^2 \\
 \frac{\partial(SE)}{\partial b} &= -2n \bar{y} + 2mn \bar{x} + 2nb = 0 \\
 &= -\bar{y} + m \bar{x} + b = 0 \\
 \bar{y} &= m \bar{x} + b \\
 \text{another point } (\bar{x}, \bar{y})
 \end{aligned}$$

Then we will find out the other because to know the slope we need 2 point we got already one. So now, we differentiate partially differentiate with respect to b so here there is no b term 0, here also, there is no b term 0 here there is a b term -2 my bar there is no b term here $2mn \bar{x}$, so here $2nb$ so equate to 0 here also this $2n$ is a constant $2n$ $2n$ so divide both side remaining $-y \bar{y} + mx + b$ equal to 0 so when you simplify y equal to $mx + b$ format.

so now this is y equal to $mx + b$ format this line passing through the $y \bar{y}$ so this line is passing through $x \bar{x}$ so the another point is $x \bar{x}$, $y \bar{y}$ so you see this is very important result. if you want to draw a best line that line has to pass through the average value of its x and average value of y, one of the point should that lines to pass through that then only that line maybe the best line. so we got the 2 point $x \bar{x}$, $y \bar{y}$ another point is $x \text{ square bar by } x \bar{x}$, $xy \bar{y} \text{ divided by } x \bar{x}$.

so, when we know this one point say another point is $x \bar{x}$, $y \bar{y}$ so if you we want to know the slope of this equation then what is the slope formula $y_2 - y_1$ divided by $x_2 - x_1$ when you use that formula you will get formula for slope. dear students we got the 2 points after using the least square principle the one point is $x \text{ square bar divided by } x \bar{x}$, $xy \bar{y} \text{ divided by } x \bar{x}$ that is one point.

Another point is $x \bar{x}$, $y \bar{y}$ when there are 2 point is there we can find out the slope. what is the slope formula of this is first point, what is the slope formula suppose a traditional we might have studied this in school. Suppose there is 2 points point 1 is x_1 , y_1 so point 2 is x_2 y_2 ok. so, that x point is $x \text{ square by } x \bar{x}$, y_1 point is divided by $xy \bar{y} \text{ divided by } x \bar{x}$. so, x_2 point is $x \bar{x}$ $y \bar{y}$ bar we know the slope formula $y_2 - y_1$ divided by $x_2 - x_1$. here the y_2 is $y \bar{y} - y_1$ $xy \bar{y} \text{ divided by } x \bar{x}$ divided by $x \bar{x}$ is $x \bar{x}$ minus actually this is i wrote x_1 y_1 for only our convenience.

so, this x_1 is different x_2 is $x \bar{x}$, x_1 is $x \text{ square by } x \bar{x}$ by $x \bar{x}$. so this is when you multiply both side numerator and denominator by $x \bar{x}$ it will become $x \bar{x}$ $y \bar{y} - xy \bar{y}$ divided by $x \bar{x}$ whole square, $x \text{ square bar}$ because it is $x \bar{x} \times x \bar{x}$ get cancelled when you bring minus this one when you multiply both side by minus $xy \bar{y} \text{ square} - x \bar{x}$ $y \bar{y}$ divided by $x \text{ square bar} -$

\bar{x} whole square. this is the formula for slope actually this slope is nothing but when you look at the numerator, there is nothing but the covariance of (x, y) the denominator is variance of x i will explain how this numerator is covariance.

so, we know that in our probability class you study the covariance of x, y is expected value of $x - \bar{x}$ by $y - \bar{y}$ and you simplify this you bring this side $xy - xy - xy \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y}$ bar we you bring e inside? it will become $e \cdot xy$ because \bar{y} is number when you bring your e of x again it is \bar{x} - \bar{x} is a number when you bring e inside it will become \bar{y} because \bar{y} is a number so that will be as it is. so, when you bring e here it becomes $xy \bar{y}$ so $- \bar{x} \bar{y} \bar{y} - \bar{x} \bar{y} \bar{y} + \bar{x} \bar{y} \bar{y}$ you can cancel it plus and minus.

the reminder is $xy \bar{y} - \bar{x} \bar{y} \bar{y}$ that is nothing but the numerator and going back you see that $xy \bar{y} - \bar{x} \bar{y} \bar{y}$ is numerator. so, that in the slope formula numerator is nothing but covariance of x, y so the variance of x is this also we studied from school $x - \bar{x}$ whole square you expand $x^2 - 2x\bar{x} + \bar{x}^2$ when you bring x inside this will become e of $x^2 - 2x\bar{x}$ is number the expected value of a number is number itself and \bar{x} becomes \bar{x} bar this is number itself will keep as it is.

when we keep e inside become $x^2 \bar{y}$ so there are $- 2x\bar{x} \bar{y} + \bar{x}^2 \bar{y}$ square. so when you subtracted it the remaining is 1. so, $x^2 - \bar{x}^2$ even when i go back when we look at this, this formula is sampling with this denominator. so, the slope formula the numerator is nothing but write numerator is nothing but the covariance denominator nothing, but the variance of x .

Actually this variance covariance and correlation coefficient regression slope all are having some relationship. See that the variance formula we know $\sigma^2_x = \frac{\sum (x - \bar{x})^2}{n - 1}$ that is only one variable. in the covariance there are 2 variable x, y , so $\sigma_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$ even this can be written as $\sigma_{xy} = \frac{\sum (x - \bar{x}) \sum (y - \bar{y})}{(n - 1)^2}$ there 2 variables is there instead of another $x - \bar{x}$ you can write $\sigma_{xy} = \frac{\sum (x - \bar{x}) \sum (y - \bar{y})}{(n - 1)^2}$ divided by $n - 1$.

so, the correlation coefficient is nothing but when you divide this covariance divided by its own standard deviation, it will get correlation coefficient. so, but the slope is the ratio of covariance divided by variance of x you see the variances. the covariance is this $\frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$ divided by $\frac{1}{n-1} \sum (x_i - \bar{x})^2$ assume that the equal same degrees of freedom. the denominator is $\frac{1}{n-1} \sum (x_i - \bar{x})^2$ of whole square n -1. so, this n -1 get cancelled the formula for slope is $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$.

(refer slide time: 33:14)

Least Squares Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

that is why we got this formula b_1 so we need not it is easy way to remember this formula for the slope is nothing but covariance by variance. in this class we started about the regression analysis, I have explained the importance of regression, how the regression is different from the our traditional hypothesis testing. in the regression equation there is a y intercept is there and slope there by using least square method. i have derived the formula for finding the slope and the y intercept.

The formula for slope is nothing but the covariance by variance then I have interlinked how variance covariance and correlation coefficient and regression coefficients. All these are interrelated. The advantage here is you need not remember the formula, formula is; if you know the variance formula and covariance formula easily, you can find out the slope of regression equation. We will continue the next class by taking an example; I will explain how to use this regression equation for prediction purpose. Thank you very much.