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

Lecture – 46
Multiple Regression Analysis in SPSS - II

Welcome friends to the class of Marketing Research and Analysis, in the last lecture ah we have started with multiple regression, so we understand now that multiple regression is helpful for prediction or predicting or some outcome, where ah the ah outcome is dependent on certain independent variables also called as the predictors. So, the effect of the independent variables or the predictors on the dependent variable is for interest to always measure in any research study.

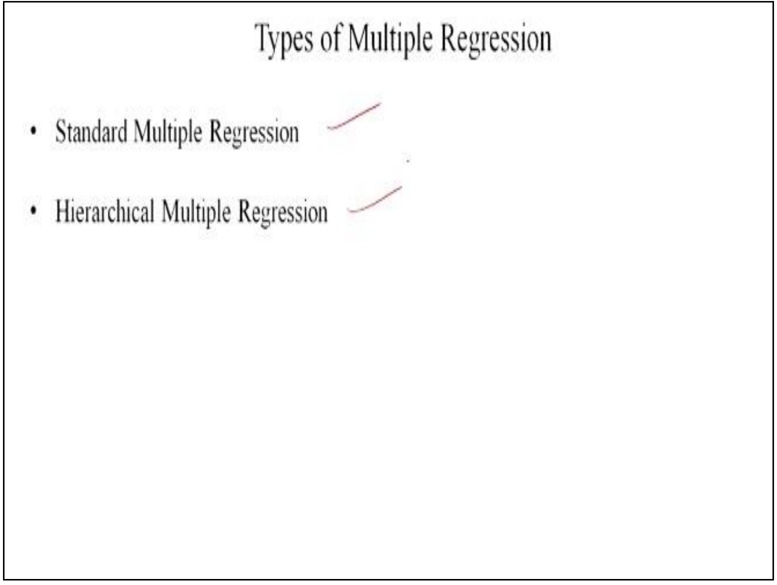
That means what I means to say is there are suppose 3 variables or 2 variables and we want to see the effect of these 2 variables on the dependent variable and we want to see which one of them x 1 and x 2 suppose are the independent variables then whether x 1 has the larger impact on the dependant variable y or x 2 has the larger impact on the dependent variable y.

So these are things that is of the interest and then we also would like to know how much x 1 and x 2 are explaining the dependent variable and how much is left as a unexplained variants which is called as the error or residual well ok. So, there while discussing about multiple regression I remember we had discussed about the concept called multicollinearity, so we said the presence of multicollinearity which is generally very generally ah not we know ah which comes from by understanding that if there is a high correlation among the independent variables generally about .9 also generally speaking.

Then problem of multi collide exist but there are also instances when the combined effect of certain variables does have an effect of multicollinearity so in that case we can see that even a poor correlation that means not a high correlation as high as .9 but even a correlation . 6 or .7 can also have a multicollinearity issue, so to do that I explained how to check that how to measure that multicollinearity problem through a techniques like ah variance inflation factor call VIF or tolerance, then how to calculate that also I explained in the last lecture.

So today will be continue from there and trying to solve the few problems and also doing it in the SPSS, ah this is a concept of multiple regression.

(Refer Slide Time: 02:50)



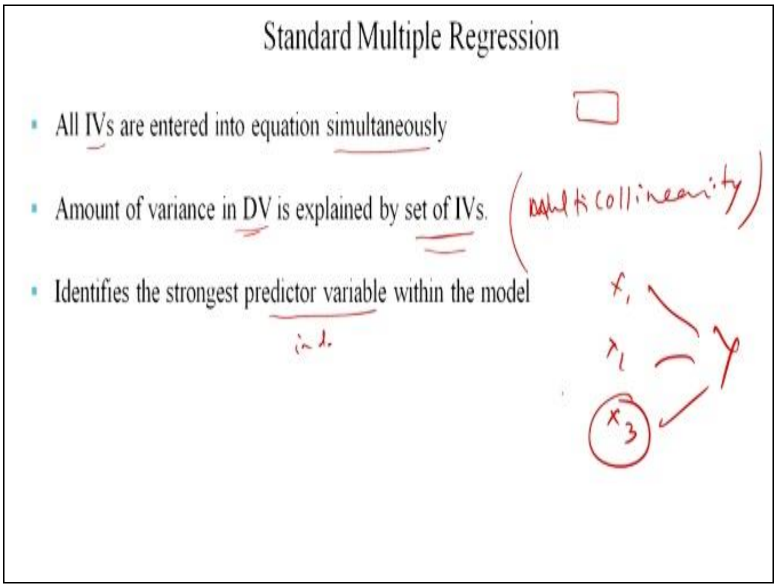
Types of Multiple Regression

- Standard Multiple Regression
- Hierarchical Multiple Regression

So, the 2 types of multiple regression basically we talk off one is Standard Multiple Regression and other is the Hierarchical Multiple Regression. What is the Standard Multiple Regression? And what is the Hierarchical Multiple Regression? The Standard Multiple Regression is a type of regression ah which assumes that all the variables are equal important there is nothing to be control.

But in the Hierarchical Multiple Regression on the other hand they is we are trying to create blocks and trying to control certain variables and trying to see then the ah effect on the dependent variable, that means when I am controlling a variable and using others suppose there 3 variables suppose I am controlling x 1 and leaving x 2 and x 3, then what is the effect on the dependent variable and once I take the all 3 variables together what is the effect on the dependent variable. So, when I see such effects to that help me to understand the effect of individually the variables so let us started with Standard Multiple Regression ok.

(Refer Slide Time: 03:55)



Standard Multiple Regression

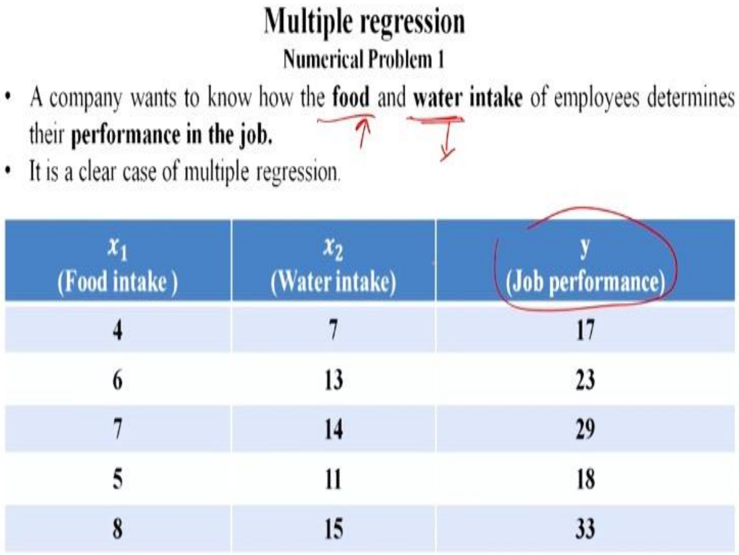
- All IVs are entered into equation simultaneously
- Amount of variance in DV is explained by set of IVs. (multicollinearity)
- Identifies the strongest predictor variable within the model

ind.

So here all the independent variables are entered into the equation simultaneously so as I said like in the Hierarchal regression you do in the block form so there is nothing like it here, so all IVs are entered simultaneously. The amount of variance in the dependent variable is explained by the set of independent variables but I had explained you and I had warned you that you should be careful of the problem of multicollinearity problem of multicollinearity, you should always check it.

If you do not taken there is a presence of multicollinearity what will happen may be you will have very insignificant beta weights and you can have you know hypothesis that some independent variable is influencing dependent variable might come wrong, so this is what will ah happen then it helps in identifying the strongest predictor variable or the strongest independent variable within the model suppose we have X 1, X 2, X 3 which have an effect on the y now out of these which one has the largest impact that can be measured with the help of this Standard Multiple Regression.

(Refer Slide Time: 05:00)



Multiple regression

Numerical Problem 1

- A company wants to know how the **food** and **water intake** of employees determines their **performance in the job.**
- It is a clear case of multiple regression.

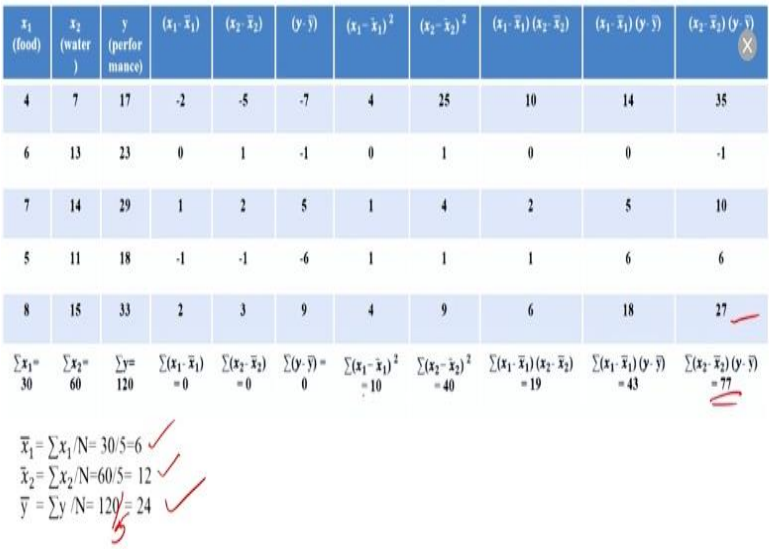
x_1 (Food intake)	x_2 (Water intake)	y (Job performance)
4	7	17
6	13	23
7	14	29
5	11	18
8	15	33

Let us take this case, will solve this problem, this is numerical problem will solved by hand. A company wants to know how the food and water intake of employees determine the performance in their job. So the company is bothered or it is interested to know how how is the performance in the job connected with the intake of food and water, that means does some time we should we also know that after your lunch your efficiency goes down, you tend to your body also starts to using this energy for digestion and all.

So you become lethargy and tend to sleep and something but if your water intake is low on the other hand inverse relationship then you might be dehydrated and your energy your body will start trying to save the energy so in that case water intake should be high and also not very high that frequently visit the wash room may be other loo and food intake should not be very high on the other hand,

So it should not this one should not be very high and this one should not be very low to dehydrate you, so we want to see how this affects the job performance, and let us see this case ok.

(**Refer Slide Time: 06:18)**



x_1 (food)	x_2 (water)	y (performance)	$(x_1 - \bar{x}_1)$	$(x_2 - \bar{x}_2)$	$(y - \bar{y})$	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)^2$	$(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$	$(x_1 - \bar{x}_1)(y - \bar{y})$	$(x_2 - \bar{x}_2)(y - \bar{y})$
4	7	17	-2	-5	-7	4	25	10	14	35
6	13	23	0	1	-1	0	1	0	0	-1
7	14	29	1	2	5	1	4	2	5	10
5	11	18	-1	-1	-6	1	1	1	6	6
8	15	33	2	3	9	4	9	6	18	27
$\sum x_1 =$ 30	$\sum x_2 =$ 60	$\sum y =$ 120	$\sum(x_1 - \bar{x}_1) = 0$	$\sum(x_2 - \bar{x}_2) = 0$	$\sum(y - \bar{y}) =$ 0	$\sum(x_1 - \bar{x}_1)^2 = 10$	$\sum(x_2 - \bar{x}_2)^2 = 40$	$\sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) = 19$	$\sum(x_1 - \bar{x}_1)(y - \bar{y}) = 43$	$\sum(x_2 - \bar{x}_2)(y - \bar{y}) = 77$

$\bar{x}_1 = \sum x_1 / N = 30/5 = 6$ ✓

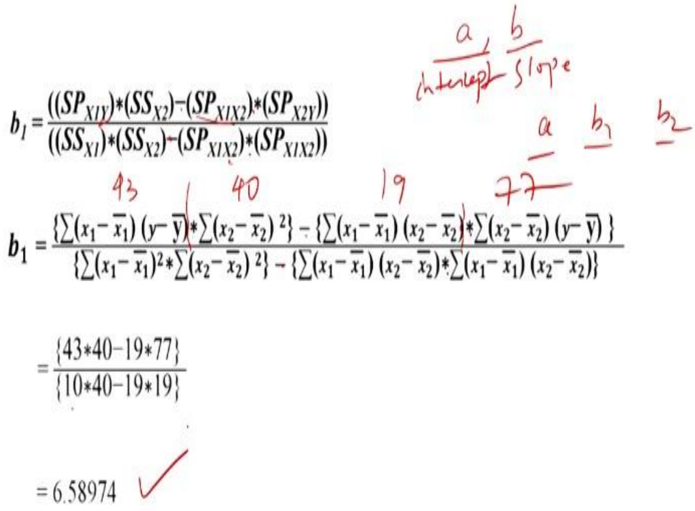
$\bar{x}_2 = \sum x_2 / N = 60/5 = 12$ ✓

$\bar{y} = \sum y / N = 120 = 24$ ✓

5

So how do you do this so I have x 1 food, x 2 is water, performance is my y. So this are all the value that you see, so summation of the x 1 is 30, summation of x 2 is 60, summation of y is 120 now x 1 – x 1 bar, so if I show you the formula first.

(Refer Slide Time: 06:40)



a, b

intercept slope

a b1 b2

$$b_1 = \frac{((SP_{X1Y})*(SS_{X2}) - (SP_{X1X2})*(SP_{X2Y}))}{((SS_{X1})*(SS_{X2}) - (SP_{X1X2})*(SP_{X1X2}))}$$

43 40 19 77

$$b_1 = \frac{\{\sum(x_1 - \bar{x}_1)(y - \bar{y}) * \sum(x_2 - \bar{x}_2)^2\} - \{\sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) * \sum(x_2 - \bar{x}_2)(y - \bar{y})\}}{\{\sum(x_1 - \bar{x}_1)^2 * \sum(x_2 - \bar{x}_2)^2\} - \{\sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) * \sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)\}}$$

$$= \frac{\{43*40 - 19*77\}}{\{10*40 - 19*19\}}$$

$$= 6.58974$$ ✓

Why do I do it, so this is the formula how do I measure so you remember that when any regression is simple regression we had this thinks to which we needed to calculate a which is my intercept then I needed b slope. So I needed this slope and intercept in the simple regression intercept but during the multiple regression suppose there are 2 variables then I will have a, b 1 and b 2, so now I need to calculate this b 1 and b 2.

So how which is b 1 calculated it says sum of products of x 1 into y so sum of SS is Sum of square of x 2 - sum of product of x 1 and x 2 so here it was x and y here x 1 and x 2 into sum of product of x 2 and y divided by over all sum of square of x 1 into sum of square x 2 – sum of product of x 1, x 2 into sum of product of x 1, x 2.

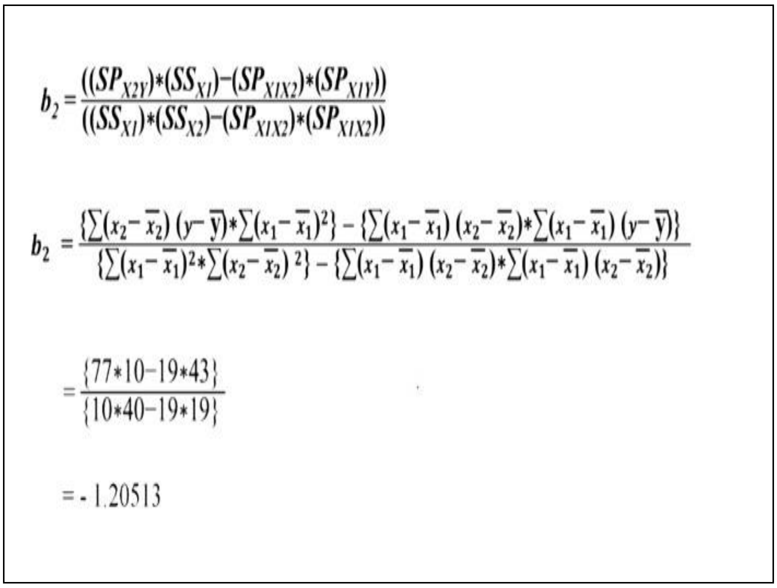
So where I, when we do this will require the values for some example you see sum of product of x and y, so x 1 into y here we are taking x 1 as x 1 – x 1 bar so this is the mean, so the observe value - the mean the summation of it, so for y we are taking y – y bar into sum of square of what x 2 so x 2 – x 2 bar mean so square because this only summation square of x 2 – so this is – sum of product of x 1, x 2 so sum of x 1 first x 1 is x 1-x 1 bar * x 2 – x 2 bar, so this one is this part up to this much then into which comes here is sum of product of x 2, y.

So x 2 is x 2 – x 2 bar into y is y – y bar divided by similarly sum of square of x 1 what it will be x 1 is x 1 – x 1 bar since it is the square sum of square it is square into sum of square of x 2 so x 2 – x 2 bar square, now this minus come here sum of product of x 1, x 2 so x 1 will be x 1 – x 1 bar into x 2 will be x 2 - x 2 bar into this into x 1, x 2 so x 1- x 1 bar into x 2 – x 2 bar so when I do this for that when I calculate now let me I go back.

So x 1 – x 1 bar x 2 – x 2 bar y - y bar x 1 – x 1 bar square of it x 2 - x 2 bar square now x 1 – x1 bar into x 2 – x 2 bar, x 1 – x 1 bar into y - y bar x 2 - x 2 bar y - y bar these are all we require so when we get this we have found out all the values so now x 1 summation of x is 30 the mean of x 1 is how much 6, x 2 is summation of x 2 60 so divided by 5 N is 5 so 12 summation of mean of y is summation of y divided by N so how much 120. 120 divided by 5, I have forgotten to write it, so N is 5 so that is equal to 24 ok.

So all the 3 we have got now let us use this when I use this formula put it in the formula how much am I getting x 1 – x 1 bar let us go back, so how much is it, so x 1- x 1 bar into y – y bar you required this value so this part, this part is how much 43 then x 2 - x 2 bar square this one, so this is 40 then x 1 – x 1 bar into x 2 – x 2 bar this one is 19 and x 2 - x 2 bar into y - y bar this one how much 77, so 77 then putting similarly the at the below x 1 - x 1 bar square this one 10, so 10 into 40 again we have seen at here from here x 2 – x 2 bar is 40 and 19th we have already got it here so now taking this 19 into 19 so because is the 2, so b 1 we are getting 6.58974.

(Refer Slide Time: 11:31)



$$b_2 = \frac{((SP_{X2Y})*(SS_{X1})-(SP_{X1X2})*(SP_{X1Y}))}{((SS_{X1})*(SS_{X2})-(SP_{X1X2})*(SP_{X1X2}))}$$

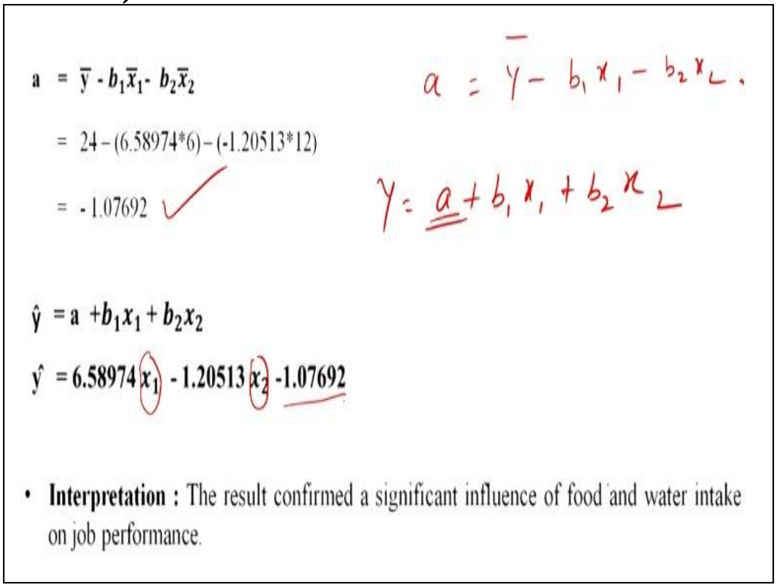
$$b_2 = \frac{\{\sum(x_2-\bar{x}_2)\,(y-\bar{y})*\sum(x_1-\bar{x}_1)^2\} - \{\sum(x_1-\bar{x}_1)\,(x_2-\bar{x}_2)*\sum(x_1-\bar{x}_1)\,(y-\bar{y})\}}{\{\sum(x_1-\bar{x}_1)^2*\sum(x_2-\bar{x}_2)^2\} - \{\sum(x_1-\bar{x}_1)\,(x_2-\bar{x}_2)*\sum(x_1-\bar{x}_1)\,(x_2-\bar{x}_2)\}}$$

$$= \frac{\{77*10-19*43\}}{\{10*40-19*19\}}$$

$$= -\ 1.20513$$

So we are calculated b 1 similarly you can calculate for b 2, now just see the difference between the formulas, here X 1 Y here X 2 Y here X1 here X 2 this is same now this one here X 2 Y here it will be X 1 Y and other things also slightly if you use your little slowly or logic you can understand the changes and you can make the formula, so SS X 1 into SS X 2, SP sum of product X 1 X 2 twice. So now this value we have got is -1.20513 so now we have got b 1 we have got b 2.

(Refer Slide Time: 12:08)



$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

$$= 24-(6.58974*6)-(-1.20513*12)$$

$$= -1.07692 \checkmark$$

$$a = \bar{y} - b_1 x_1 - b_2 x_2 .$$

$$y = \underline{a} + b_1 x_1 + b_2 x_2$$

$$\hat{y} = a + b_1x_1 + b_2x_2$$

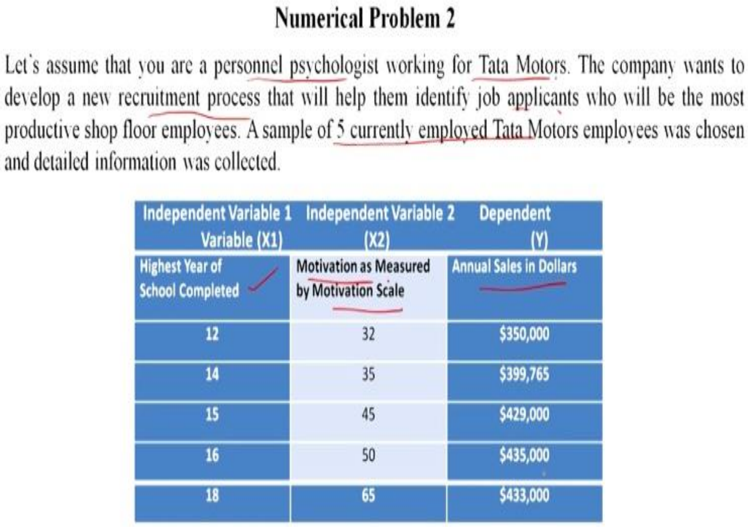
$$\hat{y} = 6.58974\,x_1 - 1.20513\,x_2 \underline{-1.07692}$$

- **Interpretation :** The result confirmed a significant influence of food and water intake on job performance.

Now a we know is equal to how much a = y bar - b 1 x 1 - b 2 x 2 now this how it is come from here y = a + b 1 x 1 + b 2 x 2 correct, so if I take this side this is what I am getting a = y – b 1 x 1 – b 2 x 2 so doing this we have see y bar was how much y bar was 24, so putting this formula we have got a is equal to this much so now what is our estimated y so estimated y is y is equal to now putting this a is equal to how much a is equal to this much – 1.07692 so + b 1 x 1 so b 1 is how much b 1 is 6.58974 x 1 – obviously we will get 1.2 – so – x 2. So now if you get a value of x 1 and x 2 then you can easily calculate for y, so interpretation the result conformed a significant influence of food and water intake on job performance.

I will show you this I have brought you know data share with me but let you show you one more way of conducting multiple regression then I will show you that, so this is a very simple method we just need to understand it you can do the next one more problem I have brought.

(Refer Slide Time: 13:44)



Numerical Problem 2

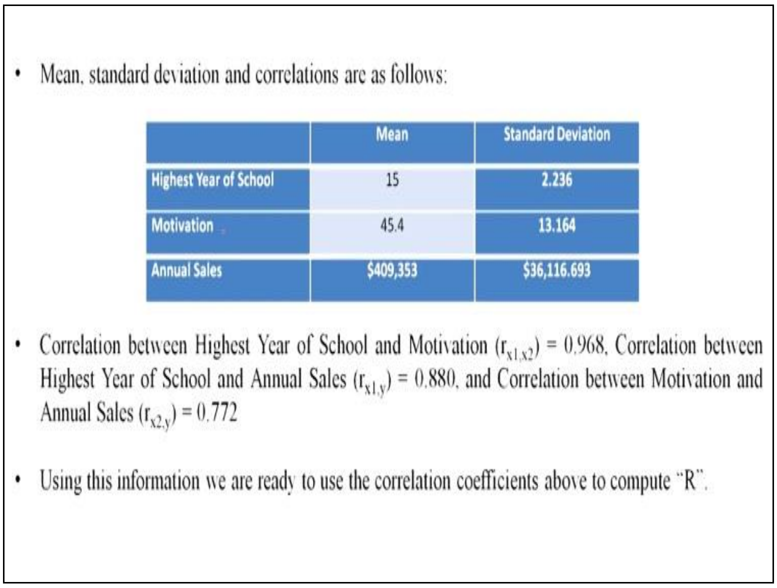
Let's assume that you are a personnel psychologist working for Tata Motors. The company wants to develop a new recruitment process that will help them identify job applicants who will be the most productive shop floor employees. A sample of 5 currently employed Tata Motors employees was chosen and detailed information was collected.

Independent Variable 1 Variable (X1)	Independent Variable 2 (X2)	Dependent (Y)
Highest Year of School Completed ✓	Motivation as Measured by Motivation Scale	Annual Sales in Dollars
12	32	$350,000
14	35	$399,765
15	45	$429,000
16	50	$435,000
18	65	$433,000

So this is what I say this a personnel psychologist who is working for Tata motors, the company wants to develop a recruitment process that will help them identify job applicants who will be the most productive employees 5 currently employed employees was chosen and detailed information was collected what information theirs school, how much highest year of school completed what is the motivation level and motivation scale it has been measured.

And how much there earning in dollar that has been measured, not earning this will be how much they are helping in the sales.

(Refer Slide Time: 14:22)



- Mean, standard deviation and correlations are as follows:

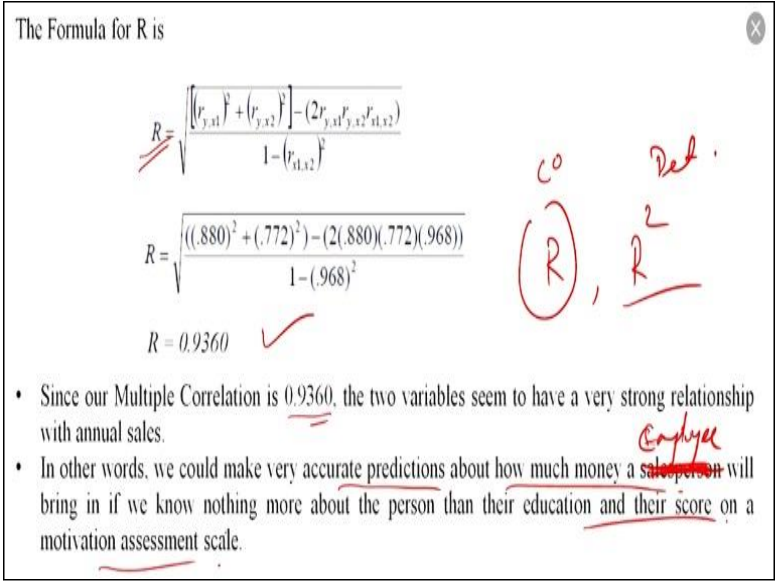
	Mean	Standard Deviation
Highest Year of School	15	2.236
Motivation	45.4	13.164
Annual Sales	$409,353	$36,116.693

- Correlation between Highest Year of School and Motivation ($r_{x1,x2}$) = 0.968, Correlation between Highest Year of School and Annual Sales ($r_{x1,y}$) = 0.880, and Correlation between Motivation and Annual Sales ($r_{x2,y}$) = 0.772
- Using this information we are ready to use the correlation coefficients above to compute "R".

So, the mean standard deviation and correlations are given, so highest year of school the Mean is 15, Motivation is 45.4 the Mean for Motivation and the Sales Mean is this much. The standard deviation also given to you, now the correlation between highest year of school and motivation are x 1, x 2. Let us see R x 1 by x 2 school and motivation is this much. Correlation between a year of School and Sales is this much.

And third Correlation between Motivation and Annual Sales is this much, so this I think I am sure you can calculate if you have followed my last to last lecture where we have, where I showed how to calculate by hand the correlation value between 2 variables. So, you can calculate now using this values we will see.

(Refer Slide Time: 15:06)



The Formula for R is

$$R = \sqrt{\frac{\left[(r_{y,x1})^2 + (r_{y,x2})^2\right] - (2r_{y,x1}r_{y,x2}r_{x1,x2})}{1-(r_{x1,x2})^2}}$$

$$R = \sqrt{\frac{((.880)^2 + (.772)^2) - (2(.880)(.772)(.968))}{1-(.968)^2}}$$

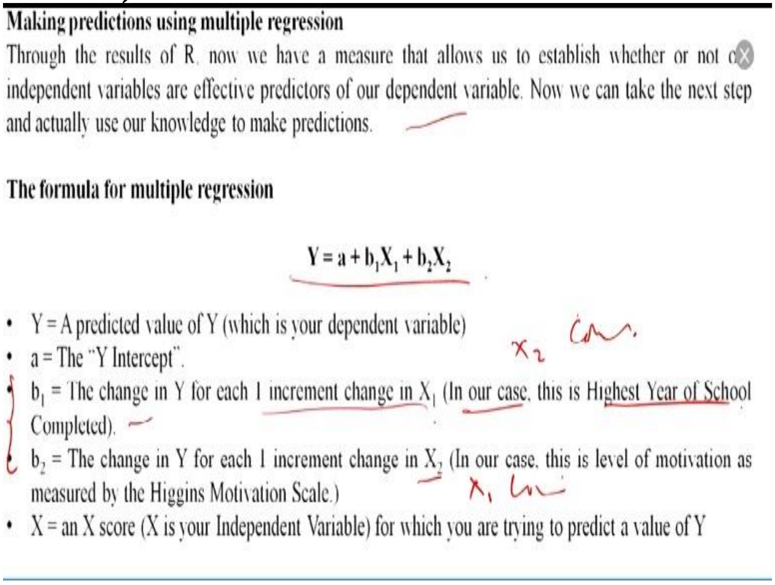
$$R = 0.9360$$

- Since our Multiple Correlation is 0.9360, the two variables seem to have a very strong relationship with annual sales.
- In other words, we could make very accurate predictions about how much money a ~~salesperson~~ Employee will bring in if we know nothing more about the person than their education and their score on a motivation assessment scale.

So, what is the formula for R? what is the R the multiple correlation now what is the formula the correlation between y and x 1 square + the correlation between y and x 2 square – 2 the correlation between y and x 1 into the correlation between y and x 2 and the correlation between x 1 and x 2 divided by 1 – the square of correlation between x 1 and x 2 square, if I do this I am getting the R value has 0.9360, since our Multiple Correlation is 0.9360 so you can understand if R is this much what will be the R square you can measure which is my this is my relation correlation, this is my determination.

So, my R square will be 9, 9's are 81 something around 384 correct, in other words we could make a very accurate prediction about how much money not a sales person sorry this is the employee will bring in if we know nothing more about the person than education and their score on a motivation scale, so if I know the motivation level and the persons school or year of education then I can say how much annual sales they can bring into the company the Tata Motors, So what is says making prediction

(Refer Slide Time: 16:36)



Making predictions using multiple regression

Through the results of R, now we have a measure that allows us to establish whether or not o independent variables are effective predictors of our dependent variable. Now we can take the next step and actually use our knowledge to make predictions.

The formula for multiple regression

$$Y = a + b_1X_1 + b_2X_2$$

- Y = A predicted value of Y (which is your dependent variable)
- a = The "Y Intercept".
- b_1 = The change in Y for each 1 increment change in X_1 (In our case, this is Highest Year of School Completed).
- b_2 = The change in Y for each 1 increment change in X_2 (In our case, this is level of motivation as measured by the Higgins Motivation Scale.)
- X = an X score (X is your Independent Variable) for which you are trying to predict a value of Y

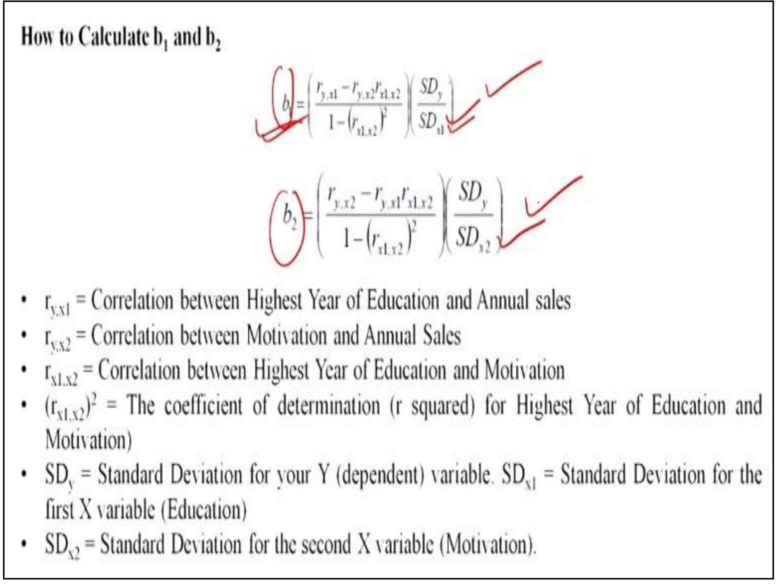
x_2 con.

x_1 con

Using Multiple Regression through the results of R, we have measure that allows us to establish whether or not our independent variables are effective predictors of our dependent variable. Now we can take the next step and actually use our knowledge to make predictions. So Y is equal to this much this is the formula so prediction, so Y is equal to predicted value of Y, a is the Intercept b 1 the change in Y for each 1 increment change in X 1.

So, in our case, this is Highest Year of School, school was the X 1, b 2 is the change in Y this is very very important unit to reread it and understand, this is very simple and very important, the change in Y for each 1 increment change in X 2 that means this case the X 2 was kept constant, X 2 was constant in this case the X 1 is constant this the meaning X is equal to an X score for which you are trying to predict a value of Y.

(Refer Slide Time: 17:34)



How to Calculate b_1 and b_2

$$b_1 = \left(\frac{r_{y.x1} - r_{y.x2}r_{x1.x2}}{1-(r_{x1.x2})^2}\right)\left(\frac{SD_y}{SD_{x1}}\right)$$

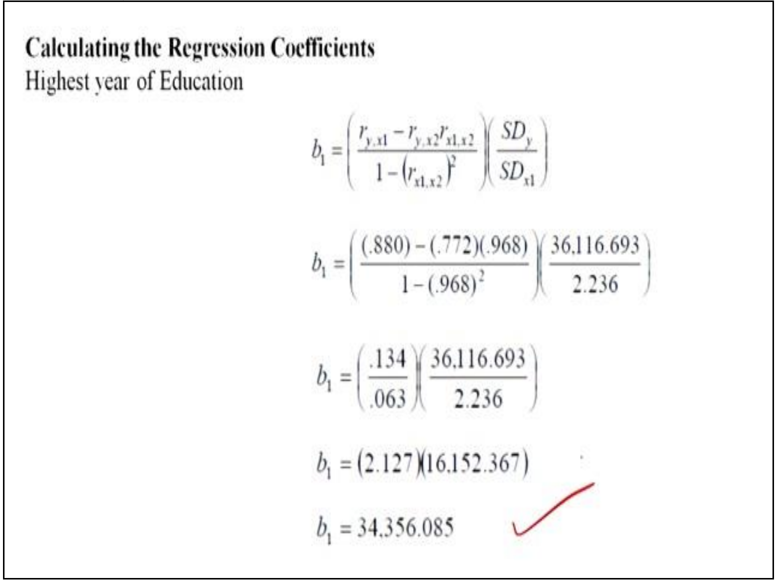
$$b_2 = \left(\frac{r_{y.x2} - r_{y.x1}r_{x1.x2}}{1-(r_{x1.x2})^2}\right)\left(\frac{SD_y}{SD_{x2}}\right)$$

- $r_{y.x1}$ = Correlation between Highest Year of Education and Annual sales
- $r_{y.x2}$ = Correlation between Motivation and Annual Sales
- $r_{x1.x2}$ = Correlation between Highest Year of Education and Motivation
- $(r_{x1.x2})^2$ = The coefficient of determination (r squared) for Highest Year of Education and Motivation)
- SD_y = Standard Deviation for your Y (dependent) variable. SD_{x1} = Standard Deviation for the first X variable (Education)
- SD_{x2} = Standard Deviation for the second X variable (Motivation).

Now how to calculate this b 1, b 2 this formula now we have now b 1 is equal to this is the formula if you use the standard deviation all the notations are given so r y x 1 is the Correlation between Highest Year of Education and Annual sales why because y is my sales, x 1 my education – r relation between y sales and x 2 is my motivation into education and mmotivation t is into divided by this 1 – education and motivation square into the standard deviation of y that is sales divided by standard deviation of x 1.

Why because we are measuring for b 1 so this is we measuring for x 1, so measuring b 1 we will be using x 1, for b 2 will be using x 2. So when we use this formula this another way we also you can do it through the correlation.

(Refer Slide Time: 18:28)



Calculating the Regression Coefficients

Highest year of Education

$$b_1 = \left(\frac{r_{y.x1} - r_{y.x2}r_{x1.x2}}{1-(r_{x1.x2})^2}\right)\left(\frac{SD_y}{SD_{x1}}\right)$$

$$b_1 = \left(\frac{(.880)-(.772)(.968)}{1-(.968)^2}\right)\left(\frac{36,116.693}{2.236}\right)$$

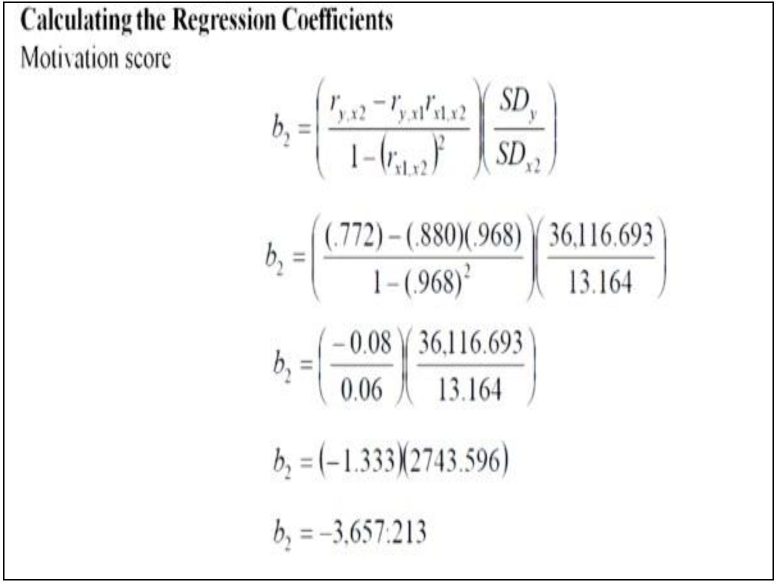
$$b_1 = \left(\frac{.134}{.063}\right)\left(\frac{36,116.693}{2.236}\right)$$

$$b_1 = (2.127)(16,152.367)$$

$$b_1 = 34,356.085$$

So, if you use this so how much is the b1 in this case we have got so b1 by putting in this formula we have got 34356.

(Refer Slide Time: 18:39)



Calculating the Regression Coefficients

Motivation score

$$b_2 = \left(\frac{r_{y.x2} - r_{y.x1}r_{x1.x2}}{1-(r_{x1.x2})^2}\right)\left(\frac{SD_y}{SD_{x2}}\right)$$

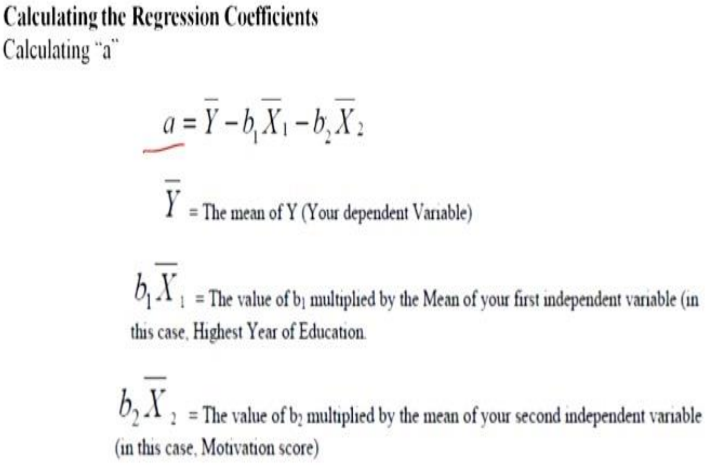
$$b_2 = \left(\frac{(.772)-(.880)(.968)}{1-(.968)^2}\right)\left(\frac{36{,}116.693}{13.164}\right)$$

$$b_2 = \left(\frac{-0.08}{0.06}\right)\left(\frac{36{,}116.693}{13.164}\right)$$

$$b_2 = (-1.333)(2743.596)$$

$$b_2 = -3{,}657{:}213$$

(Refer Slide Time: 18:44)



Calculating the Regression Coefficients

Calculating "a"

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

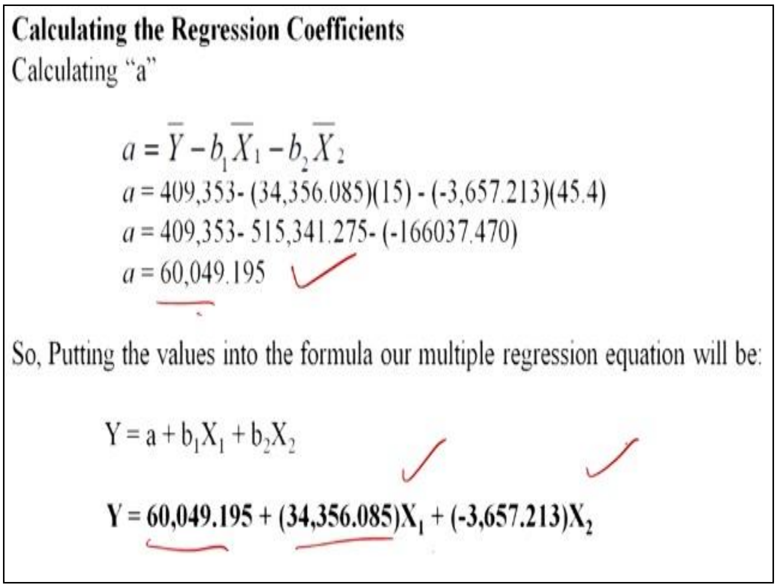
$\overline{Y}$ = The mean of Y (Your dependent Variable)

$b_1\overline{X}_1$ = The value of b_1 multiplied by the Mean of your first independent variable (in this case, Highest Year of Education.

$b_2\overline{X}_2$ = The value of b_2 multiplied by the mean of your second independent variable (in this case, Motivation score)

Similarly b 2 is – 3657. So if I use and now calculate a by using the same formula a = y – b 1 X 1-b 2 X 2 since X 1, X 2 only means are known to use them, so all the 3 means.

(Refer Slide Time: 19:02)



Calculating the Regression Coefficients

Calculating “a”

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

$$a = 409{,}353 - (34{,}356.085)(15) - (-3{,}657.213)(45.4)$$

$$a = 409{,}353 - 515{,}341.275 - (-166037.470)$$

$$a = 60{,}049.195$$

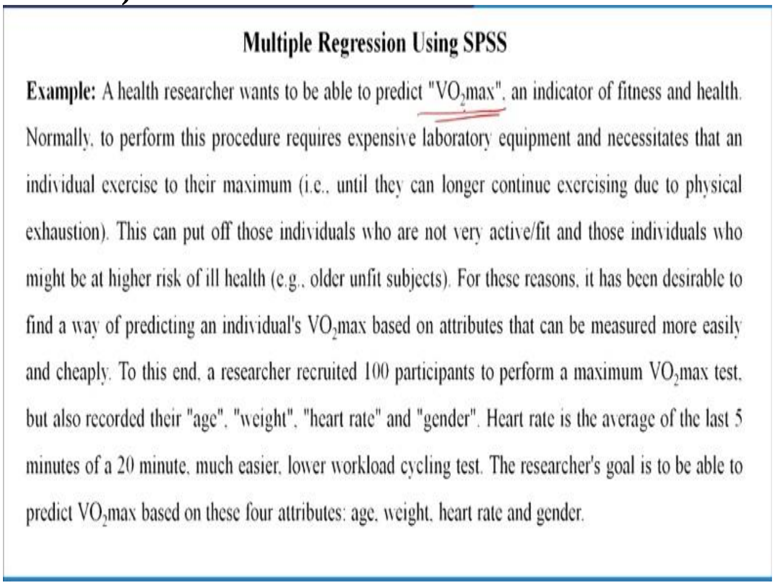
So, Putting the values into the formula our multiple regression equation will be:

$$Y = a + b_1X_1 + b_2X_2$$

$$\mathbf{Y = 60{,}049.195 + (34{,}356.085)X_1 + (-3{,}657.213)X_2}$$

So, then we have by using this way calculated a is 60,049, so now if I put the values into the formula how it will look a + b 1 X 1 + this is the – values so – b 2 that means b 2 value whatever it is into X 2, now whatever now X 1 and X 2 whatever the education level and the motivation level are there if you know about the new employee then you can predict then how much of sales will happen due to this particular employee, so this is what Multiple Regression helps you to do. So, we are done this by and now I explain how to do it in you know the SPSS. Let us understand this case,

(Refer Slide Time: 19:53)



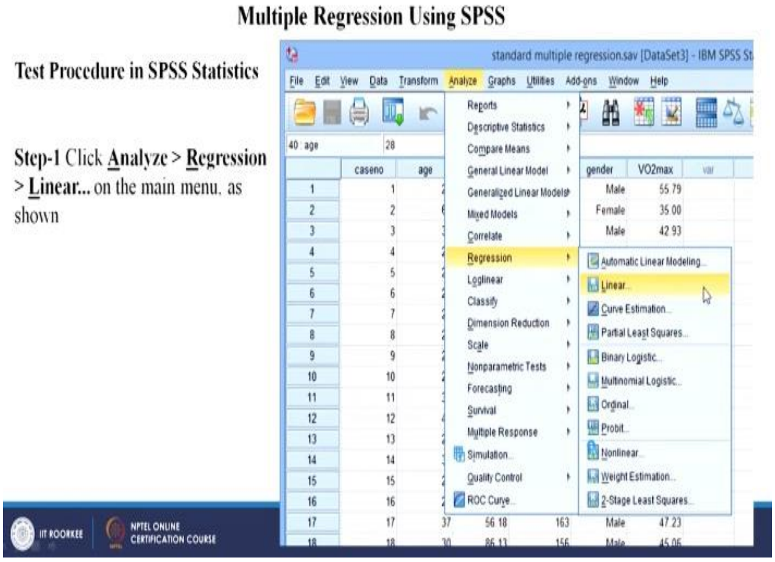
Multiple Regression Using SPSS

Example: A health researcher wants to be able to predict "VO_2max", an indicator of fitness and health. Normally, to perform this procedure requires expensive laboratory equipment and necessitates that an individual exercise to their maximum (i.e., until they can longer continue exercising due to physical exhaustion). This can put off those individuals who are not very active/fit and those individuals who might be at higher risk of ill health (e.g., older unfit subjects). For these reasons, it has been desirable to find a way of predicting an individual's VO_2max based on attributes that can be measured more easily and cheaply. To this end, a researcher recruited 100 participants to perform a maximum VO_2max test, but also recorded their "age", "weight", "heart rate" and "gender". Heart rate is the average of the last 5 minutes of a 20 minute, much easier, lower workload cycling test. The researcher's goal is to be able to predict VO_2max based on these four attributes: age, weight, heart rate and gender.

A health researcher wants to be able to predict VO 2max now this earlier also we had use this case VO 2 is basically utilizing the oxygen or ability of body to utilize oxygen during intense situation. Which is the indicator of fitness and health, so normally to performance this procedure expensive laboratory equipments are required, so this is the case where 100 participants have been taken to perform a maximum VO 2max test their age, weight, heart rate and gender was recorded to see if gender is influence the VO 2 utilization, if heart rate affected it, if weight affected it, if age affected it.

Heart rate is the average have the last 5 minutes of a 20 minute much easier lower workload cycling test, heart rate has been measured that way, the researchers goal is to predict maximum oxygen utility based on these 4 attributes, now let us see this case.

(Refer Slide Time: 20:54)



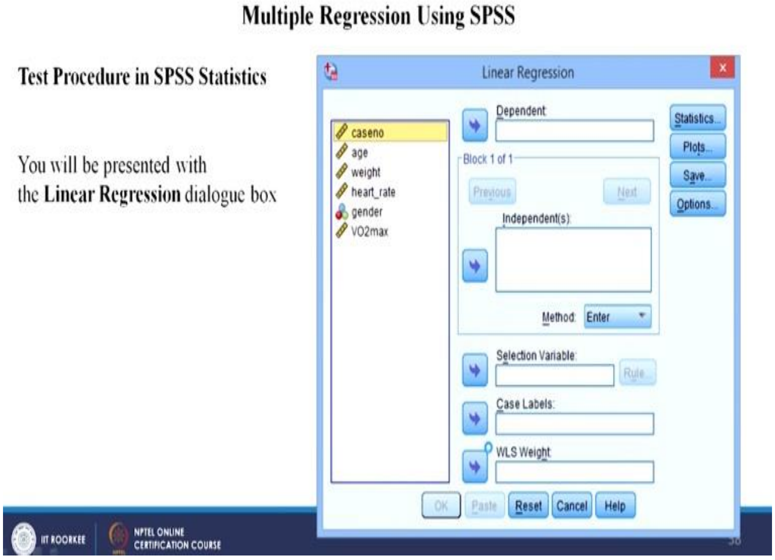
Multiple Regression Using SPSS

Test Procedure in SPSS Statistics

Step-1 Click **Analyze** > **Regression** > **Linear...** on the main menu, as shown

Then now how do you do it, now I am going straight away to the multiple regression case. So how you should do it, I will show it through another example,

(Refer Slide Time: 21:04)



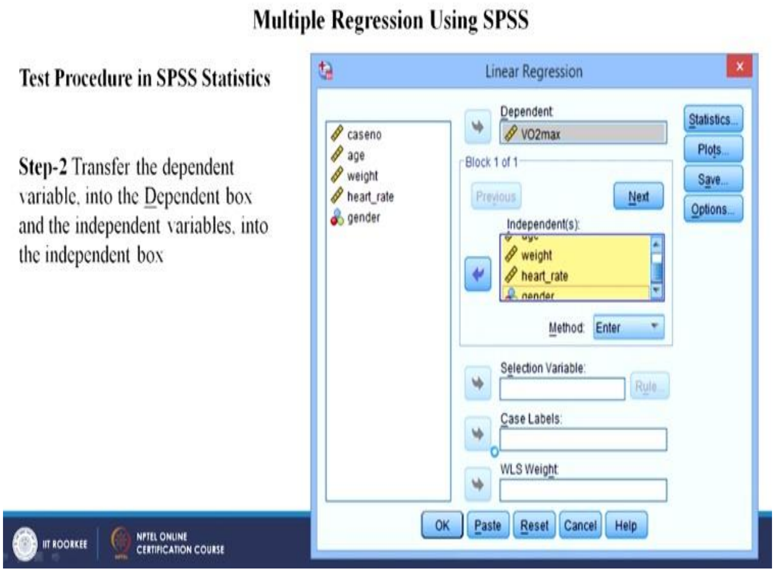
Multiple Regression Using SPSS

Test Procedure in SPSS Statistics

You will be presented with the **Linear Regression** dialogue box

So this is how you take it to the regression so that problem also you can understand,

(Refer Slide Time: 21:09)



Multiple Regression Using SPSS

Test Procedure in SPSS Statistics

Step-2 Transfer the dependent variable, into the Dependent box and the independent variables, into the independent box

Linear Regression

Dependent: VO2max

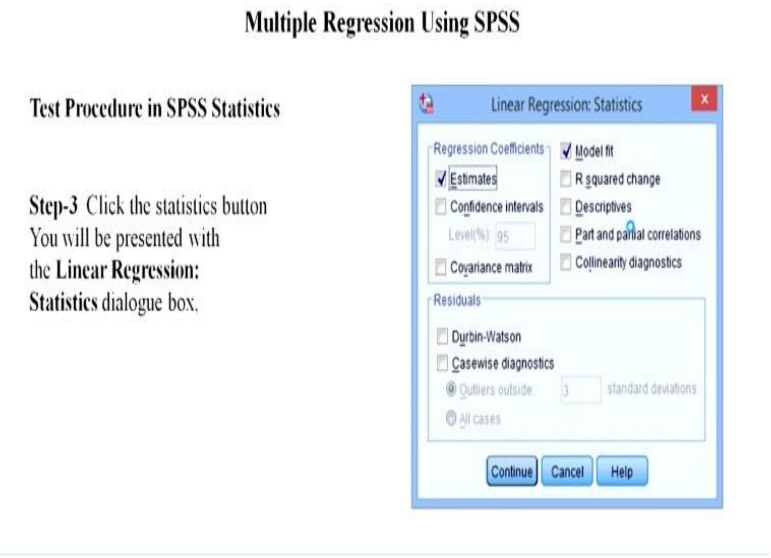
Block 1 of 1

Independent(s): age, weight, heart_rate, gender

Method: Enter

So you see the VO 2max was taken a dependent variable and other things are taken as my independent variable.

(Refer Slide Time: 21:19)



Multiple Regression Using SPSS

Test Procedure in SPSS Statistics

Step-3 Click the statistics button You will be presented with the **Linear Regression: Statistics** dialogue box.

Linear Regression: Statistics

Regression Coefficients

☑ Estimates

□ Confidence intervals

Level(%): 95

□ Covariance matrix

☑ Model fit

□ R squared change

□ Descriptives

□ Part and partial correlations

□ Collinearity diagnostics

Residuals

□ Durbin-Watson

□ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

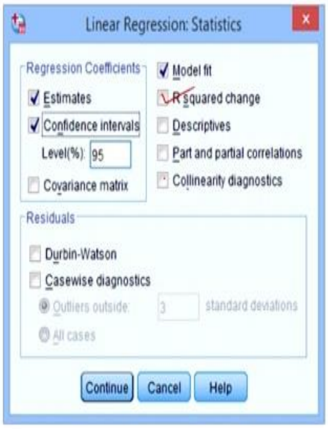
(Refer Slide Time: 21:29)

Multiple Regression Using SPSS

Test Procedure in SPSS Statistics

Step-4 n addition to the options that are selected by default, select confidence intervals

Step-5 Click continue and then Ok



Linear Regression: Statistics

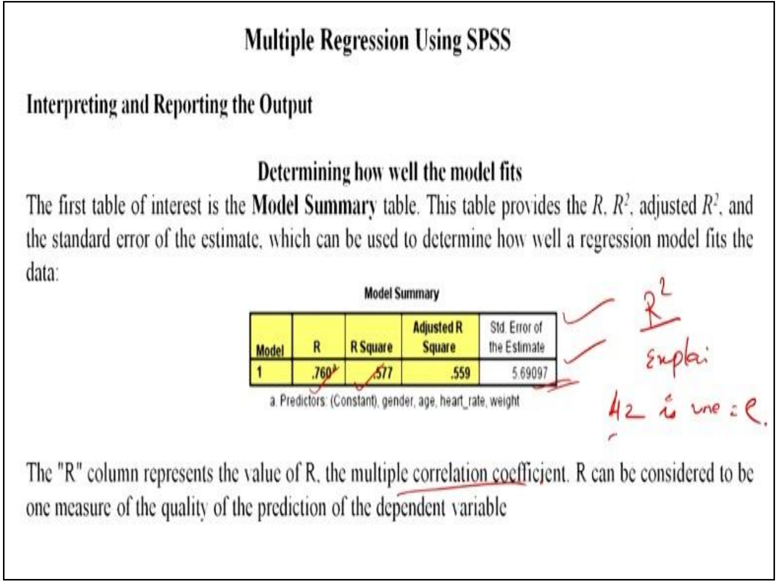
Regression Coefficients
- ☑ Estimates
- ☑ Confidence intervals
- Level(%): 95
- □ Covariance matrix

- ☑ Model fit
- ☑ R squared change
- □ Descriptives
- □ Part and partial correlations
- □ Collinearity diagnostics

Residuals
- □ Durbin-Watson
- □ Casewise diagnostics
- ● Outliers outside: 3 standard deviations
- ○ All cases

So all those steps are explained to you in this case, so what do I want estimates, module fit, R square change you can also take this R square change even you can check the collinearity diagnosing in case there is a multi collinearity problem, to check that are u take it, descriptive also you can use.

(Refer Slide Time: 22:34)



Multiple Regression Using SPSS

Interpreting and Reporting the Output

Determining how well the model fits

The first table of interest is the **Model Summary** table. This table provides the *R*, R^2, adjusted R^2, and the standard error of the estimate, which can be used to determine how well a regression model fits the data:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760^b	.577	.559	5.69097

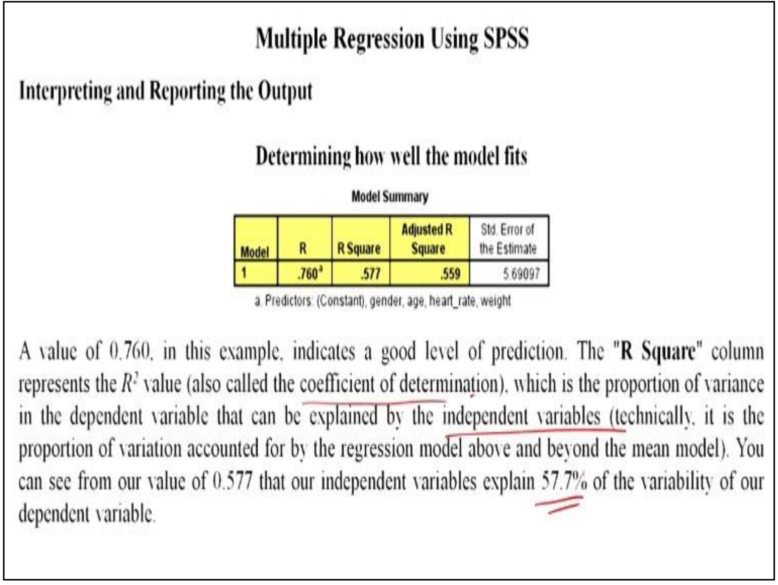
a. Predictors: (Constant), gender, age, heart_rate, weight

The "R" column represents the value of R, the multiple correlation coefficient. R can be considered to be one measure of the quality of the prediction of the dependent variable

Now this was which I done it on my own, so the first table of interest, this was table through my the data which I had when I calculated I found the data that I had was R square R was coming .760 and R square was square of it so .77 standard error of estimate is .69, so the R column it represents the value of R the multiple correlation coefficient R can be considered to be 1 measure of the quality of prediction of the dependent variable.

So as you understand R square what is the R square? R square is the explained variance that means in this case 57 – 1 - .77 how much into .577 is equal to let say .42 almost .42 so 42% is unexplained which is the error is equal to unexplained which is called is error and 57% is only explained in this case.

(Refer Slide Time: 22:39)



Multiple Regression Using SPSS

Interpreting and Reporting the Output

Determining how well the model fits

Model Summary

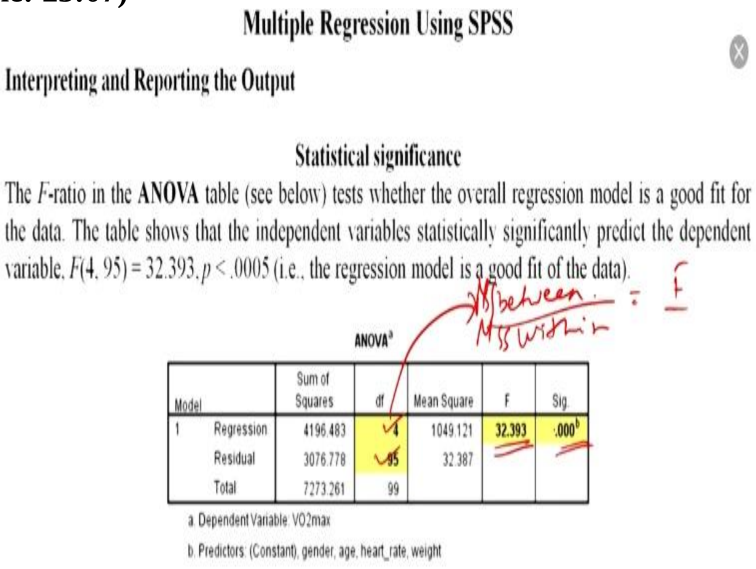
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760[a]	.577	.559	5.69097

a. Predictors: (Constant), gender, age, heart_rate, weight

A value of 0.760, in this example, indicates a good level of prediction. The "**R Square**" column represents the R^2 value (also called the coefficient of determination), which is the proportion of variance in the dependent variable that can be explained by the independent variables (technically, it is the proportion of variation accounted for by the regression model above and beyond the mean model). You can see from our value of 0.577 that our independent variables explain 57.7% of the variability of our dependent variable.

How do you interpreting this results so a value of .760 in this example indicates a good level of prediction, the R Square called the coefficient of determination is the proportion variance in the dependent variable that can be explained by the independent variables , you can see from are here in this case says that independent the explains 57.7%.

(Refer Slide Time: 23:07)



Multiple Regression Using SPSS

Interpreting and Reporting the Output

Statistical significance

The *F*-ratio in the **ANOVA** table (see below) tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, $F(4, 95) = 32.393, p < .0005$ (i.e., the regression model is a good fit of the data).

$$\frac{MSS\ between}{MSS\ within} = F$$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4196.483	4	1049.121	32.393	.000[b]
	Residual	3076.778	95	32.387		
	Total	7273.261	99			

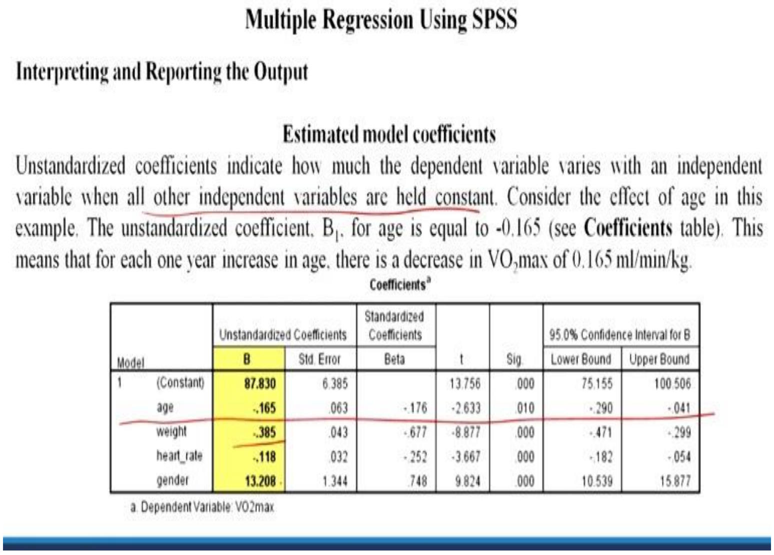
a. Dependent Variable: VO2max

b. Predictors: (Constant), gender, age, heart_rate, weight

Now how do you interpret the output the F-ratio in the ANOVA table you get ANOVA table which I will show you when I do a test, so shows that the overall regression model is a good fit, if you data something like this suppose this is your table ANOVA table now it is a good fit why because it is coming significant. The table shows that the independent variables statistically significantly predict the dependent variable at what level another F value at 4 and 95.

Now I think you are by now have attended by all my lectures if you seriously seen it this 4 and 95 are from what this 4 is from my between the variances between the group divided by 95 is my within. So this is the between by the within, F ratio = F ratio, means sum of square between mean sum of square within is my F-ratio, so F(4,95) is equal to how much 32.39 and significant at a level of .005.

(Refer Slide Time: 24:17)



Multiple Regression Using SPSS

Interpreting and Reporting the Output

Estimated model coefficients

Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. Consider the effect of age in this example. The unstandardized coefficient, B_1, for age is equal to -0.165 (see **Coefficients** table). This means that for each one year increase in age, there is a decrease in VO_2max of 0.165 ml/min/kg.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	87.830	6.385		13.756	.000	75.155	100.506
	age	-.165	.063	-.176	-2.633	.010	-.290	-.041
	weight	-.385	.043	-.677	-8.877	.000	-.471	-.299
	heart_rate	-.118	.032	-.252	-3.667	.000	-.182	-.054
	gender	13.208	1.344	.748	9.824	.000	10.539	15.877

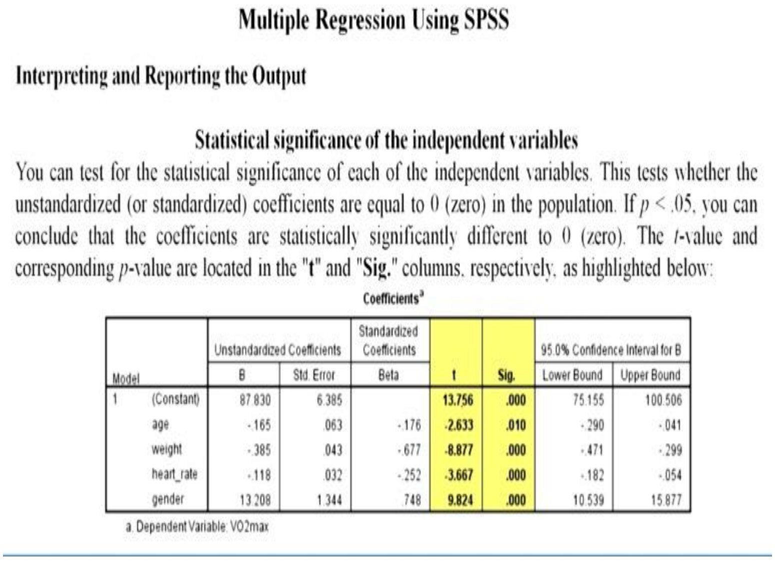
a. Dependent Variable: VO2max

So this is how you should report, now look at this continues if you look the Beta coefficients now what it is saying, what is the beta coefficient is basically B 1, B 2 to the slopes. The unstandardized coefficients there are 2 things that when u see use the SPSS something will be reporting 2 things. One is unstandardized and another is standardized. Now unstandardized coefficients basically are when the raw data has been used and standardized is to in fact the standardized in the way so the means is 0 and standardize is 1, so that now every variable can be compared with each other.

So unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. So, this is the beauty of the regression that you are trying to see the effect of one variable keeping the others constant, considered the effect of age in this example the unstandardized coefficient B 1 now age let us go for age this one, for age is equal to how much 0.165 the unstandardized.

This mean that for each one year increase in age there is a decrease why because minus, VO 2max of .165 millimetre per minute per kg this is what it understands. If similarly weight increases by 1 kilo then -.385 is the change in the VO 2 utilizing similarly other things but if gender is increasing so I do not know what was 0 and what was 1 so let us say 0 is female and 1 is male, so males that means will utilize better the oxygen in compared to females.

(Refer Slide Time: 26:04)



Multiple Regression Using SPSS

Interpreting and Reporting the Output

Statistical significance of the independent variables

You can test for the statistical significance of each of the independent variables. This tests whether the unstandardized (or standardized) coefficients are equal to 0 (zero) in the population. If $p < .05$, you can conclude that the coefficients are statistically significantly different to 0 (zero). The *t*-value and corresponding *p*-value are located in the "**t**" and "**Sig.**" columns, respectively, as highlighted below:

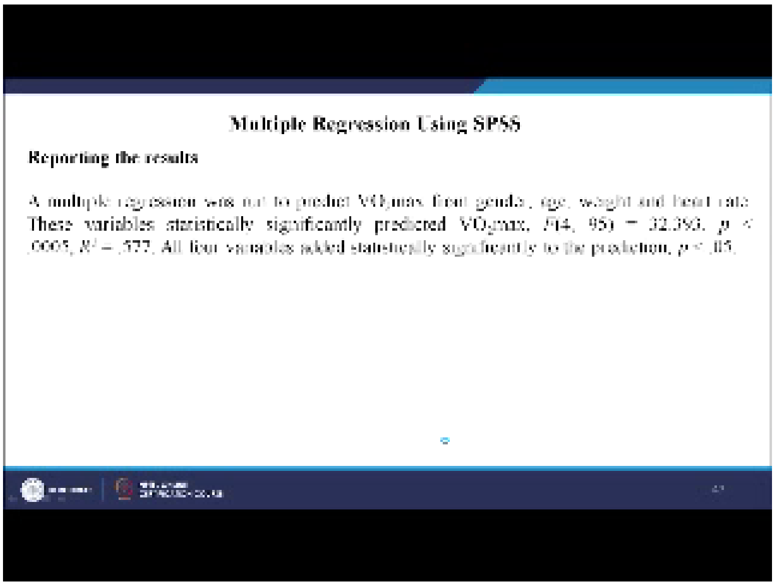
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	**t**	**Sig.**	Lower Bound	Upper Bound
1	(Constant)	87.830	6.385		**13.756**	**.000**	75.155	100.506
	age	-.165	.063	-.176	**-2.633**	**.010**	-.290	-.041
	weight	-.385	.043	-.677	**-8.877**	**.000**	-.471	-.299
	heart_rate	-.118	.032	-.252	**-3.667**	**.000**	-.182	-.054
	gender	13.208	1.344	.748	**9.824**	**.000**	10.539	15.877

a. Dependent Variable: VO2max

Now look at this table also statistical significant of the independent variables. You can test for the statistical significance of each of the independent variables now how now we have taken this look at this column this 2 columns now t =-2.63 for age -8.873 all are above you can see 2. So if it is above 2 you can be very sure that it will be significant. So, what would you says you can test now if you see the twist as this test whether the unstandardized coefficients or standardized coefficients equal to 0 in the population. If p is less than .05 you can conclude that the statistical the coefficients or statistically significant which is the happening in our case. The t value and corresponding p value located in the this column,

(Refer Slide Time: 26:59)



Multiple Regression Using SPSS

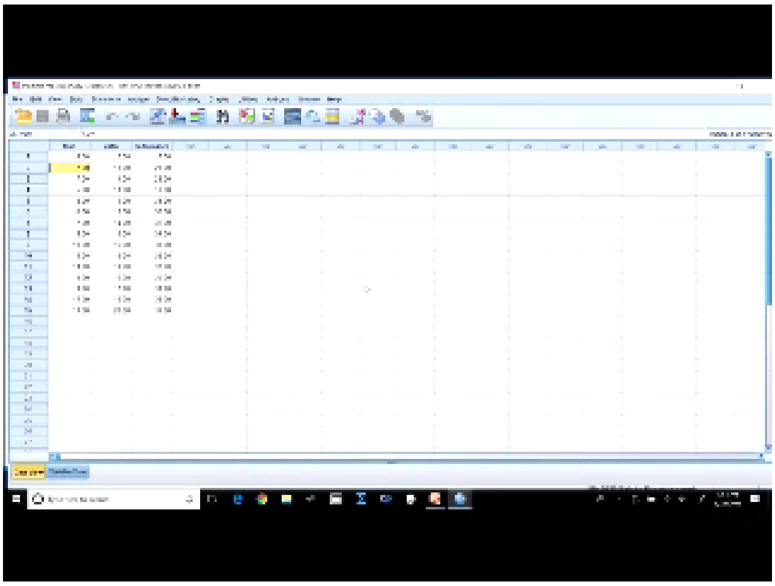
Reporting the results

A multiple regression was run to predict VO_2max from gender, age, weight and heart rate. These variables statistically significantly predicted VO_2max, $F(4, 95) = 32.393$, $p < .0005$, $R^2 = .577$. All four variables added statistically significantly to the prediction, $p < .05$.

So, the how finally when you write in the research report you should write like this, a multiple regression was run to predict this from gender, age, weight and heart rate. These variables statistically significantly predicted VO 2max you should write like this and this one and R square this one or all 4 variable added statistically significant to the prediction at this level. If suppose some of them could not have been significant then you should have written that and you can also write from here you see if you look at these values. The which one has the highest contribution.

If you see now -8 point this is one 9.824 then gender the highest then followed by weight then followed by heart rate and then followed by age, this is what it says. Now let us do one problem which have brought for you.

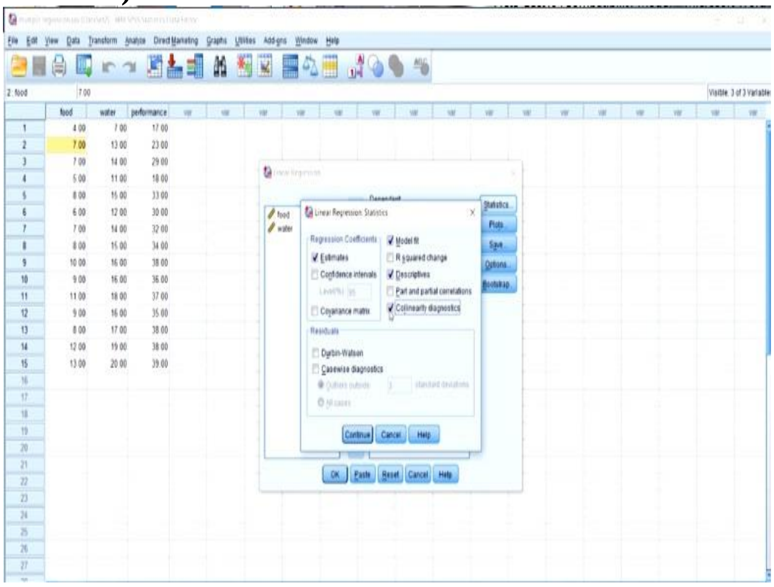
(Refer Slide Time: 27:51)



So, this is case for the first example which we are done. I have not brought their data for second month. So this is the first one so the food and water influencing the performance how do I do go to analyse, go to regression, so I need a just linear regression what is my dependent variable performance what are my independent variables food and water so now I want to see how they are affecting.

So, there are you see there is method step wise, remove, backward, forward but am not do anything at the moment and just doing enter that means all the independent variables are taken simultaneously. If I use the forward are backward it will be taken in different manners, for example forward one by one it will be adding and backward all the independent variables will be taken at one go and then will be eliminated one by one, statistics now I go to statistics I want to see the estimates on the model fit I may use the descriptive for my requirement.

(Refer Slide Time: 28:53)



2: food 7.00

	food	water	performance
1	4.00	7.00	17.00
2	7.00	13.00	23.00
3	7.00	14.00	29.00
4	5.00	11.00	18.00
5	8.00	15.00	33.00
6	6.00	12.00	30.00
7	7.00	14.00	32.00
8	8.00	15.00	34.00
9	10.00	16.00	38.00
10	9.00	16.00	36.00
11	11.00	18.00	37.00
12	9.00	16.00	35.00
13	8.00	17.00	38.00
14	12.00	19.00	38.00
15	13.00	20.00	39.00

Linear Regression Statistics

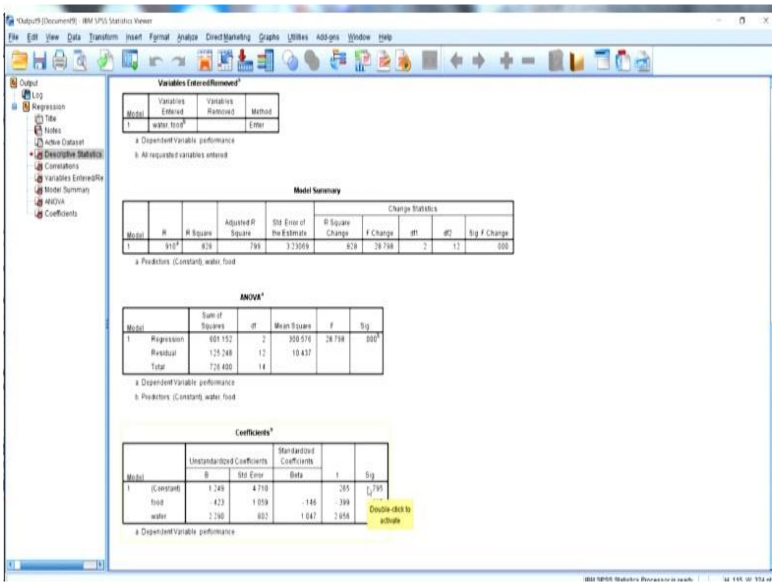
Regression Coefficients
☑ Estimates
□ Confidence intervals
Level(%): 95
□ Covariance matrix

☑ Model fit
□ R squared change
☑ Descriptives
□ Part and partial correlations
☑ Collinearity diagnostics

Residuals
□ Durbin-Watson
□ Casewise diagnostics
● Outliers outside 3 standard deviations
○ All cases

If you want to check for multicollinearity I can check, so if I already done it you may not use it, you can use the R square change also.

(Refer Slide Time: 29:05)



*Output9 [Document9] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Regression
Title
Notes
Active Dataset
Descriptive Statistics
Correlations
Variables Entered/Re
Model Summary
ANOVA
Coefficients
Double-click to activate

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	water, food[b]		Enter

a. Dependent Variable: performance

b. All requested variables entered.

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.910[a]	.828	.799	3.23069	.828	28.798	2	12	.000

a. Predictors: (Constant), water, food

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	601.152	2	300.576	28.798	.000[b]
	Residual	125.248	12	10.437		
	Total	726.400	14			

a. Dependent Variable: performance

b. Predictors: (Constant), water, food

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.249	4.710		.265	.795
	food	-.423	1.059	-.146	-.399	
	water	2.290	.802	1.047	2.856	

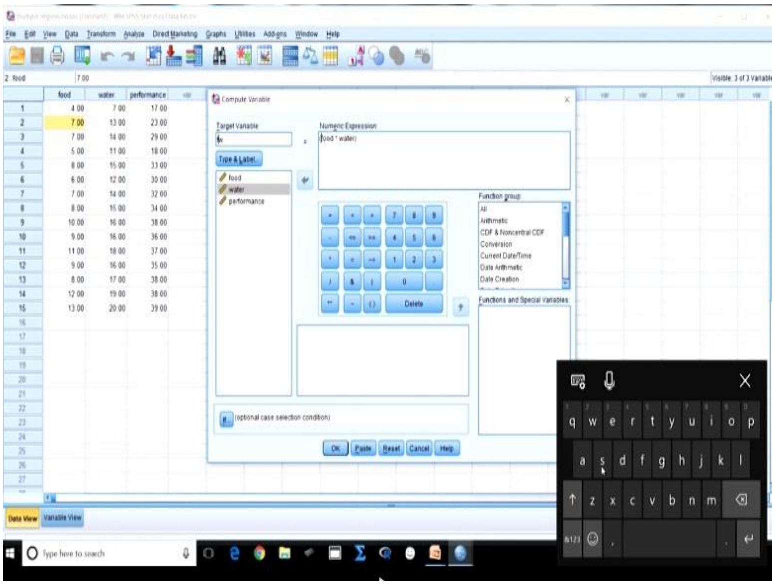
a. Dependent Variable: performance

So now let us look at it, so let us look at the output. so if you look at the output, for example the performance is 31.8 and with the standard deviation of 7.2 the mean standard deviation given, food it is given so there is n is equal to 15, now look at the correlation, the correlation between let us say performance and food is how much .843 performance and water is .908, so performance is more correlated with water.

Food and let us say food and water for example .945 is very high correlation, so let us see the overall model if you see R is coming how much that means these 2 variables are explaining the performance by 82% since the correlation, multiple correlation is . 91 the R square is .82, 82% of performance is explained by these 2 factors water and food intake and this is significant. Now looking at the you know this ANOVA, so if you look at this model is overall significant now let us look at food and water individually.

Now if I see here I see one interesting think is that the food and water in case if you see food and water out of which the t value for food is very low that is much below 2 the value of 2 and therefore and you can see it is non-significant that means we cannot conclude what is null hypothesis in this case food has no effect on the performance and what is my null hypothesis second hypothesis the water has no impact on the performance.

(Refer Slide Time: 31:09)



2 : food 7.00

	food	water	performance
1	4.00	7.00	17.00
2	7.00	13.00	23.00
3	7.00	14.00	29.00
4	5.00	11.00	18.00
5	8.00	15.00	33.00
6	6.00	12.00	30.00
7	7.00	14.00	32.00
8	8.00	15.00	34.00
9	10.00	16.00	38.00
10	9.00	16.00	36.00
11	11.00	18.00	37.00
12	9.00	16.00	35.00
13	8.00	17.00	38.00
14	12.00	19.00	38.00
15	13.00	20.00	39.00

Target Variable: fw

Numeric Expression: (food * water)

So now in some condition you can use this by you can check even an interaction effect also which I am not taken but you can take for example in this also I can show you, you can take and check also for example compute transform compute variable I will create a new variable let us say ok, food into let say water I am did so how do I do this is food multiplied with water so this is my new variable so I am creating, I think we should give a back at here, food I can take FW, leave it FW, so issues so now I will ok new variable must be created here you see FW.

So I want to see the interaction effect also not let me take analyse, regression, linear and I will take this variables also here ok and let me see in whether there is an impact, if we see there is collinearity also I will take, now look at is new table, now when I took this 3 so what is the relationship of performance and food, performance and water and the combine effect, food similarly with this in the combine effect, now when I done and I want to seek whether the what is the model impact first of all the model has improved and if I see the VIF look at the VIF of food it is 83.73 that means since that is the very high correlation between food and water.

So I need to remove one of them and I think water had higher correlation with performance so I remove food ok. So, if I remove food it will be different and I will show you that also, but let us understand here now if you look at the coefficients, now if you coefficients here food has improved earlier it was very bad .6 something, now water is significant and the interaction effect is at the moment not showing any impact but if I do one thing if I take this and I you know run it again for you linear and I take this food out ok and I run it.

Now I do not know whether in change will come or not are not but I hope there should be some change, you see the inflation you know various inflation factor this has improved earlier it was 85 something it has improved but the significance not change much anyway so that does not add anything new to us but one thing is for sure we understood that through a regression equation or through regression helps in predicting and helps in predicting the change in the dependent variable due to a change in the independent variable but this gets highly effected by the presence of the multi collinearity in our case was the multi collinearity, so the food and water.

So that is why it showed very different kind of result but ok. anyway to understand to write you can say here from the early one if you omit this part of you know the interaction part and you take this only you omit this you can see you can say that well food has showing significant effect on performance but water does have an effect of performance and we have also established and relationship for it why we say that the food ones you take the food then generally one takes food after you know lunch time we have seen the people get drowsy and then get tired do not feel like working so that drastically advisedly affects the performance.

But continues water intake keeps people dehydrate you know away from dehydration and keeps them going. so this is something we can logically explain ok, but anyway I think you are clear with what is the impact how to measure multiple regression what thinks you should take care of and how do explain it and how do report it, so well this is all for the multiple regression, standard multiple regression in the next lecture will talk about the hierarchical regression how to do the hierarchical regression and then will be moving to some others forms of regression. So Thank You very much have a nice day.