Business Statistics

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

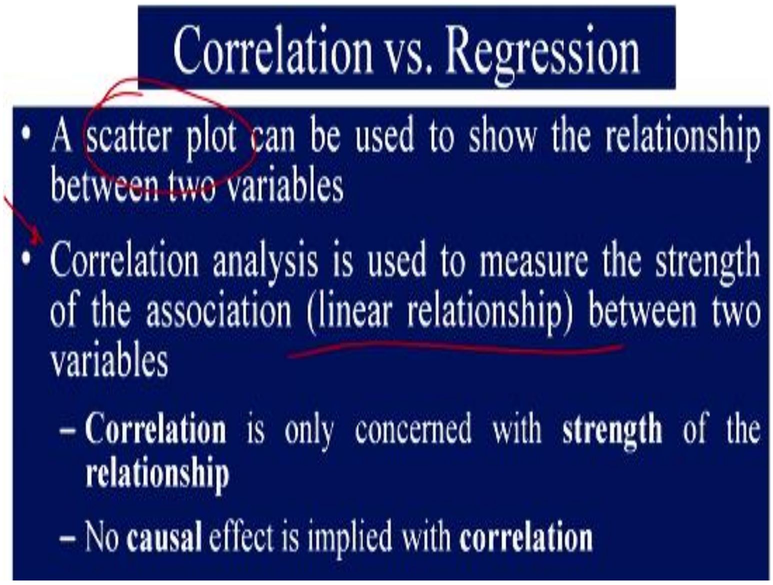
Lecture - 56
Simple Linear Regression-I

Hello friends. I welcome you all in this session. As you are aware in previous session we were discussing about chi-square test and we have seen chi-square test of independence wherein we have worked out several questions. In today's session I am going to talk about regression simple linear regression. So whenever we talk about regression we always talk about correlation as well in fact regression and correlation go hand-in-hand.

We have seen correlation when we talked about the measures of dispersion where we have seen there is something called coefficient of variation right. So apart from coefficient of variation there were other measures of dispersion which we have seen, but in today's class we will look at regression So what is regression? In simple term regression is something which gives us association between 2 sets of numerical data or 2 time series data

So you can have association between let us say age and weight or let us say you can have relationship between expenditure and income higher the income higher the expenditure right so you can have this sort of relationship. So relationship can be either positive or negative or you may have a situation of no relationship at all right. So regression is something which gives you association between 2 variables which are measured on metric scale. So you if you want to know association between any 2 variables.

(Refer Slide Time: 02:27)



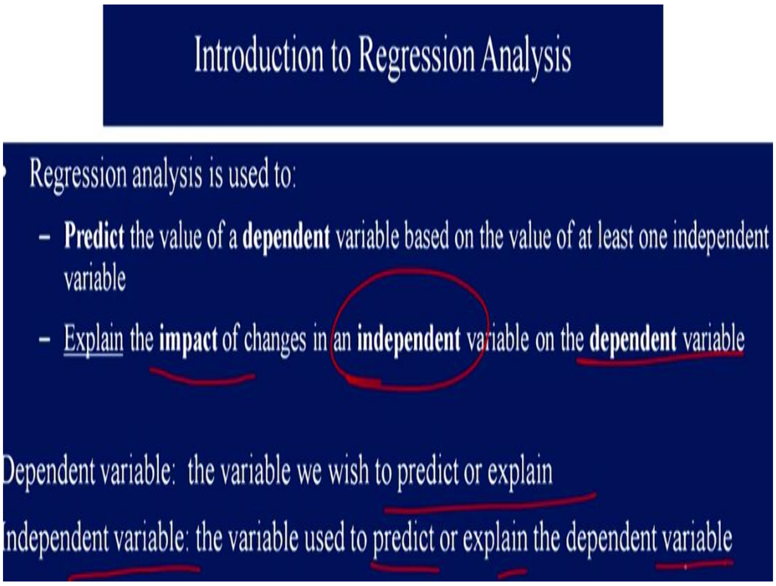
Correlation vs. Regression

- A scatter plot can be used to show the relationship between two variables
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables
  - **Correlation** is only concerned with **strength** of the **relationship**
  - No **causal** effect is implied with **correlation**

Then the best thing is you should plot something called scatter plot. So it gives you an idea about association between 2 variables. Correlation is something which measures how strong that association is so it measures degree of relationship between 2 variables and it measures only linear relationship. We are not talking about nonlinear relationship so how the strength of association is gets measured by correlation.

It has nothing to do with causation right keep in mind. Correlation has nothing to do with causation when I say causation means there is something called cause and effect relationship. So we are saying that let us say there is high correlation between income and expenditure, but it does not mean that there is income and the cause is expenditure it is not like that right. So correlation has nothing to do anything with causation.

(Refer Slide Time: 03:41)



Introduction to Regression Analysis

- Regression analysis is used to:
  - **Predict** the value of a **dependent** variable based on the value of at least one independent variable
  - Explain the **impact** of changes in an **independent** variable on the **dependent** variable

Dependent variable: the variable we wish to predict or explain

Independent variable: the variable used to predict or explain the dependent variable

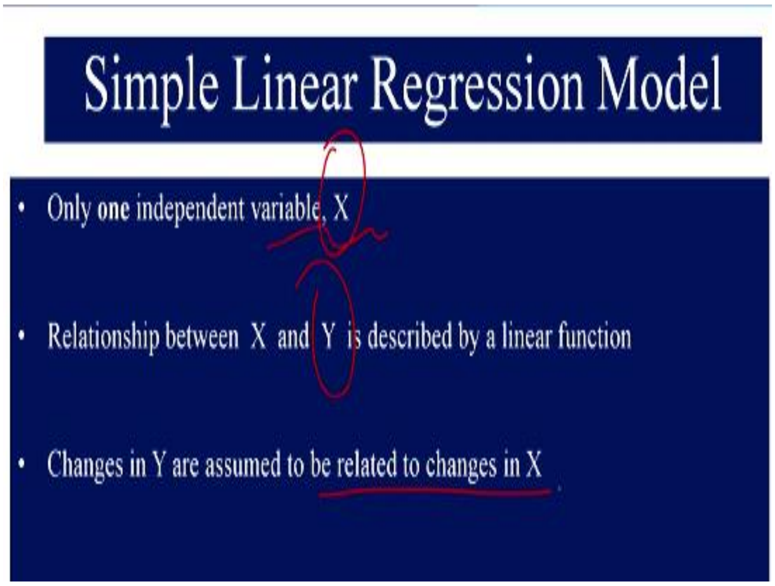
So regression analysis has got 2 important functions, 2 important applications either it can be used for prediction or it can be used for explanation. So prediction function and explanation function are the 2 functions of regression. So when I say prediction function it means it predicts dependent variable let us say profit is a function of sales. So profit is dependent variable and sales is independent variable.

So if we have sales data we can perform experiment on let us say independent variable by increasing or decreasing sales and we will see the effect of that experiment on profit right. It explains the impact of change in independent variable on the dependent. So the second function is explanation function right. It will help you in knowing how much change in dependent variable occur if there is a change in independent variable.

So in regression you will always have at least one dependent variable and one independent variable. So if there is a dependent variable and an independent variable it would be a simple regression. It is a bivariate regression; you may have multiple regression as well wherein you can have more than one independent variables right. So dependent variable is something which we wish to predict or explain right.

As I said profit is a function of sales so sales is independent variable and profit is dependent variable. Independent variable is something which we always use to predict or explain the dependent variable right.

(Refer Slide Time: 05:37)

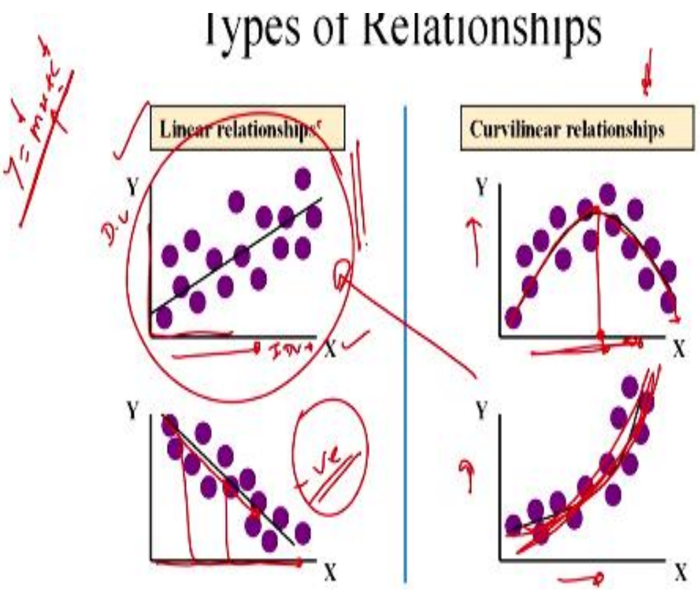


Simple Linear Regression Model

- Only one independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X

So when we talk about simple linear regression model; we will have just one independent variable and the relationship would be linear between X and Y so X is independent and Y is dependent variable. So change in Y are assumed to be related to changes in X okay. So whenever there is one unit change in independent variable causes one unit change in dependent variable then it would be a perfect linear regression model right.

(Refer Slide Time: 06:17)



Types of Relationships

Linear relationships

Curvilinear relationships

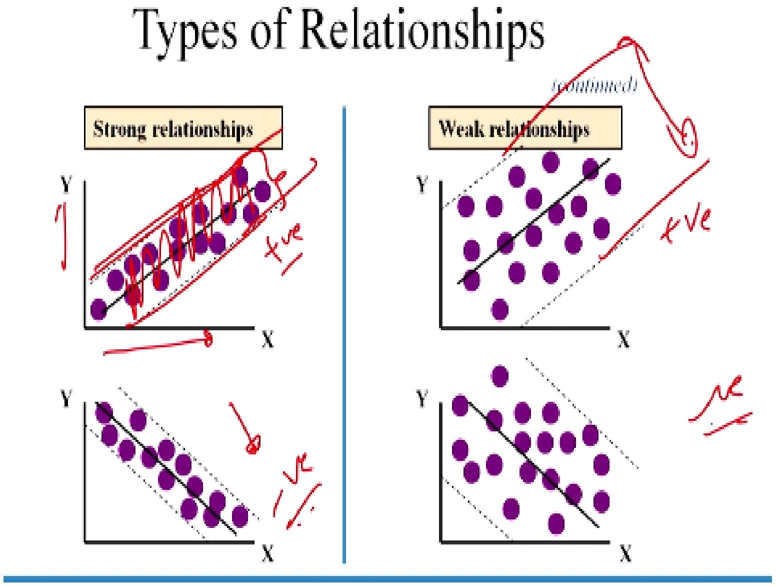
Let us look at couple of relationships so you have got independent variable over here on X axis. So independent variable X is independent variable this is dependent variable right on Y axis. So you can just simply write an equation Y = mx +c is nothing but equation of line right. So here c is constant right m is the slope and x is the independent variable right. So there is a linear relationship.

There is a positive linear relationship the moment we increase X the value of Y also increases. So this is linear relationship so linear relationship can be positive or negative now this is a negative one right So the moment you increase let us say X the value of Y decreases is not it. So this is a negative relationship, but linear right. You may have a situation where relationship is Curvilinear.

But in most of the times you will have a curvilinear relationship right. So this is how the value of Y is changing with respect to X. So when X increases value of Y increases, but after a particular point let us say after this if you further increase value of X then Y starts decreasing right. This is a kind of exponential relationship right. So with increase in X the value of Y is increasing.

But the rate of increment of Y is more in this case compared to this case is not it. So these are couple of relationships.

(Refer Slide Time: 08:26)

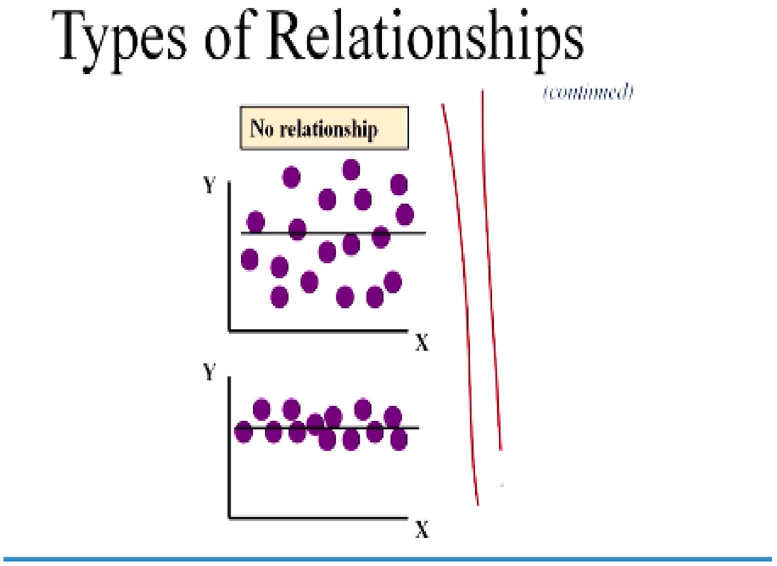


Types of Relationships

(continued)

So you can have strong relationships. So with increase in X, Y is increasing but this relationship is positive and strong relationship. Why we are calling it strong relationship because this dispersion or the standard deviation is very less compared to this standard deviation is it not. So if standard deviation is less it would be a stronger relationship. This is again strong relationship, but negative one so this is positive one this is negative one and again weak relationship positive and negative right.

(Refer Slide Time: 09:10)

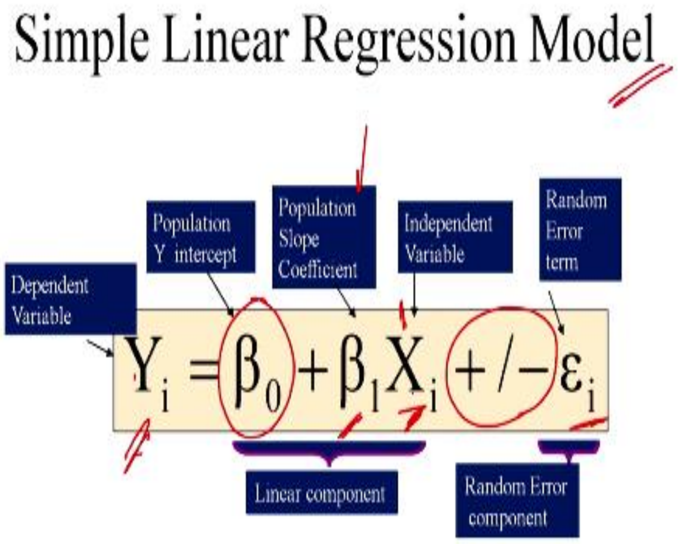


Types of Relationships

(continued)

You can have a situation of no relationship at all so X and Y are independent they are not related to each other a situation like this okay so there is no relationship. So we have seen linear relationship, curvilinear relationship and no relationship.

(Refer Slide Time: 09:33)



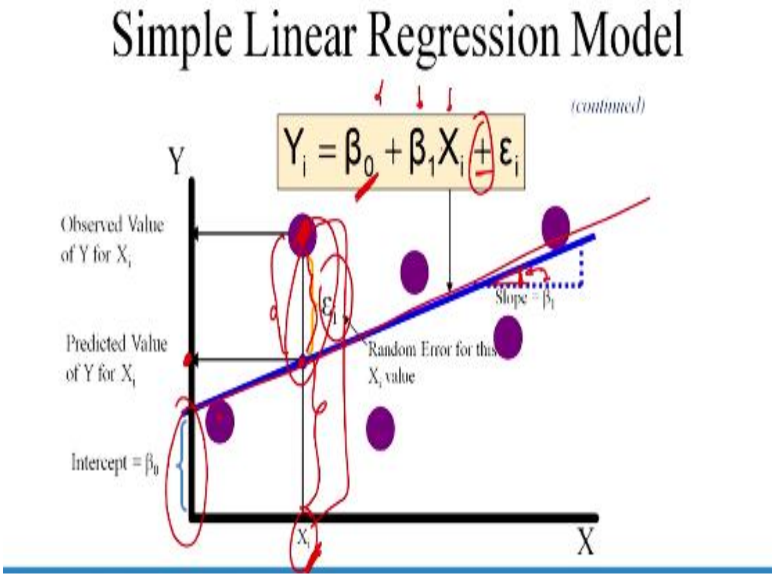
Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i +/- \varepsilon_i$$

So this is how the simple linear regression model looks like So this is Y= or Yi= B_0+ B_1 Xi ± error ei. So this is your dependent variable, this is intercept. Intercept is something which you will get if you X=0 in this equation. So for a 0 value of X whatever is B_0 is nothing but intercept. So this is your independent variable population slope coefficient right. So this is coefficient of independent variable X1.

You can call it you can have different multiple independent variables right. So this is linear component of the relationship and this is error component right.

(Refer Slide Time: 10:41)



Simple Linear Regression Model

(continued)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

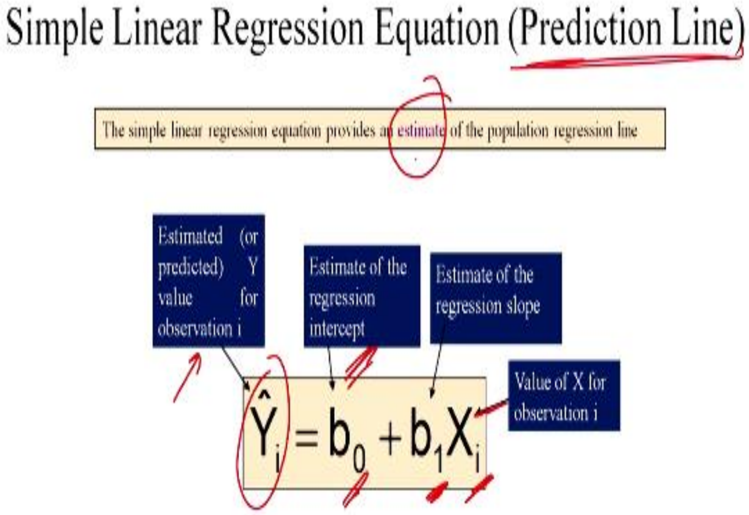
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

Simple linear regression model we will continue this model so Yi = this is slope; this is intercept right B0 this one right Bi, B1 is slope which is this one this angle is nothing but Bi right this is Bi right. Xi is you can have more than one independent variables right and this is nothing but error term. So let us say you have got these let us say 4 data points right and let us say this is the line passing through these 4 data points right. So this is predicted value of Y for given Xi value.

So the value of Y is this for a given value of X. If you look at this particular point for this particular point which is at X1 let us call X1 or let us say let it be 3 for simplicity. If it is at 3 then there is some value of Y let us say 10 then this is the error term in it and this is explained variance this is known as unexplained variance and this is known as total variance right. So you have got error term now keep in mind this error term is + and – both.

It is you cannot have just positive error; error can be negative as well right just see here right. So always keep in mind that the error term will be positive and negative both.

(Refer Slide Time: 12:44)



Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

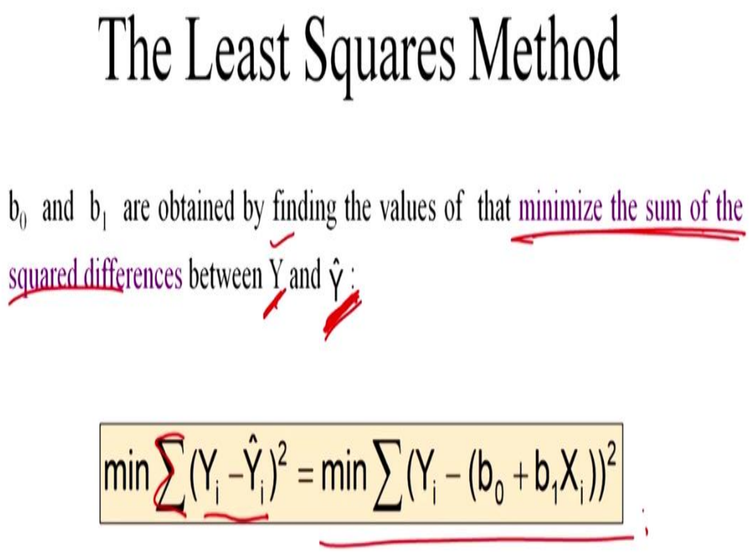
Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

Now let us talk about prediction line and the line which will predict the value of dependent variable for a given value of independent variable. So this Y so estimated Y value for observation i right. This is again intercept b0 this b1 and this is independent variable Xi right. So this is simple linear regression equation provide an estimate of the population regression line right. So this is estimation equation.

(Refer Slide Time: 13:28)



The Least Squares Method

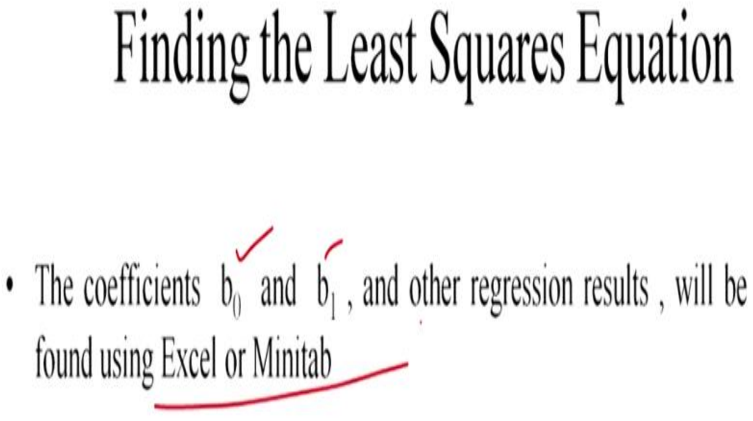
b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\boxed{\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2}$$

You can draw a line which we will call it line of best fit passing through several data points right. In fact, you can think of this line as a line of best fit right. So this line of best fit can be drawn. So the least square method is this so you can have b0 and b1 are obtained by finding the values that minimizes the sum of the square differences between Y and Y right. So this is the estimated value of Y and this is given value of Y.

So we want to minimize the sum of squared differences right just see this is sum of squared and difference and this is how you can minimize.

(Refer Slide Time: 14:33)

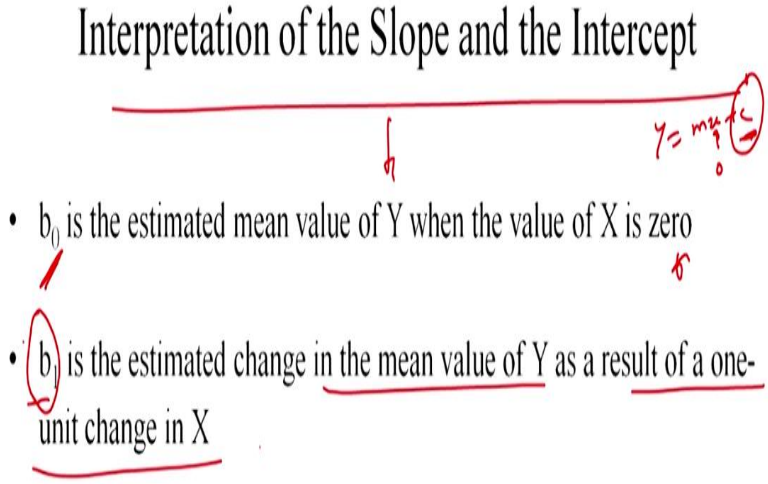


Finding the Least Squares Equation

- The coefficients b_0 and b_1 , and other regression results , will be found using Excel or Minitab

So the coefficient b0 and b1 and other regression results will be found using Excel or Minitab or in fact there are formula available for finding b0 and b1 but you can always use a software may be SPSS, SASS or let us say Minitab or any other statistical software.

(Refer Slide Time: 15:09)



Interpretation of the Slope and the Intercept

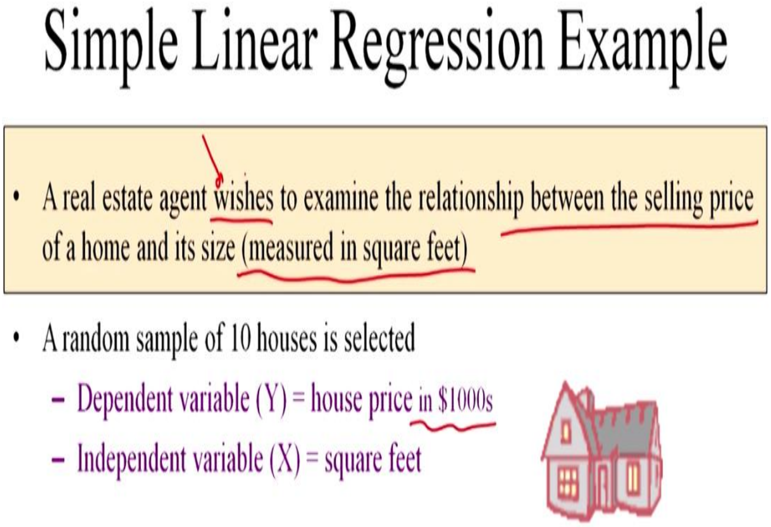
y = mx + c

0

- b_0 is the estimated mean value of Y when the value of X is zero
- b_1 is the estimated change in the mean value of Y as a result of a one-unit change in X

How to interpret slope and the intercept? So as I said, b0 is the value Y when X= 0 right. As I said y = mx + c. So if I keep X=0 then this becomes Y=C so this becomes b0 right which is this right b1 is the estimate change in the mean value of Y as a result of one unit change in X. So if you change independent variable X1 let us say there is only one independent variable X if you change its value by one unit then the change in value of Y will be given by b1 which is known as coefficient of independent variable.

(Refer Slide Time: 16:10)

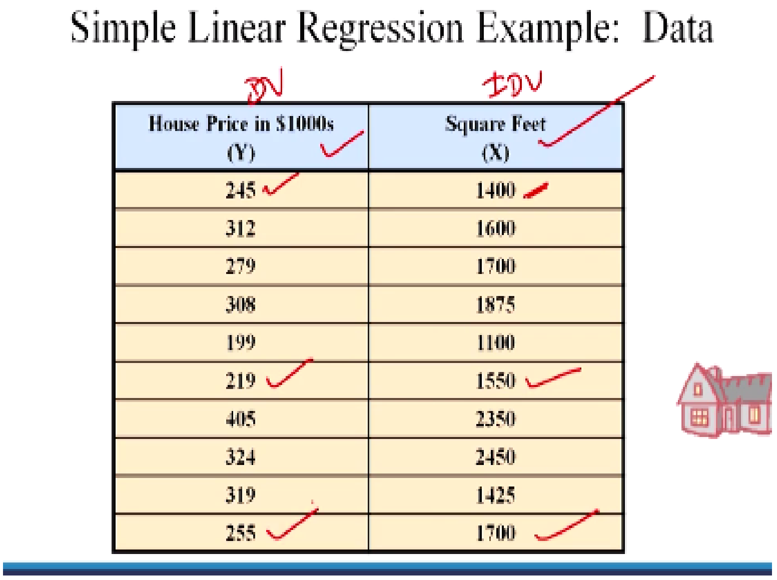


Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

So let us solve a question, a real estate agent wishes to examine the relationship between selling price of a home and its size. So size is measured in terms of square feet and selling price is measured in terms of dollars. So independent variable is size of the plot and dependent variable is price. So the agent, state agent wishes to examine relationship, so is there any relationship between these 2.

(Refer Slide Time: 16:50)



Simple Linear Regression Example: Data

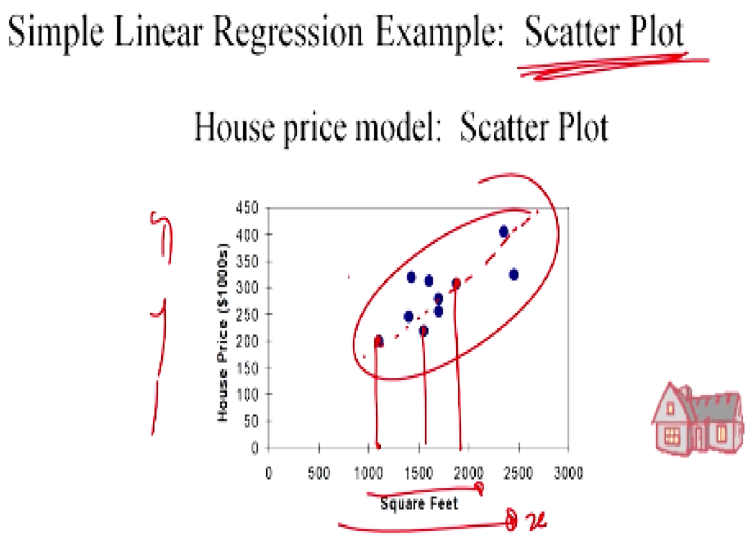
DV IDV

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

So some data were collected on house price and square feet the size of the plot. So this is your independent variable and this is dependent variable. So when size is 1400 square feet price is this much 245,000 for simplicity we are not writing 00 and 0 here right. So when size is this price is this when size is this price is this. So we want to know is there any relationship between these 2.

So for that first of all you can get an idea whether there is any relationship exists between 2 variables or not just by drawing scatter plot.

(Refer Slide Time: 17:45)



Simple Linear Regression Example: Scatter Plot

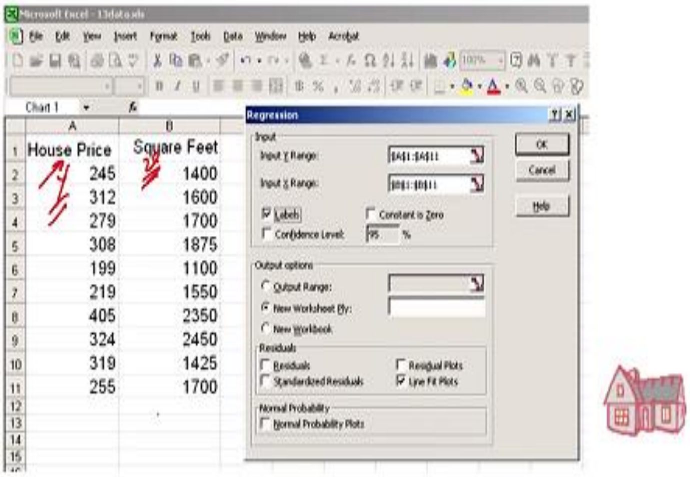
House price model: Scatter Plot

So draw a scatter plot first of all. So if you look at scatter plot where this is your X independent variable and this is Y which is dependent variable house price right. So you can see that when this is the size this is the price when this is the size this is the price when this is the size this is the price right. So if you look at there is some trend which is in positive direction is not it.

We do not whether this relationship is strong or weak, but it seems that there is some positive relationship right because as you increase size of the plot the house price increases. So there is some positive relationship So how to check that positive relationship.

(Refer Slide Time: 18:44)

Simple Linear Regression Example: Using Excel



Microsoft Excel - 13data.xls
File Edit View Insert Format Tools Data Window Help Acrobat
Chart 1
A
B
Regression
OK
Cancel
Help

	House Price	Square Feet
1	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700

Input

Input Y Range: $A$1:$A$11

Input X Range: $B$1:$B$11

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

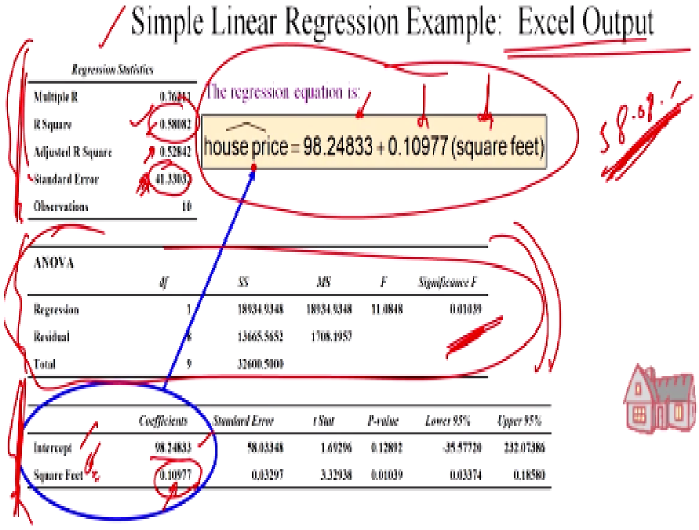
□ Standardized Residuals ☑ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Let us look at this question so you can solve this question using just excel. So you have got house price which is Y dependent variable and X is size of the plot right square feet. So you can just give inputs as Y= this and X= this right. After that you just click it okay you will get output.

(Refer Slide Time: 19:16)



Simple Linear Regression Example: Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

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

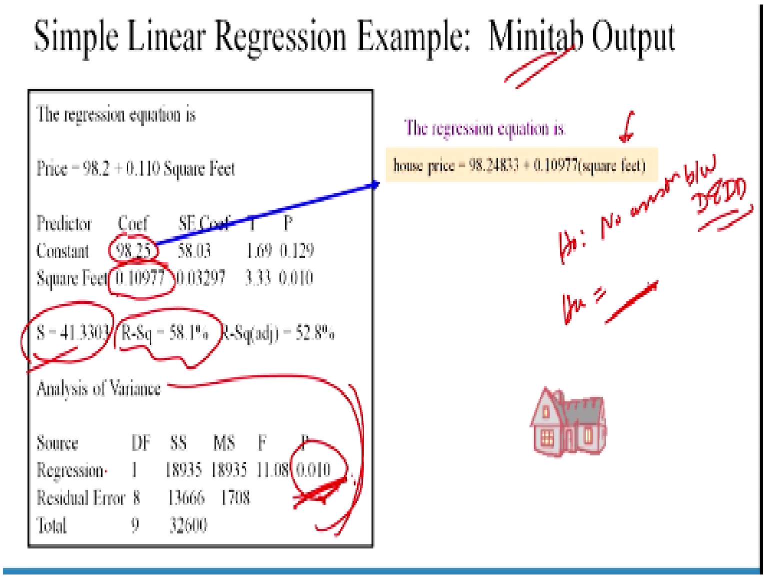
So this is the output of excel. So if you look at output it has got 3 different parts this is first part of the table, this is the second part and this is third part right. So the first part you are getting regression statistics. So regression statistics, in fact you need to remember R square and adjusted R square and standard error right. So we will say that if you look at intercept over here so intercept is this 98.24 + this is your coefficient right and this is your independent variable and this is your Y value.

So house price = 98.24+ 0.109 square feet so this is our regression equation right and we have got one more information over here so R square is this 58.082 percentage right. So in other words we can say yeah it is 58.082 right percentage. So we say that the variation in dependent variable are this much variation in dependent variable is explained by independent variable plot size right and this is error.

So we will say that this is explained variance this is unexplained variance and if you add these 2 it should be total variance right. So look at regression statistics specially R square adjusted R square and standard error and you will coefficient over here intercept and the coefficient of independent variable tight. You will get ANOVA table as well which will tell us how good this model is right.

So for the time being in fact you can skip this part of the table, but keep in mind the lowest portion of the table and this part right. So this output from excel.

(Refer Slide Time: 22:00)



Simple Linear Regression Example: Minitab Output

```
The regression equation is

Price = 98.2 + 0.110 Square Feet

Predictor      Coef     SE Coef     T     P
Constant      98.25     58.03      1.69  0.129
Square Feet  0.10977   0.03297     3.33  0.010

S = 41.3303   R-Sq = 58.1%   R-Sq(adj) = 52.8%

Analysis of Variance

Source          DF    SS     MS     F      P
Regression       1   18935  18935  11.08  0.010
Residual Error   8   13666   1708
Total            9   32600
```

The regression equation is:

house price = 98.24833 + 0.10977(square feet)

H_0: No asson b/w D&ID

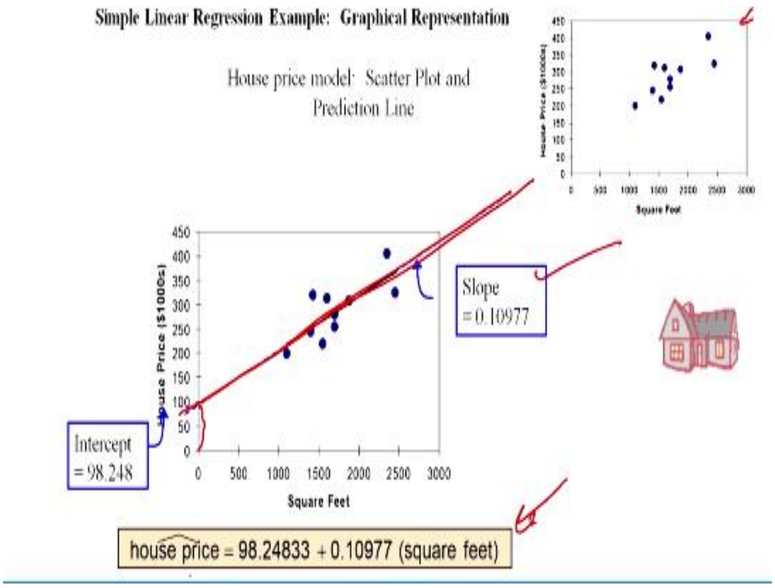
H_a =

You will get almost similar output from Minitab. So you will get the same coefficient right so just see this 98.25 this is coefficient of independent variable .10977 square feet. And if you look at R square which 58.1% is not it and standard error is this much. So we should have a model in such a way that R square should approach towards 100% and standard error should be as low as possible of course ANOVA table is again available here.

So we will say that when we say this regression P is 0.01 so we will say that there is a significant relationship between independent variable and dependent variable. In fact, you can frame and you can always frame a null hypothesis that there is no association. So H0 will be no association between and independent variables. So alternative hypothesis is there is association so you are rejecting null hypothesis it means there is association right.

So this is your scatter plot and this is your line of best fit which you have just calculated right this is nothing but your line of best fit right.

(Refer Slide Time: 23:49)



Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line

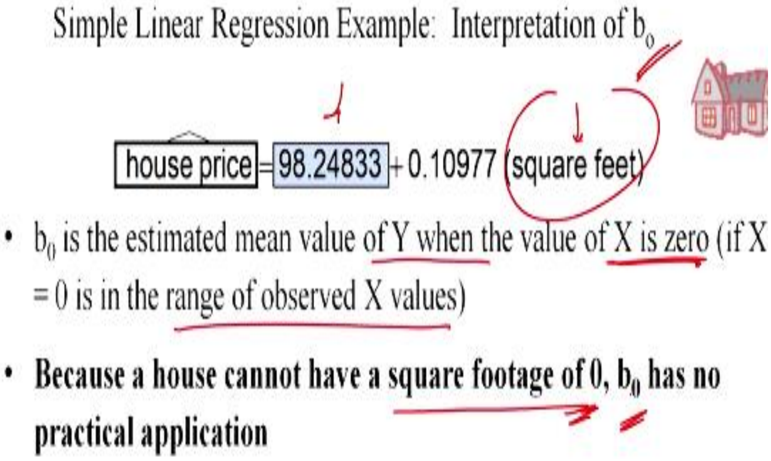
Slope = 0.10977

Intercept = 98.248

$\widehat{\text{house price}} = 98.24833 + 0.10977\text{ (square feet)}$

So this is how this line is passing from here. So slope is given intercept is this 98.2 just see this value this is 98.24 it is not exactly 100 okay. So if you have got this equation then you can find out house price for a given value of square feet okay.

(Refer Slide Time: 24:16)



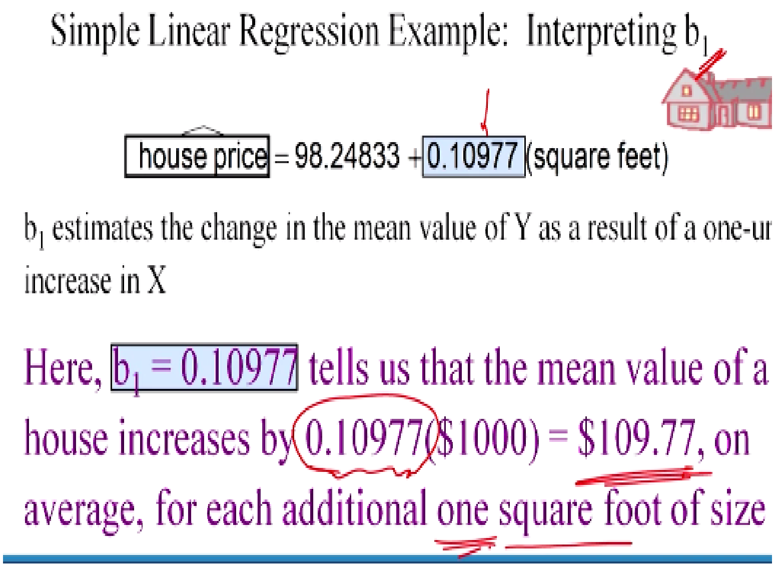
Simple Linear Regression Example: Interpretation of b_0

$\boxed{\widehat{\text{house price}}} = \boxed{98.24833} + 0.10977\text{ (square feet)}$

- b_0 is the estimated mean value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- **Because a house cannot have a square footage of 0, b_0 has no practical application**

Let us look at interpretation of this b0 once again. So b0 is something when you put square feet= 0 right so b0 is estimated when mean of mean value of Y when the value of X is 0. So if X is 0 in the range of observed value of X so if you put this =0 this is nothing but b0 value. So because a house cannot have a square footage of 0, b0 has no practical implication as far as this question is concerned.

(Refer Slide Time: 24:57)



Simple Linear Regression Example: Interpreting b_1

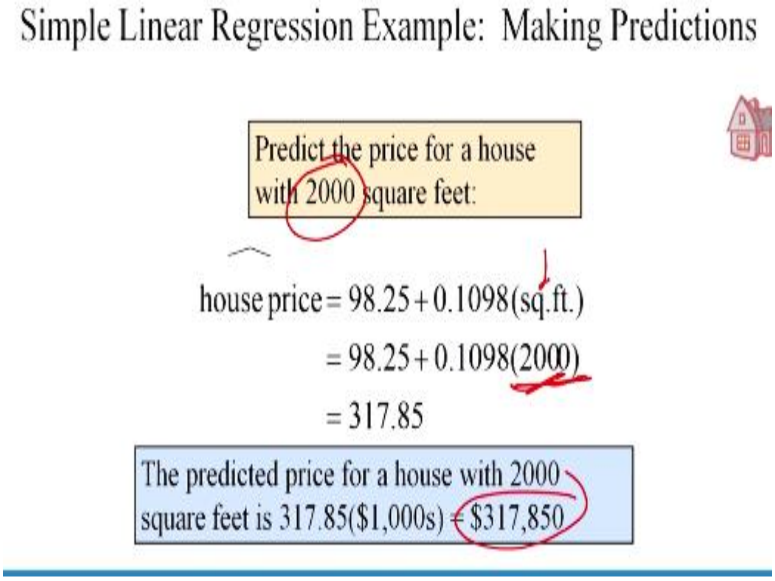
$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

b_1 estimates the change in the mean value of Y as a result of a one-ur increase in X

Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by 0.10977($1000) = $109.77, on average, for each additional one square foot of size

Let us look at how to interpret b1 which is the coefficient of independent variable so b1= this right. It tells us that the mean value of house increase by this much rupees or this much dollar. If we increase size of the plot by one square feet this is the meaning of b1.

(Refer Slide Time: 25:31)



Simple Linear Regression Example: Making Predictions

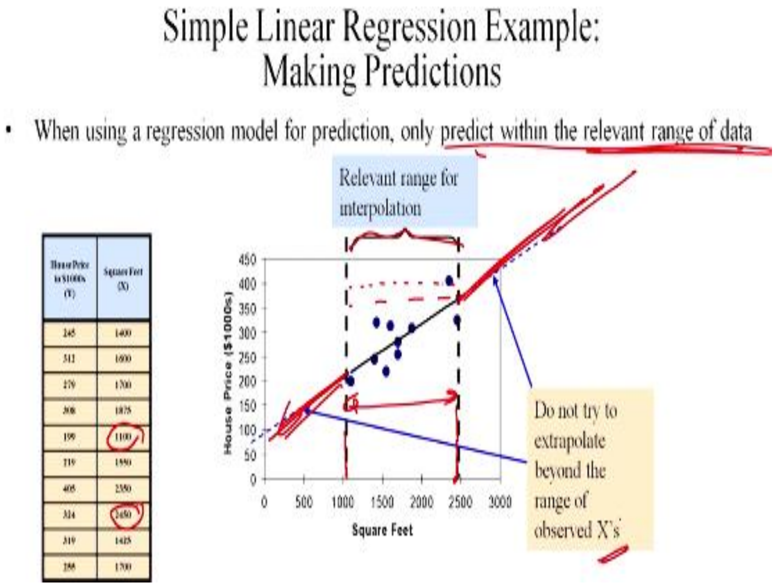
Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$
$$= 98.25 + 0.1098\,(2000)$$
$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

So let us say if you want to predict the price for house with size this much 2000 square feet. So just put over here 2000 in this equation. So the house price would be this much dollars right. So you can find out price of house for any given size of the plot.

(Refer Slide Time: 26:08)



Simple Linear Regression Example: Making Predictions

- When using a regression model for prediction, only predict within the relevant range of data

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

Relevant range for interpolation

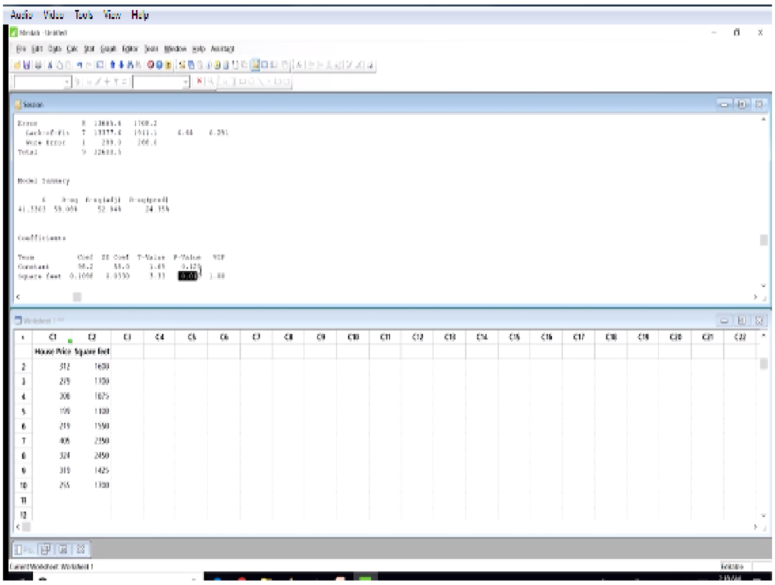
Do not try to extrapolate beyond the range of observed X's

Now it is always suggested to predict in such a way that the value of x should not beyond this limit. It should not go beyond these 2 limits otherwise there would be more and more chances of error. So we should always predict within the relevant range of data and this is nothing but relevant range of data. How did you get this is the minimum and maximum value of X. So if you look at this maximum value of X is somewhere here this one right.

This is something 2450 right this point this line and the lowest is approximately this one right it is 1100 right. So it is always suggested that you should find out the house price for any given value in this area. Do not try to extrapolate beyond this limit or interpolate below this limit okay. So do not try to extrapolate beyond the range of observed axis why you can always extrapolate or interpolate there is not wrong with this equation.

But only the point is the error term would be higher in that case. So try to predict in relevant range of data.

(Refer Slide Time: 28:02)


Audio Video Tools View Help
Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Worksheet 1 ***
Current Worksheet: Worksheet 1

```
Error            8  13665.6   1708.2
  Lack-of-Fit    7  13377.6   1911.1    6.64    0.295
  Pure Error     1    288.0    288.0
Total            9  32600.5


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
41.3303  58.08%     52.84%      34.35%


Coefficients

Term           Coef  SE Coef  T-Value  P-Value   VIF
Constant       98.2     58.0     1.69    0.129
Square Feet  0.1098   0.0330     3.33    0.010  1.00
```

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22
	House Price	Square Feet																				
2	312	1600																				
3	279	1700																				
4	308	1875																				
5	199	1100																				
6	219	1550																				
7	405	2350																				
8	324	2450																				
9	319	1425																				
10	255	1700																				
11																						
12																						

So let us look at how to solve this question using Minitab right. So this is the question right and let us solve this question using Minitab. So house price is dependent variable and square feet is independent variable so 245. So first of all you should enter all data points of first series right it is called house price. Second series is size of the plot of in terms of square feet. So you have got in fact this is the sample data right. So sample size is 10 over here.

So 1875, 1100, 1550, 2350, 2450, 1425 and 1700. So you have got this dependent variable and this one is independent variable right. So go to stat just go to regression then just go to simple regression right regression and fit regression model right. So here response variable is c1 right house prices response variable and predictor variable is square feet right. In fact, either you call it independent variable or predictor variable one and the same thing right.

So in fact you can have a situation where independent variable may be categorical as well right. So we will work out examples when your independent variable is categorical. So just click okay yeah. So let us look at just compare the output of what we have seen with this output. So if you look at R square first of all is this just see this it is 58.08 while in our case what it was, it was also 58.08 same value same answer right from excel output.

Let us compare it with Minitab output right. So this was 58.1% right which is close to 58.08 right. What about standard error standard error is 41.33 here it is 41.33 right same. If you look at the P value of square feet it is 0.010 right let us look at this. So for yeah this is the value P value for square feet for independent variable. It means this is a significant independent variable while for constant if you look at this 0.129 so this insignificant.

So this is how you can solve a question on regression using Minitab software. In next class, we will some more examples on regression and we will take couple of examples on multiple regression as well. So with this I complete my session over here. Thank you very much.