Thank you very much for attending this session. Today, I will talk about regression analysis.

I will introduce three tools for regression analysis. Data Amalysis ToolPack, Scatter Graph Capabilities, and Excel functions.

Data Analysis ToolPack is a useful add-ins for regression analysis. But it is not the best tool for linear regression between two variables, because whenever you change even one piece of data, you need to create a new report. We usually use Data Analysis ToolPack for multi-variable regression, or when we need some more advanced information about a linear regression between two variables.

For linear regression analysis between two variables, we usually use excel functions like intercept, slope, RSQ, and STEXY. Or scatter graph capabilities like add trend line and display R-Squared.

introduce regression through this data analysis tool pack and it is also

very useful for multiple regulation so i give a brief introduction to recreation

i introduce you to scada graph and data analysis tool and then we set the stage for our

future sessions when we mainly use excel functions here are some

introductory words for regression analysis regression analysis is used

both for prediction forecasting like moving average like exponential smoothing

for prediction and it is also used for association association between

a dependent variable and one or more independent variable the relationship could be

linear or non-linear we will discuss all situations we may use

regression for prediction for example we may take past data of port of los angeles and long beach

or pass data of incoming student to csun based on say 10 years 20 years 25 years data

we may predict what would be the volume of activities next year in

san pedro bay ports ports of los angeles and long beach or what would be the number of incoming

student at csun so that is prediction implementation

but besides prediction we can also use it for association association between two variables or between one

variable and more than one other variable for example

in the association context we may use regression to find the relationship between price and sales

we may use regression to find the relationship between total cost and quantity produced

you may consider relationship association between the number of hours a student put into his or her

education and the academic performance he or she will have by the end of the semester

multi-valuable regulation analysis for example between performance of a student in a semester and

his or her previous gpa number of hours he or she puts in his coursework in this specific semester

number of informal prerequisites that the student may have taken for this course for

example for operations management som 306 the only prerequisite is statistics

but if we analyze we may find it out that the student who also take gateway before this course

perform better that is multi-variable regression analysis which may be linear or non-linear

we use regression for relationship between one variable and one other variable

or between one variable and several other variables the relationship may be linear or

non-linear let me clarify in a very simple language what do we mean by linear and non-linear

in a linear regression the amount of increase in y for each unit increase in x is constant

for example for one unit increase in x we may have 0.2 units increase in y that is a linear relationship and in a

linear relationship we only have constant and constant multiplied by a variable

so we have b0 plus b 1 x equal to y or b 0 plus b 1 x 1 plus b

2 x 2 is equal to y in a non-linear relationship for example like

exponential relationship it is not the amount of the change which remains fixed but the

percentage of change for each unit increase in x is two percent of the previous value

therefore the relationship would be like this this is an example of the differences between

linear relationship and non-linear relationship linear relationships are

only composed of a constant value plus a constant multiplied by a variable y is equal to b0

plus b 1 x 1 plus b 2 x 2 plus b 3 x 3 that is the relationship between

one independent variable and three dependent variables of x1 x2 x3 which we may have 10 observations

20 observations 30 observations for each one and that relationship is linear

so here are different relationships that we may consider relationship between two

variables more than two variables when independent variable is time

and when independent variable is not time in linear fashion and in non-linear fashion if we study

the relationship between two variables usually we benefit from scatter graph scatter graphs can show the linear or

non-linear relationship between two variables usually we refer to the independent

variable as x and dependent variable as y if you have several independent

variables we call them x1 x2 x3 or you can call them x z t p whatever you like

x1 and x2 and x3 does not mean 3 observations it means 3 independent variables for each of those

independent variables we may have 10 period 20 period 30 period observation

for two variables we can use scatter graph with three variables one dependent and two independent

variables we may use bubble chart which in one direction shows

x1 y is on y direction and then the second x x 2 will be shown by bubbles or

circles inside the graph but when we go beyond three variables we cannot visualize them

because our visualization space or physical visualization space is limited

to three dimensions of course as we become better people our visualization

space will expand when we like other people when we love other people

and we try to help other people throughout our life our visualization space in general

will expand but our physical visualization space is still limited to

three timing so we cannot show more than three variables but we can always get one of the independent variables

and observe its relationship with the dependent variable using a scatter graph we can do it

for all other independent variables too now consider this set of data

this is the data that we have for 23 years performance in ports of los angeles and

long beach these are yearly volume of activities and they are in terms of t e

use that means 20 equivalent units so the total volume of

container handling in the ports of los angeles and long beach in 2019 was around

17 million unfortunately after this 17 millions about half of them are full containers

which come from far east united states and about half of them are empty containers

which are going back of course not 50 50 but maybe 50 which comes from

far east is loaded filled with merchandise and the 50 which goes back maybe 10

of it is loaded the rest is empty that means the balance of trade between the united states and china as

far as it goes to container handling is quiet to the benefit of china we don't

send them as much as we import from them the first thing i can do here is insert and then i'll go to

charts and i may pick one of these charts these are scatter charts and these are

bubble charts so we go here then insert then we go to this one and that

will lead to something like this we click on this chart and this table comes up and we

write at trendline something like this comes up on the right hand side

this is non-linear non-linear non-linear non-linear the only linear one is this one

and we click on this the trend line is here it was inserted over there i can go here and i

can say display the equation and also display the

coefficient of determination or something that we call it r square and i will

have them on my graph click on this the equation comes up this equation tells us that y

is equal to almost 397 x minus seven eight three three eight

three you may find it a little bit strange but don't forget our excess start from 1997.

the x that we put it over there is 20 20. but usually in regression analysis we may subtract this one from 1996

and then we get 1 2 3 and here we get a 23 and if we click on this one then

the r square or coefficient of determination will appear over there

what r square says is how close the dots that you have are to a linear pattern or

close to a line if those dots those blue dots that we have if this dots

if they are exactly on the line r square comes out equal to one if they are up and down and don't show a

specific pattern then r square comes out something close to zero

what we like to see is r squares which are closer to one large larger than point seven indeed r squared

represent what percentage of changes in y can be explained in terms of changes

in x let me explain what is behind regression analysis regression analysis tries

to minimize the gap between regression line and the actual data

no matter we use intercept and slope functions of excel or we insert the line

on the graph or using data analysis tool pack what list square method tries is to minimize

the gap if this is actual data and if this is the regression line

the least square method tries to minimize the square of the gap between these two

so it tries to minimize the summation of these squares and now we want to use data analysis

toolpak of excel to find the best line a line and also the best line are identified using two parameters

one of the parameters is here this height and we call it b0 b0 is the intercept of the line with

y-axis and then if we go one unit to the right and then go up that is the slope of the

line and we show it using parameter b1 and we represent any line

as y equal to b 0 plus b 1 x constant and a constant

multiplied by a variable if we have several variables if you are talking three-dimensional four-dimensional space

then it would be y equal to b 0 plus b 1 x 1 plus b 2

x 2 plus b 3 x 3 and so on and so forth what the regression line tries

to do is to find those coefficients and give us the equation

for the line here is 23 years data of san pedro bay ports the combination of ports of los angeles and long beach

which are considered the 10th largest containerized port in the world so i have downloaded this data but one

thing i want to tell you here is unless you are in the last

stage of preparing a report and presenting it to a superior never left justify right justify

senator justify numbers never because when you download things from web sometimes you think they are

numbers but they are not they are text and if you do operations on them they will not come out correct

so let them as they are unless you really want to clean up them and present it to a superhero in the

last stage if we look at the format format here is general these numbers are

1997 to 2019 to make our job a little bit simpler i have decided to transform 1997

into one and then go down to get 23 for 2019 but if you write b2 is equal to

a2 minus 1996 and copy down you will see that it doesn't work this is what you will see not one two

three four five six seven and when you look at the format you see it is text

because text are left justified and numbers are right justified so if i apply a

function on text it doesn't come out as i expect to be applied on a number

the first thing you may do is to mark this column and change this text to number

in a small percentage of time it works in a large percentage of time it doesn't work

the next thing you may try and it works most of the time is type this formula here b2

i type equal to a2 minus 1996 and i expect that b2

to come out 1. then i mark the whole column click on data and then i click on text

to column and then i click finish and usually it works those texts become number

sometimes you need to leave it under limited sometimes you may need to change it to

fixed width sometimes you may need to push next and then finish sometimes from the beginning you can push finish

you may need to try all of these possibilities to change those texts into numbers here you may be

able to type 23 pieces of data which i can't i can type them and make sure that i

have typed them correctly but if they were 230 pieces or 2300 or 23 000

then perhaps you would have not been able to type them and therefore we need to

find a way to change those texts to numbers if you are aware of other ways

please share it and then if i do that after i click on finish then i will see one here

and then when i copy down i will have 1 to 23

type over there data text to column finish or maybe you need to change this one or

maybe you need to first click on next and then the other

one and finally and hopefully you come up with this and then you can transform this one into this one by copying it

down so finally we have our data finally we know that we are using

what technique click on data then on data analysis data data analysis

this window comes up click on regression okay this window comes up make sure

you enter y values first dependent values first these values then x values

these values i changed x values from 1997 to one and two and three instead of four digits

and therefore the equation will differ whether my x is from one or it is from

1997. slope remains the same but intercept will be different however when you enter the appropriate number

in these two different equations you will get the same results so we enter this xs

and they are there then it will ask if you want the result on a new page

or on this page i want it on this page so i'll click on output ring click over there

and as i click that button will go there and now where do you want me to put it cell e1

and then okay in this graph in this table whatever is in

black was prepared by data analysis what is in green was prepared by data analysis and i

painted them green to say these are the most important items that you need to know

and red parts are my explanations i will explain it one by one again green and black parts

are output of data analysis the table that you get as many decimal points but when you prepare a report you don't

report 10 decimal points therefore i have a little bit organized the

output of data analysis just by a little bit adjusting the columns or reducing the number of decimal points

the first two items that we look in this table are intercept and

x variable one intercept is b0 x variable 1 is b1 and the equation is stated as

y equal to b0 which in this case is 8500 433 plus

396.761 multiplied by x that is the regression line

equation we need both numbers next we look at multiple r

multiple r is what we also call it correlation coefficient and shows the strengths of the relationship between x

and y it ranges from negative one to positive one but negative one is as good as positive one both of them both

negative one and positive one tell us that there is a very tight relationship hundred percent

relationship between x and y but when it is negative the relationship is downward

when it is positive is upward however a close to negative one or a close to one

correlation coefficient indicates that the observations that we have are around the regression line they are

not far away the regression r square or coefficient of determination or what we show it in excel as

rsq that is just square of correlation coefficient and therefore coefficient of determination r square

which is the square of correlation coefficient is always between 0 and 1 because

correlation coefficient is always between negative 1 and positive 1

and this is the square of the other one so it's always between zero and one correlation coefficient is

used only for analysis of the linear relationship between two variables coefficient of

determination can be used for analysis of the relationship

linear or non-linear relationship between two or more than two variables coefficient

of determination r square has more applications but they both are targeted towards

the same conclusions that if coefficient of determination is closer to 1

or if correlation coefficient is closer to 1 or negative 1 that means we can rely on that

relationship no matter if that relationship is upward or downward so here

coefficient of determination r squared is 0.77 and correlation coefficient or r

or what we call it here multiple r is always sine of b1 here is positive multiplied by

square root of r square and the square root of r squared here is point

eight eight correlation coefficient is always square root of coefficient of determination and then

we need to put sine of b in front of it if the relationship is downward we need to put a negative in front of it

if it is upward the positive is already p-value we want p-value small we want r squared large

but p-value small definitely less than that is what we call it 95 confidence level

i need to explain one other thing here e minus zero eight and then a decimal point is

in front of it three point twenty seven e minus zero eight

e zero eight means 1 here and then 8 0's in front of it

e negative 0 8 is equal to 1 divided by 1 and 8

0's in front of it so it is decimal point seven zeros and a one over there which

is quite small compared to say 0.1 or 0.01 or 0.0

25 or compared to 0.05 all of these numbers are much bigger than that

and we like small values small value means we cannot deny a relationship between x and y

or a relationship between y and several x's suppose we have been asked to forecast

for next period then we need one other piece of information that is

the number of observations that we have had up to now which is here 23 observations but we also want

to know whether we have started from one or from something greater than one therefore if in this problem

i tell you suppose the first period is period one what is your forecast for the next

period period one we have had 23 periods therefore next period is period

24. 24 must be multiplied by this one multiplication and then added to this one and that is my forecast

for the next period so linear relationship between two variables is stated like this this

is x y intercept slope and so on and so forth r and b 1 have the

same sign r and b 1 are positive r and b 1 are negative and this is the situation

when r and p one are zero or close to zero correlation coefficient or or what is in

output known as multiple r is something between negative one and one we have already gone through

this discussions that plus one and negative one have the same meaning

they both say that the observations are close to the regression line from the regression line

correlation coefficient is used for relationship linear relationship between two

variables coefficient of determination for linear and non-linear relationship between

two or more variables we can also say that coefficient of determination which we also

call it as r squared or r sq it is number between zero and one when we multiply it by hundred

it is between zero and hundred percent it will tell us this percent of changes

in y can be explained based on changes in x so if r squared is equal to 1 that means

if you give me x i could give you y value with hundred percent precision

hundred percent of behavior of y is explained in terms of x if it is zero

that means you give me x because no matter what data you enter into data analysis or other

functions you get some line you get some equation but if r squared is small

that or p value is high that means if you give me a x i will give you y but i don't know if that y is a real y

because you give me a small x i will give you this y and you give me very large x and i give

you again a y that is close to the one for a small value

so when r squared goes down our trust in the regression line also goes down our trust in the

forecast of the regression line we have already discussed this you may just read it quickly we have

also discussed this one one other piece of information that is

important in the output of data analysis is standard error this is

also a very important outcome standard error is like 1.25 math or like square root of

mse remember we were using 1.25 mad and the square root of mse as the standard deviation of our

forecast for next period here standard error has the same meaning therefore if you ask me

to provide you with average and standard deviation of my forecast for next period that is what i will do

i will go to the number of observations and that is 23 and if observation started from period 1

i know next period is period 24. so i have 24 here and i know that the slope is this one

so i multiply it and i add it to it and that is what i have done here so this would

be my forecast for next period but that is a shallow statement then this

is my forecast for next period we always say that you are free to say that but if we think

in depth that is the average of our forecast for next period

but that average also has a standard deviation and this is the standard deviation my

forecast for next period has an average of 1874

and its standard deviation is 1491 or 1492. we said that we never have a precise

forecasting technique therefore always our forecast must be accompanied by a

standard deviation of the forecast but there is a little bit difference between standard error

and square root of mse both of them are based on square of the error square of the deviations

but there is a small difference i will expect look at this

table this is my actual data and based on the computations that i had in previous page and i found

p0 and p1 these are my forecasts and now what i have done

i have found a minus f and i have squared it and i have put it

over there exactly the same thing which you are doing for mse in moving average and exponential smoothing

a minus f square a minus f squared so these are all a minus

f squared values and then in mse we have had averaged an

average is equal to sum [Music] of a minus f

squared divided by the number of observations and that is what we were doing in moving

average and exponential smoothing and we were calling it mse and then the square root of msc was

standard deviation of all forecasts for the next period standard error has a small difference

this is summation a minus f

squared this is standard error if i divide this summation by n

it doesn't come out equal to standard error if i divide it by n minus 1

it doesn't come out if i divide it by n minus 2 it is exactly the same value so summation of the

square of the deviation square of the errors divided by n minus 2 it is standard

error and we use it as the standard deviation of our forecast for the next period

why should we divide it by n minus 2 whenever you find the difference of two things

and you have found some parameters of this thing based on these things for each parameter

you have estimated you need to deduct one degrees of freedom we use the actual data to estimate b0

and b1 two parameters therefore if i am dividing the gaps i cannot divide it by n

because i lose 2 degrees of freedom and therefore i must divide it by n minus 2.

you don't need to be worried about that because the output prepares for you but i just want you to know what

is the standard error and what is mse and why they are a little bit different mse

in the concept that we used it before otherwise correct msc is this one because we have

estimated two parameters and we lose two degrees of freedom

remember when you manually were using in some of your other courses because in this course we don't do those

things manually if you remember to compute a standard deviation to compute variance

you should have divided it by n minus 1 not by n that is because you estimate variance based on your

estimate of average and since you have estimated the average you lose

1 degrees of freedom and that is why in computing variance of the sample instead of

dividing it by the number of observations we divide it by

number of observations minus 1. and this is another example that you may go through

i hope we have had a good review we will go through more in-depth concepts

in future discussions thank you very much for attending this session

i hope we can have a good learning environment together English (auto-generated)