

Marketing Research and Analysis-II
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology Roorkee

Lecture-44
Simple Regression Analysis in SPSS

Hello friends, I welcome you all to the course marketing research and analysis. In the last lecture, we covered concepts of correlation. So, we have understood what is correlation and how it impacts or how it explains the relationship between 2 variables or more than 2 variables. Sometimes, we just find out the relationship between 2 variables and more and sometimes we partial out or we try to control one variable and see the effect of other variables which was the case of partial correlation.

Today we are trying to discuss an another important concept, very, very important concept and very highly utilized in all spheres of education or academics be it management or be it social science, be it engineering wherever you can think of and this is regression. So regression basically is very interesting because this concept actually means to regress towards the mean or to moves towards the mean so what it means something that there is a law of nature tends to average out everything. So, if something has happened you got some high value that the next time when you do the same exercise the chances are fair that you will get a low or lesser value which is closer to the mean.

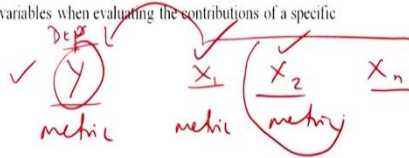
So, regression means to regress or to where does it regress towards the mean, that what it means, so what exactly is regression? It is a very powerful and very flexible approach or procedure for analysing associative relationships between a metric dependent variable and one or more independent variables.

(Refer Slide Time: 02:11)

Regression analysis

Regression analysis is a powerful and flexible procedure for analysing associative relationships between a metric-dependent variable and one or more independent variables. It can be used in the following ways:

1. To determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.
2. To determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.
3. To predict the values of the dependent variable.
4. To control for other independent variables when evaluating the contributions of a specific variable or set of variables.



But it has to be it is like you know Y and X and both this Y let us say could X 1, X 2 up to X n, for example so these are both in metric, this is a metric variable and these also have to be metric this is how it is defined. So it can be used the following ways to determine whether the independent variables, so these are my independent variables, explain a significant variation in the dependent variable. So my dependent variable; so any change in my dependent variable will be explained by a change in the independent variable so what is the relationship, how much of the variation of the variance in the dependent variable is explained through the independent variables.



Second to determine how much of the variation in the dependent variable can be explained by the independent variables first it is whether the relationship exists and second the strength. The third to predict the values of the dependent variable you need to find out calculate how much what is the change in the what is the Y value for a particular X 1, X 2 or particular X value and the third and the last is to control for other independent variables when evaluating the contributions of a specific variable or set of variables supposed I want to find out the effect of the X 1 and Y, I want to control X 2 or I want to take X 1, X 2 and I want to control may be X 3.

(Refer Slide Time: 03:50)

Simple Linear Regression Model

→ Predictive

Linear regression is the next step up after correlation. It is used when we want to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable (or sometimes, the outcome variable). The variable we are using to predict the other variable's value is called the independent variable (or sometimes, the predictor variable). **For example**, you could use linear regression to understand whether exam performance can be predicted based on revision time; whether cigarette consumption can be predicted based on smoking duration; and so forth. If you have two or more independent variables, rather than just one, you need to use multiple regression.



4

So we will start with the simple linear regression model, so as it says linear regression is the next step after correlation, so you have understood correlation, very clearly you have understood correlation. Now it is the next step, so it is used when we want to predict so this regression analysis is also called as predictive analysis, sometimes you must have heard the terms predictive analysis. To predict the value of a variable, which variable now Y the dependent variable based on the value of another variable which is the X the independent variable. The variable we want to predict is called the dependent variable or the outcome variable.

The variable, which was using to predict the X is called the independent variable or the predictor variable. For example you could use linear regression to understand whether exam performance can be predicted based on the revision time. So how much time a student is using for revision for revising his subject that will have an impact on how he is performing in the exam. Another example whether cigarette consumption can be predicted, can we predict how many cigarettes does a person consume?

Based on the time he is spending on smoking, duration of smoking and so for. If you have 2 or more independent variables, rather than just one we say it is a case of a multiple regression. Since there are multiple independent variables, they are called multiple regression.

(Refer Slide Time: 05:35)



Simple Linear Regression in SPSS

Assumption Y X

Assumption #1: Your two variables should be measured at the continuous level (i.e., they are either interval or ratio variables). Examples of continuous variables include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

Assumption #2: There needs to be a linear relationship between the two variables.

Assumption #3: There should be no significant outliers.



5

Now what are the assumptions of simple linear regression or regression for example, a 2 variables should be measured at the continuous level I said that both Y and X we need to be measured in a continuous or metric variable so which is either interval or ratio, examples of continuous variables include for example, revision time measured in hours, intelligence measured in IQ score, performance measured from 0 to 100% or something, weight measured in kg etcetera.

Assumption 2 there needs to be a linear relationship between the 2 variables. That means what we are saying the linear relationship means they are moving linearly that means, they have a proportionate the movement the change, there is an equal change there is a proportionate change. Third there should not be any significant outliers because if there is outlier it will drastically distort the entire relationship.

(Refer Slide Time: 06:41)

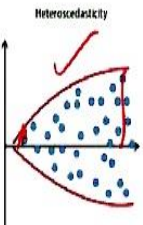
Simple Linear Regression in SPSS

Assumption #4: You should have independence of observations, which can be easily checked using the Durbin-Watson statistic, which is a simple test to run using SPSS Statistics.

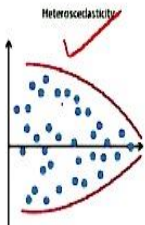
autocorrelation

Assumption #5: Your data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line.

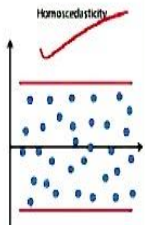
Heteroscedasticity



Heteroscedasticity



Homoscedasticity



You should have independence of observations so this independence of observation can be checked by the Durbin-Watson statistics test which is also a measure of autocorrelation. Now just forget this autocorrelation. So if you want to check it you can check through this also, whether the presence of autocorrelation is there or not and this helps you to find out independence of observation but just imagine and understand independence of observation we mean that no variable or no respondent is repeated for more than one time that means it has; it gets a chance to be a part of the study for once only.

Until and unless we are doing a repeated measured design, so you had understood earlier experimental design so in that we are saying in our factorial design whatever we are doing so there we are using the independence of observation assumption that means one respondent can be part of study for once only and if the sample is repeated for let say 2 times, so once for year 1 once for year 2 then it is called a repetition of repeated measured observation.

The last assumption is the data need to be homoscedastic. What is homoscedastic? This is where the variances along the line of best fit remain similar, now look at to this case, so homo means similar, so similar variance that means the variance is similar and if the variance is not similar it is hetro so these 2 cases are hetroscedastic you see the variance here and you see the variance here.

So there is a large gap, this is significant difference. But look at the gap of the plots are the data points here and here everywhere it is more or less the same. This homogeneity of variance is very; very important assumption right, homoscedastic or the homogeneity of variance is very important assumption in any statistical analysis.

(Refer Slide Time: 08:42)

Simple Linear Regression in SPSS

Assumption #6: Finally, you need to check that the residuals (errors) of the regression line are approximately normally distributed (we explain these terms in our enhanced linear regression guide). Two common methods to check this assumption include using either a histogram (with a superimposed normal curve) or a Normal P-P Plot.

Example

A salesperson for a large car brand wants to determine whether there is a relationship between an individual's income and the price they pay for a car. As such, the individual's "income" is the independent variable and the "price" they pay for a car is the dependent variable. The salesperson wants to use this information to determine which cars to offer potential customers in new areas where average income is known.

One more thing is you need to check the residuals or errors. Now what are the errors? Please understand them, many of times students get confused. The error is basically the unexplained part in a study. Now let me explain it through may be a diagram if possible so what I will do is I will draw it here, I have some space. So Let us say this is my X and Y, this is my Y, this is my X, now let us say I have this is my Y mean, so the value of Y is the average value of Y.

Now let us take a variable and we are assuming, now this is where I am saying is my Y estimated, now what is Y estimated? Means the Y which you have calculated, the calculated value of Y the dependent variable that you have calculated, now let us take a variable here, now this is my Y observed, so although my estimated calculated Y is this much, my mean is this much but this is my real Y or my observed Y.

Now if you look at this part, this entire thing, this total from here observed minus the mean, this part is called my total variance. Now so $Y = Y - \bar{Y}$ summation of this, similarly if you look at the Y, this \bar{Y} estimated Y - the \bar{Y} , So Y estimated - \bar{Y} mean this gives me what, now this is the one which we call as the explained variance, so explained variance. So this is called sum of square of regression also it is sometimes it is donated as sum of square regression. What is this part? The Y the actual Y - the estimated Y, now this part is what is called my unexplained variance and this unexplained variance is called as sum of square of error or residual.

So, when I said this keep residual is nothing but this part, so you need to check the residuals of the regression line are approximately normally distributed, this should be normally

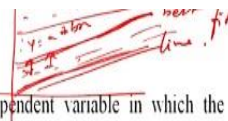
distributed, 2 common methods to check is through a normal PP plot or histogram or I have shown explained when I was taking about data purification you can easily go and calculate and find out whether the live in the range of -2 to +2 and if it is yes then it is within the range and then we say it is normally distributed if it is beyond -2 to +2 then we say is not normally distributed.

Otherwise just you can go through a histogram or a normal plot and PP plot and find out. Let us take this case. A salesperson for a large car brand wants to determine whether there is a relationship between an individual's income and the price they pay for a car. So you watch now is there is any relationship between individual income and price.

As such, the individual income is the independent variable and the price they pay is my dependent variable. The salesperson wants to use this information to determine which cars to offer to the potential customers. Suppose a customer is a very income customer so which car should we will be shown, so to offer potential customer, new area where average income is known to them.

(Refer Slide Time: 12:27)

Simple Linear Regression Model



Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line is called simple linear regression.

Regression analysis involving two or more independent variables is called multiple regression analysis.

Simple Linear Regression: Example

Anand's Pizza Parlors is a chain of Indian-food restaurants located in a five-state area. Anand's most successful locations are near college campuses. The managers believe that quarterly sales for these restaurants (denoted by y) are related positively to the size of the student population (denoted by x); that is, restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population. Using regression analysis, we can develop an equation showing how the dependent variable y is related to the independent variable x.

So as it says regression analysis involves one independent variable and one dependent variable which you have understood, in which the relationship is approximated by a straight line. Now what is this straight line? So let me explain that point, so for example if you see is something like this so we say this is the regression line let say and this line is called the best fit line. Why it is called the best fit line?

Because there could be infinite lines but this is line which we are saying the best line why because if you take all the data points and calculate the variances, the distance from this line then you will see it is in this line that the variance is the minimum, if you would had taken any other line like say this one or let us say here or somewhere here then the variance would be more in comparison to when you take this line as the regression line.

So we have understood that this is a case of simple linear regression and when we are using more than 2 independent variables, 2 or more then we say it is a case of multiple regression analysis. Now let us go to the example Anand pizza parlour is a chain is a food restaurants located in 5 states. The most successful locations are near college campuses.

The managers believe that quarterly sales for these restaurants are related positively to the size of the student population, which is X, that is restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population obviously you know pizza's are more liked by the students. So now using regression analysis, we can develop an equation showing how the dependent variable Y is related to the independent variable X. So what is Y? Sales, what is X? My size or population.

(Refer Slide Time: 14: 34)

Simple Linear Regression Model

Regression Model and Regression Equation

In the Anand's Pizza Parlors example, the population consists of all the Anand's restaurants. For every restaurant in the population, there is a value of x (student population) and a corresponding value of y (quarterly sales). The equation that describes how y is related to x and an error term is called the regression model. The regression model used in simple linear regression follows.

SIMPLE LINEAR REGRESSION MODEL

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

β_0 intercept + β_1 slope

β_0 and β_1 are referred to as the parameters of the model, and the Greek letter epsilon is a random variable referred to as the error term. The error term accounts for the variability in y that cannot be explained by the linear relationship between x and y .

So now in Anand's example the population consists of all the Anand's restaurant for every restaurant there is a value of x student population and a corresponding value of y sales. The equation that describes how y is related to x and error term in the model, so you see $y = b_0$ or this b right, $+ b_1x + e$, now what is this let us explain where b_0 or b_0 and b_1 are referred to as

the parameters of the model so this parameter is called my intercept and this is called my slope.

So I did you had more number of x for example x 1, x 2, x 3 so there would be b1, b2, b3 goes on but what is this error I just now explained, the unexplained variance or the residuals or the errors ok, so these are the thing we have the Greek letter epsilon is a random variable referred to as the error term this one. The error term accounts for the variability in y that cannot be explained by the relationship between x and y.

So let us understand again let us draw a line this is my let say this is a regression line. Now this is called my intercept, the one we choose as seen here b0 and this is my b1, the slope, so y is the b1 x which we are talking about so this is the Beta 1 so the slope.

(Refer Slide Time: 16:07)

Simple Linear Regression Model

The population of all Anand's restaurants can also be viewed as a collection of subpopulations, one for each distinct value of x. For example, one subpopulation consists of all Anand's restaurants located near college campuses with 8000 students; and so on. Each subpopulation has a corresponding distribution of y values. Thus, a distribution of y values is associated with restaurants located near campuses with 8000 students; and so on. Each distribution of y values has its own mean or expected value. The equation that describes how the expected value of y, denoted $E(y)$, is related to x is called the regression equation. The regression equation for simple linear regression follows.

SIMPLE LINEAR REGRESSION EQUATION

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

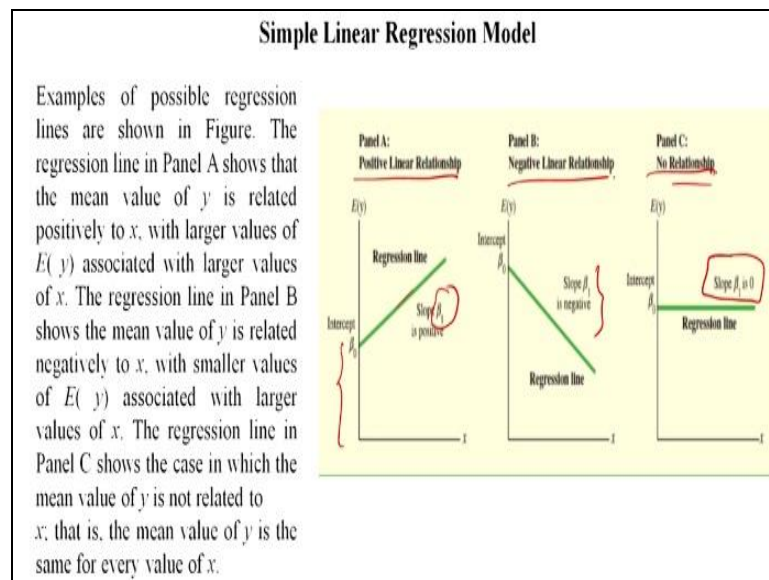
y_1 8000
 x_1 7000
 y_2 9000

Now continuing the population of all Anand's restaurant can be viewed as a collection of subpopulations, one for each distinct value of x, so x was my size of the population. For example one subpopulation consists of all the restaurants of Anand located near college campuses with 8000 students, and so on. Each subpopulation has a corresponding distribution of y values.

So, for each value of x there will be a corresponding value of y obviously, if x is 8000 the y is something let us say y 1, If it is 7000 let us say y 2, it is 9000 let us say y 3 so we have corresponding values, thus a distribution of this y values, this distribution of this y values y 1, y 2, y 3 goes on y n is associated with restaurants located near the campuses.

So each distribution of y values as it is own mean or expected value that is what I was trying to say when I started the lecture that regression is regressing towards the mean, so what is the mean or expected value the equation that describes how the expected value of y is related to x is called the regression equation. So the regression equation is $E(y) = b_0 + b_1 x$ that means my intercept + my slope * the independent variable.

(Refer Slide Time: 17:34)



Now this is about the relationship you see in correlation also you have seen this so if you see, if I take a slope between x and y , so this part is my intercept as I shown b_0 , and this is my regression line, these are the regression lines, so in this case you see this case the b_1 is positive. So when it is positive we say it is positive linear relationship that means as x is increasing y is also increasing. But in this case as x increasing you see the value of y is decreasing, so the slope is negative in this case, it is a negative linear relationship. The third case you see with the change in x , there is no change in the y , so the slope is 0, so this is no relationship.

(Refer Slide Time: 18:24)

Simple Linear Regression Model

If the values of the population parameters β_0 and β_1 were known, we could use previous equation to compute the mean value of y for a given value of x . In practice, the parameter values are not known and must be estimated using sample data. Sample statistics (denoted b_0 and b_1) are computed as estimates of the population parameters β_0 and β_1 . Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1 in the regression equation, we obtain the estimated regression equation. The estimated regression equation for simple linear regression follows.

β_0, β_1

ESTIMATED SIMPLE LINEAR REGRESSION EQUATION

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

If the values of the population parameters b_0 and b_1 the intercept and the slope were known, we could use previous equation to compute the value of y for a given value of x . Obviously if I need to find the value of y the estimated y , the estimated value of the dependent variable and I know my x but I do not these 2, I need to calculate. So in practice, the parameter values are not known and must be estimated using the sample data.

Since you cannot infer it from the population is difficult so you have the sample so let us use the sample and because it represent the population and calculate the b_0 and b_1 the slope. So sample statistics b_0 and b_1 are the reflection of the same b_0 for the population. So are computed as estimates of the b_0 and b_1 . Substituting the values of the sample statics b_0 and b_1 for β_0 so that means the intercept is 0, so if intercept is 0 that means it is starting from the origin and 1 in the regression equation. We obtain the estimated regression equation, so this is how it is looks like.

(Refer Slide Time: 19:36)

Least Squares Method

The least squares method is a procedure for using sample data to find the estimated regression equation

To illustrate the least squares method, suppose data were collected from a sample of 10 Anand's Pizza Parlor restaurants located near college campuses. For the i th observation or restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars). The values of x_i and y_i for the 10 restaurants in the sample are summarized in the Table

Restaurant i	Student Population (1000s) x_i	Quarterly Sales (\$1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

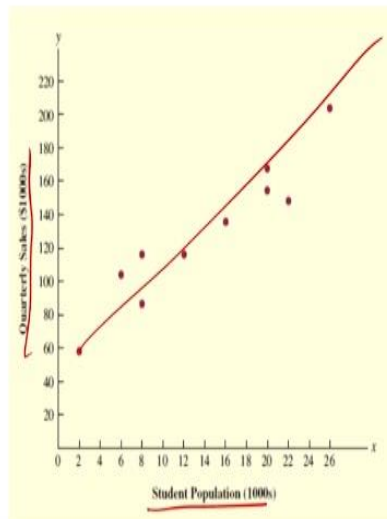
Now the least square method, so least square is a procedure for using sample data to find the estimated regression equation. So, in the case of this Anand's case there are 10 restaurants, student population in 1000 is given, 2000, 6000, 8000, 8000, 12000 up to 26000. The sales are given to us. So now it says 10 Anand's pizza restaurants data is given, and the x_i is the size of the student population in 1000's, y_i is the quarterly sales in 1000's of dollars. So the values of x_i and y_i for the 10 restaurants are given to us.

(Refer Slide Time: 20:15)

Least Squares Method

Student population is shown on the horizontal axis and quarterly sales is shown on the vertical axis. Scatter diagrams for regression analysis are constructed with the independent variable x on the horizontal axis and the dependent variable y on the vertical axis. The scatter diagram enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.

What preliminary conclusions can be drawn from Figure???



Now if I draw a just I draw the data points, if I draw the data points this is what it says, student population is shown on the horizontal axis and quarterly sales is shown in the y axis. So scatter diagram is just representation of the data points, for the regression analyses are constructed with the independent variable x and y . The scatter diagram enables us to observe the data graphically and to draw preliminary conclusion you can draw any final conclusion but preliminary conclusion about the possible relationship.

Now just see this and tell me what is the relationship you think? So this is a positive relationship, it is growing, as x is growing y is growing, if this is my regression line so this is a positive relationship, from this figure we can find out that

(Refer Slide Time: 21:10)

Least Squares Method

Quarterly sales appear to be higher at campuses with larger student populations. In addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a straight line. Indeed, a positive linear relationship is indicated between x and y .

We therefore choose the simple linear regression model to represent the relationship between quarterly sales and student population. Given that choice, our next task is to use the sample data in Table to determine the values of b_0 and b_1 in the estimated simple linear regression equation. For the i th restaurant, the estimated regression equation provides

$$\hat{y}_i = b_0 + b_1 x_i$$

where

\hat{y}_i = estimated value of quarterly sales (\$1000s) for the i th restaurant

b_0 = the y intercept of the estimated regression line

b_1 = the slope of the estimated regression line

x_i = size of the student population (1000s) for the i th restaurant

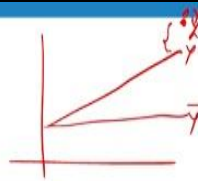
Now quarterly sales appear to be higher at campuses with large student populations, in addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a straight line, indeed, a positive relationship. We therefore choose the simple linear regression model to represent the relationship between quarterly sales and student population.

Given that choice, our next task is to use the sample data and determine the value of b_0 and b_1 the simple parameters in the estimated equation. If we find out these for suppose b_0 and b_1 then for any x_1 we can find the value of y . So y is my estimated value of quarterly sales, b_0 the y intercepts, b_1 the slope and x_1 is the student population size.

(Refer Slide Time: 22:04)

Least Squares Method

The criterion for the least squares method is given by expression



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

OV - EV = 0

So the criteria is saying y this is what I was explaining, the minimum the criteria least square criteria is the summation of $(y_i - \hat{y})^2$.

So I had explained you so least square what it is says the minimum, minimize this, what is this minimize is just go back let us go back to the diagram, so I said earlier this was my estimated y , this is my y bar and we have a observed value y , so this part was my unexplained part.

So the intension of the researcher is always to minimize this part and to reduce it as much as possible, so that means what our estimated value should be able to incorporate or accommodate the actual observed values. So y is my observed value of the dependent variable for the i th observation and y_i is my estimated value.

So if my observed value minus my estimated value is equal to 0, that means what observed value is equal to my estimated value that means what we can say that it lies on the same point which means that there is no unexplained variance, everything is explained in the study. It is a very important concept because the more unexplained you have that means the researcher has very little control over the research.

(Refer Slide Time: 23:35)

Least Squares Method

SLOPE AND y-INTERCEPT FOR THE ESTIMATED REGRESSION EQUATION

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

$$b_0 = \bar{y} - b_1\bar{x}$$

where

x_i = value of the independent variable for the *i*th observation

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

\bar{x} = mean value for the independent variable

\bar{y} = mean value for the dependent variable

n = total number of observations

So how do you calculate the slope and intercept? First let us calculate the slope, so the formula is if you see

b_1 = summation of $[(x_i - \bar{x})(y_i - \bar{y})]$ and divided by the summation of $(x_i - \bar{x})^2$.

So this is the x_i is my value for the independent variable for *i*th observation, y_i is value for the dependent variable for the *i*th observation, \bar{x} is mean value for the independent variable, \bar{y} is mean value for the dependent variable, n is my total number of observations. Once we get this and I have let us say \bar{y} and \bar{x} the estimated, the mean value of y we have and the \bar{x} value then we can calculate b_0 . So let us do this.

(Refer Slide Time: 24:27)

Least Squares Method

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

Restaurant <i>i</i>	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	$\sum x_i$	$\sum y_i$			$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x_i - \bar{x})^2$

So first we find out these are the 10, so this is my x_i the values which earlier also we have shown and this is my y . Now first we calculate the \bar{x} so \bar{x} is how much $140/10$ so that is equal to 14, \bar{y} is how much $1300 / 10 = 130$. We need $(x - \bar{x})$ so we find out $x - 14$ so $2 - 14, 6 - 14,$

8 - 14, 8 - 14, 12 - 14 goes on till 26 - 14, so $x - \bar{x}$ similarly $y - \bar{y}$ 58 - 130, 105 - 130 and goes on till 202 - 130. Now we want $(x - \bar{x}) * (y - \bar{y})$.

So we can do this, calculate this multiply and we have found out this. Now $(x - \bar{x})^2$ is what we are finding out so $(x - \bar{x})^2$, 12 square is 144, 8 square 64 goes on. So after finding everything, see this is the formula, so we have got everything now with us, so let us use the formula
(Refer Slide Time: 25:33)

Least Squares Method

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

$$= \frac{2840}{568}$$

$$= 5$$

The calculation of the y intercept (b_0) follows.

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= 130 - 5(14)$$

$$= 60$$

Thus, the estimated regression equation is

$$\hat{y} = 60 + 5x$$

So this part $(x - \bar{x}) * (y - \bar{y})$ this is how much 2840 / 568 so my slope is equal to 5. Now we calculate the intercept, so calculate the intercept b_0 we have \bar{y} is my mean, so 130 - 5 the slope * \bar{x} the mean is 14 here so this gives me the value of the intercept is 60 what is the intercept mean it means that whenever you do not have any value for x that means when x is 0 still there is some value for y and this value for y is nothing but the intercept, the meaning of this is that.

When $x = 0$ whatever value of y remains that is my intercept. So the estimated regression equation is now y estimated is equal to 60 this is my slope + 5 the slope this looks like, how do you look $b_0 + b_1x$ so b_1 is my 5 * x. So for any new variable any new value of x now you can calculate the value of y, you can estimate.

(Refer Slide Time: 27:02)

Least Squares Method

The slope of the estimated regression equation ($b_1 = 5$) is positive, implying that as student population increases, sales increase. In fact, we can conclude (based on sales measured in \$1000s and student population in 1000s) that an increase in the student population of 1000 is associated with an increase of \$5000 in expected sales; that is, quarterly sales are expected to increase by \$5 per student.

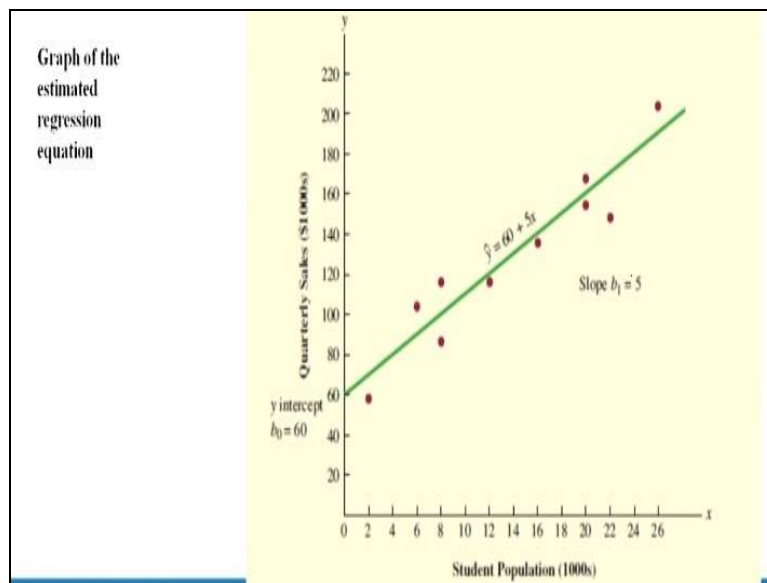
If we believe the least squares estimated regression equation adequately describes the relationship between x and y , it would seem reasonable to use the estimated regression equation to predict the value of y for a given value of x . For example, if we wanted to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\hat{y} = 60 + 5(16) = 140$$

So for example in this case so let us say b_1 is positive. Implying that as a student population increases, sales increase because the slope is increasing, in fact we can conclude that an increase in the student population for 1000 is associated with an increase of 5000 in expected sales that is quarterly sales are expected to grow by 5\$ per student. If we believe the least squares estimated regression equation adequately describes the relationship between x and y , it would seem reasonable to use the estimated regression equation to predict the value of y for a given value of x .

For example, if we wanted to predict the quarterly sales for a restaurant to be located near a campus with 16000 students, we would compute as how y is equal to estimated y is = 60 which is my intercept + 5 is my slope * 1000 we are taking it into only the numeric is 16 so 16000 because that is how you have written. So finally this becomes 140000 the sales 140000. So this is what we understand, so this is how graphically this is how it is shown the regression equation so this is my slope and this is my intercept everything is shown here.

(Refer Slide Time: 28:14)



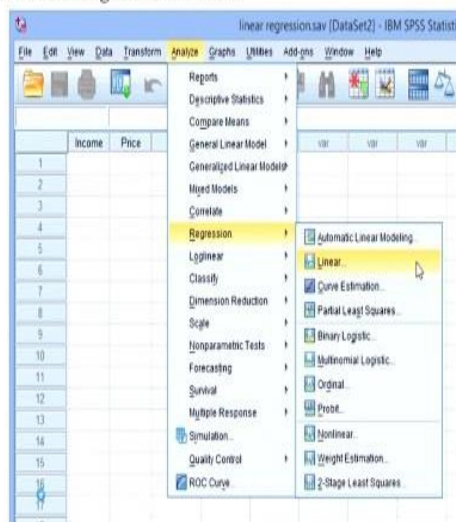
(Refer Slide Time: 28:14)

Simple Linear Regression in SPSS

Test Procedure in SPSS Statistics

Step-1

Click **Analyze**> **Regression**> **Linear...** on the top menu, as shown



Now I will show you how to just do it in the SPSS. First let me show in the slide also, so how do you do it in the regression, go to analyze regression, go to the linear model and then you take what you want as the dependent you take it here and what you to take it independent take it here. So here price and income this is just arbitrary example so we have taken and then we need to check.

(Refer Slide Time: 28:41)

Simple Linear Regression in SPSS

Test Procedure in SPSS Statistics

Step-3 You now need to check four of the assumptions discussed in the Assumptions section above: no significant outliers (assumption #3); independence of observations (assumption #4); homoscedasticity (assumption #5); and normal distribution of errors/residuals (assumptions #6). You can do this by using the statistics and plots features and then selecting the appropriate options within these two dialogue boxes.

Step-4 Click the OK button

So let me show this and the model summary, obviously this later on we will come.

(Refer Slide Time: 28:48)

Simple Linear Regression in SPSS

Output of Linear Regression Analysis

The first table of interest is the **Model Summary** table, as shown below:

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.873 ^a	.762	.749	874.779

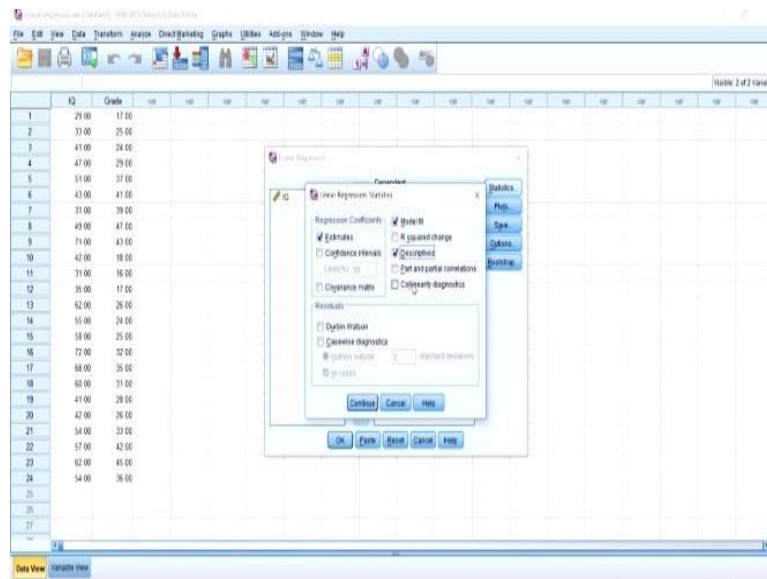
^a Predictors: (Constant), Income

This table provides the R and R^2 values. The R value represents the simple correlation and is 0.873 (the "R" Column), which indicates a high degree of correlation. The R^2 value (the "R Square" column) indicates how much of the total variation in the dependent variable (Price), can be explained by the independent variable (income). In this case, 76.2% can be explained, which is very large.

So, but if you want to understand this understand you will get such kind of descriptive table model summary which this is the R and if you remember this R we had said this is related to the multiple correlation value, anyway let us go to this what is the R square what is the adjusted R square explain. R square is nothing but the square of this value and adjusted R square is something very interesting which let I will explain that adjusted R square is a value which goes on increasing up to a particular level as you increase the number of variables but then after a certain point of time when you add more variables the adjusted R square value actually it is started declining.

That means either no change or it is start declining because the point is that adjusted R square only accommodates those variables which contribute to the data or to the dependent variable or to the study. So let us go to this simple regression.

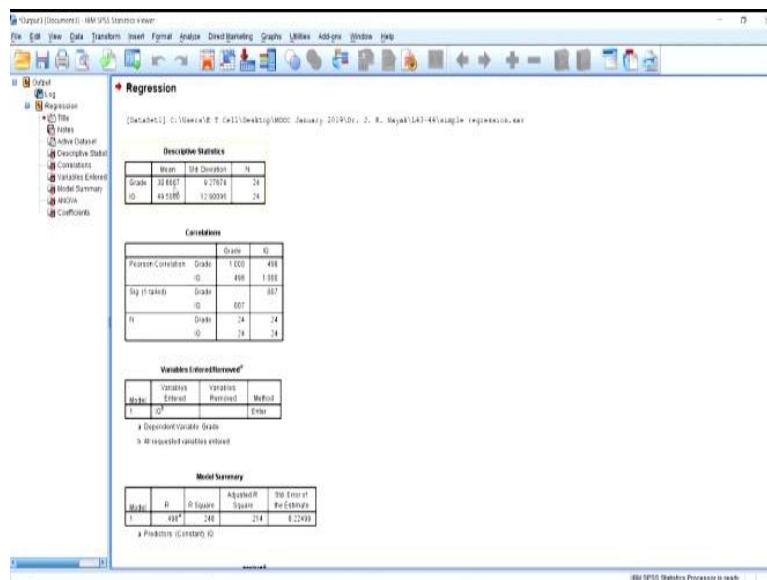
(Refer Slide Time: 29:47)



So simple regression I go to analyze, this case is IQ and Grade. So is Grade effected by IQ, so let us see so I take Grade as my dependent variable and IQ as my independent variable. So which this is a very basic regression we are doing and we are understanding. So let you can go to statics and you can see there are several things I want to check descriptive for example R square change is not required here.

For example independent observation you can check it through a Durbin Watson which is an auto correlation again. Collinearity I will explain it later on, what is the role of Collinearity, what is multiple Collinearity, multi collinearly problem I will explain it later but forget it for a moment. So you want to anything else now. I do not want to do any change here so I just want to run it.

(Refer Slide Time: 30:31)



So I see, you see if you look at this now first what it saying the Grade the mean is 30.67 and standard deviation is 9.2, IQ 49.5, 12.9. Now let us look at the correlation, what is the correlation between Grade and IQ it is 0.498 and is it significant? Yes it is significant at a 0.007 level. Now look at this the model summary, So my R can you go up and see so here you have got a correlation of .98 and here R is .98.

That is what I was trying to explain many a times student do get confused how is this R connected, is this R connected with the correlation or not so well this is the output, this is same as the correlation, multiple correlation coefficient or the correlation coefficient. So R is .948 in this case and does the R square; now R square is nothing but my coefficient of determination.

So if I divide this 1 - coefficient of determination I will get something which can be explained as strength of the test. So, .248 but look at this adjusted R square that means what, when I am taking all variables it is .248 and remember the R value will go on increasing. The R square value and the R value will go on increasing as you add more and more number of variables.

But the adjusted R square, will not, it will remain same or it will not increase. So then we say well this is how the models look like, now this is the ANOVA, ANOVA means it is variances I will explain you may be later on. These 2 terms ANOVA and regression are very strongly correlated also, you can understand each other through one from the other and now what is

the coefficient. Now there is 2 coefficients you can see unstandardized Beta coefficient and standardized Beta coefficient.

Now this is the t value and this is my significance. If you look at this IQ, so IQ the unstandardized coefficient is 0.358, the standardized coefficient is 0.498 which you already got standardized coefficient is my correlation and my t value is 2.694 and it is significant, that means what we can say that IQ plays a significant role in the Grade of a student. So, if higher the IQ because it is positive relationship, the higher is the students Grade.

So this is what it explains this you know this is the simple regression model and it has explained you how to find out the value of the dependent variable from by changing the value of the independent variable and I have explained to you what error terms mean and you should not be afraid of the word error, it is not actually the error it is an unexplained variance. So what is the relationship between explained and unexplained I have explained all.

So I think this is the just a beginning for the regression class, we will be doing more forms of regression, regression can be used in multiple ways so as I said it starts with a basic that both the dependent and independent variable have to be metric in nature but we will see several special cases where it might not be the independent variable might be in some other format may be in a categorical scale but we can still do it, so how we will do it all we will see in the later on future classes in the upcoming classes for today we will close it here, thank You very much.