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

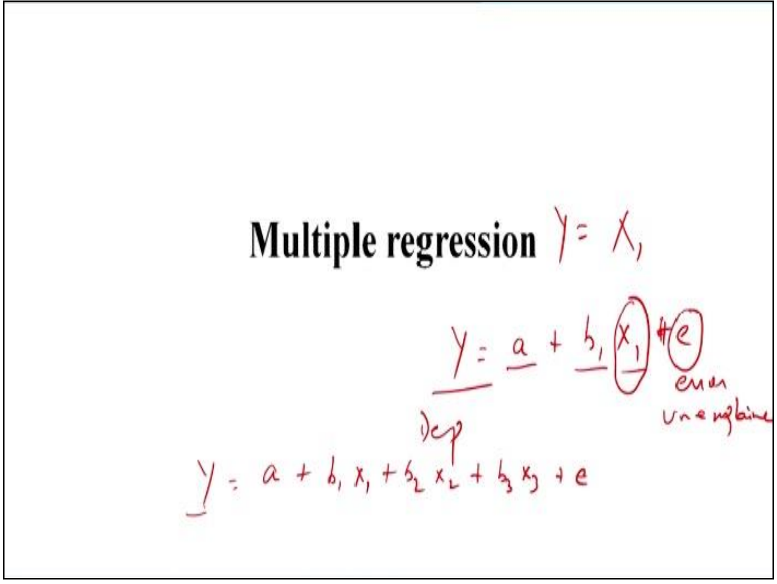
Lecture - 45
Multiple Regression Analysis in SPSS - I

Hello friends, Welcome to the class of Marketing Research and Analysis. In the last lecture we were discussing about correlation, simple correlation, multiple correlation and then we have understood then we went into the concept of regression, so we realise that regression is one we have modelling which helps you to predict some outcome or some dependent variable on basis of some impendent variables which are the predictors.

So we realise that while calculating the regression, the coefficients we realise that we need to understand the factors like you know, the concepts like the unexplained, explained and the total variance. So I drew a diagram and explain to you that what you mean by residual errors or unexplained variance and what you mean by explained variances? And some of these 2 are called the total variance.

Today we will be moving further and understanding about the little higher form of regression which is called the multiple regression, so in the simple regression we were only having one variable called x_1, so we were having x_1 which was effecting the Y, right.

(Refer Slide Time: 01:48)



Multiple regression $Y = X_1$

$$Y = a + b_1 X_1 + e$$

Dep

error unexplaine

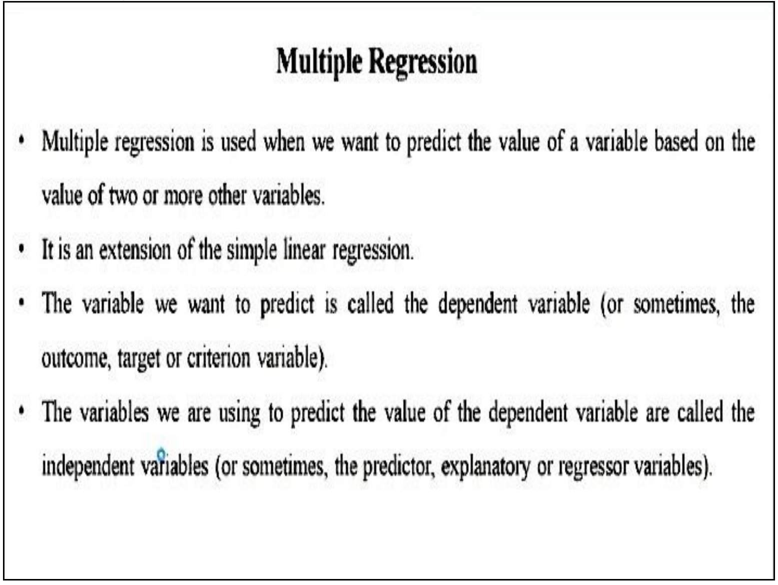
$$Y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + e$$

So what we did was we said Y = a + b_1 + x_1 + e that was the basic formula right, equation, so a was are intercept, b_1 was the slope, x_1 was the first independent variable or the only independent variable and **e** is my error or we say is the unexplained variances**.** Ok, so this is my dependent but what if; if we are having let us say more than one you know independent variable so let us say Y = it becomes b_1 x_1+ b_2 x_2 suppose there are 2 independent variables suppose let us add one more b_3 x_3+ e, so e is as usual the unexplained part.

So now we are having 3 independent variables and we know every independent variable influences the dependent variable Y, so we need to understand simply what it means, when we are saying $Y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$ we mean to say suppose we want to see the effect of X 1, then what you need to do is just say $y = a + b_1 x_1$ keeping b_2 x_2 as constant.

Similarly if you want to check individual the effect of x_2 then you to keep x_1 and x_3 as constant, suppose you want to see the effect of x_3 then you to keep x_1 and x_2 as constant, so now but the equation is become little complicated and tougher because you have 3 slopes now are the beta coefficients b_1, b_2 and b_3 which needs to be found out and then only you can put forth the values and find out the result the final outcome.

(Refer Slide Time: 03:38)



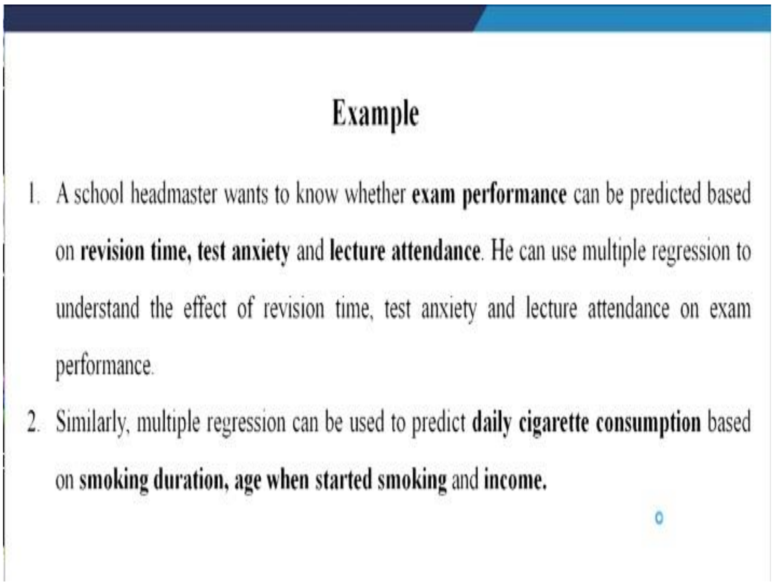
Multiple Regression

- Multiple regression is used when we want to predict the value of a variable based on the value of two or more other variables.
- It is an extension of the simple linear regression.
- The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).
- The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

So how do we proceed lets go step by step, so how multiple regression is defined what it means Multiple regression is used when we want to predict so it is also called predictive analysis, so want to predict the value of a variable based on the value of 2 or more other variables. So now this variable is the dependent variable Y, right and based on the value of 2 or more so if it was one it was simple regression.

So now this is my let us say 2 if there are 2 x_1 and x_2 are it goes up to you can go to x_n, so n number of variables and all independent variables. As it says extension of the simple linear regression the variable we want to predict is called my dependent variable or sometimes the outcome target or criterion variable. The variable we are using to predict the value the dependent variable are called the independent variables or sometimes, the predictor, explanatory or regression variables, ok.

(Refer Slide time: 04:37)



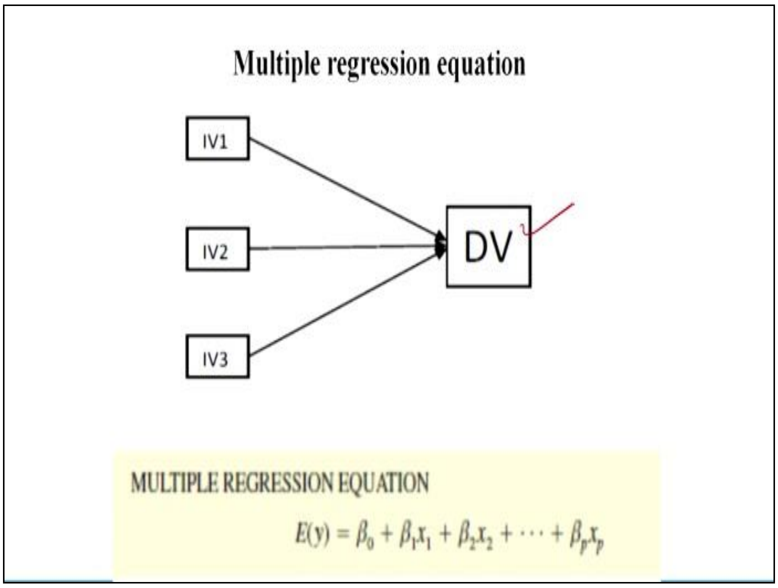
Example

1. A school headmaster wants to know whether **exam performance** can be predicted based on **revision time, test anxiety** and **lecture attendance**. He can use multiple regression to understand the effect of revision time, test anxiety and lecture attendance on exam performance.
2. Similarly, multiple regression can be used to predict **daily cigarette consumption** based on **smoking duration, age when started smoking** and **income.**

Now let us take an example to understand. A school master, headmaster wants to know whether exam performance can be predicted whether the performances in a exam can be predicted of students obviously on the revision time how anxiety and anxious the students are during an exam and how much exams classes there lectures there attended right this is generally teacher wants to know whether there is any relationship between attendance and performance. So what we can do is when we have 3 such variables which can effect the performance so he says this is he can use multiple regression to understand the effect of revision time, test anxiety and lecture attendance on the performance of the exam.

Similarly another example, somebody wants to predict whether daily cigarette consumption, is based on smoking duration, age when started smoking and income, that means how many cigarette that somebody consume does it depend on the amount the time is spends on smoking, the age when started smoking which year they started. And what is the income level, so when we are trying to merge such kind of incidents this is called this simple case of a multiple regression, ok.

(Refer Slide Time: 05:55)



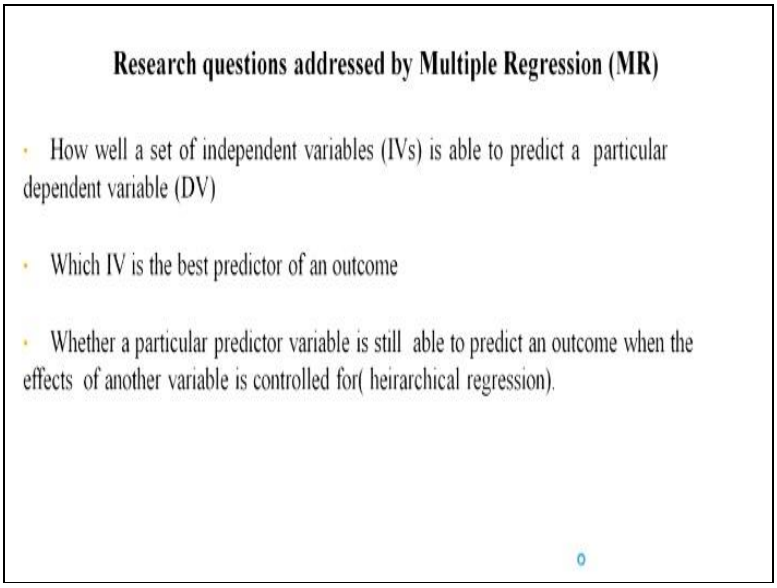
Multiple regression equation

MULTIPLE REGRESSION EQUATION

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Now let us see this, so this is my dependent variable, these are my 3 independent variables and my equation look like this, ok.

(Refer Slide Time: 06:00)



Research questions addressed by Multiple Regression (MR)

- How well a set of independent variables (IVs) is able to predict a particular dependent variable (DV)
- Which IV is the best predictor of an outcome
- Whether a particular predictor variable is still able to predict an outcome when the effects of another variable is controlled for(heirarchical regression).

So what is the research questions addressed by Multiple Regression let us see that, for the first question it says answers is how well a set of independent variables is able to predict a particular dependent variable, how well, how efficiently or effectively this independent variables are able to predict the dependent variable. Second which is more interest people researches, which independent variable suppose there are 3 X 1, X 2, X 3 which dependent variable out of these 3 has the highest influence on Y or the biggest influence or the best predictor.

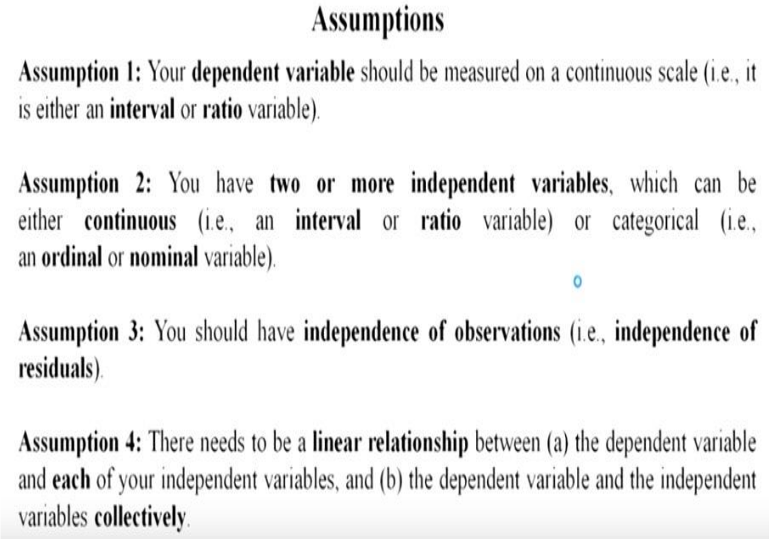
The Third question whether a particular predictor variable is still able to predict in outcome when the effects of another variable is controlled. For example suppose I want only see the effect of this and I want to control these 2, so First case I am trying to see this one and the next case I may take all the 3 and see whether the prediction ability is changing that means what we are saying is the change in R square.

So if you remember we had explained what is the R the coefficient multiple collinearity, the multiple coefficient of correlation and then we said R square is the coefficient of determination and we said how this R, what is this R square, this R square is nothing but it talks it is explained variances, so if there is a change in the explained variances when you take 2 different outcomes in blocks then we can understand what is the change due to the influence of the other new variables that you have entered. So, this method is called a, method is called the hierarchical regression methods which I will be explaining you later on.

Somewhere Assumption and multiple regressions is more or less is the same only but let us see some of them, the dependent variable should be measured on a continuous scale, it is a interval or ratio. Assumption 2: 2 or more independent variables which can be either continuous or categorical that is ordinal or nominal sometimes. Now I will Explain what happens in the may be in next to next class lecture what you do when there is categorical variable, how do you use a categorical variable? It is called a dummy variable regression I will explain that.

A third you should have independence of observation that means each observation are responded major only once, right until then there is a repetition. The fourth there needs to be a linear relationship between the dependent variable and each of your independent variables and b the dependent variable and the independent variables collectively.

(Refer Slide Time: 07:48)



Assumptions

Assumption 1: Your **dependent variable** should be measured on a continuous scale (i.e., it is either an **interval** or **ratio** variable).

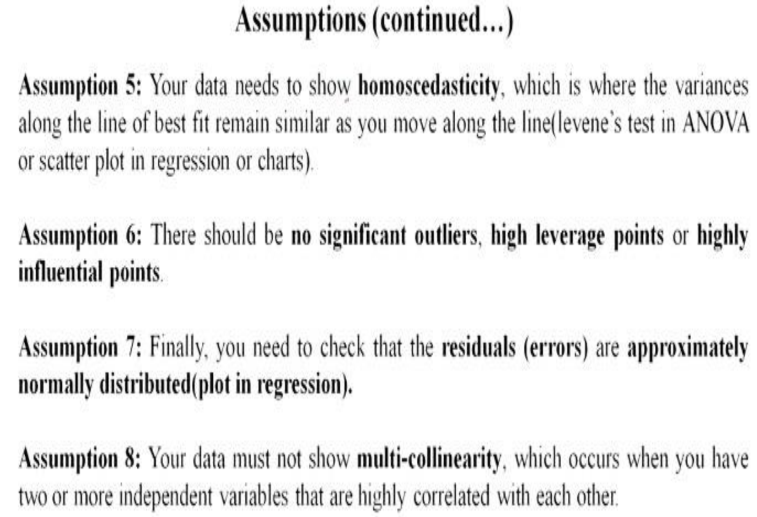
Assumption 2: You have **two or more independent variables**, which can be either **continuous** (i.e., an **interval** or **ratio** variable) or categorical (i.e., an **ordinal** or **nominal** variable).

Assumption 3: You should have **independence of observations** (i.e., **independence of residuals**).

Assumption 4: There needs to be a **linear relationship** between (a) the dependent variable and **each** of your independent variables, and (b) the dependent variable and the independent variables **collectively**.

So there needs to be a linear relationship this is one of the primary assumptions of regression or basically in any parametric test that the relationship should be more or less linear in nature, sometimes I had seen and also mentioned in my class lectures that this linear relationship actually improves when you improve just increase the sample size.

(Refer slide Time: 09:05)



Assumptions (continued...)

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(levene's test in ANOVA or scatter plot in regression or charts).

Assumption 6: There should be **no significant outliers**, **high leverage points** or **highly influential points**.

Assumption 7: Finally, you need to check that the **residuals (errors)** are **approximately normally distributed(plot in regression).**

Assumption 8: Your data must not show **multi-collinearity**, which occurs when you have two or more independent variables that are highly correlated with each other.

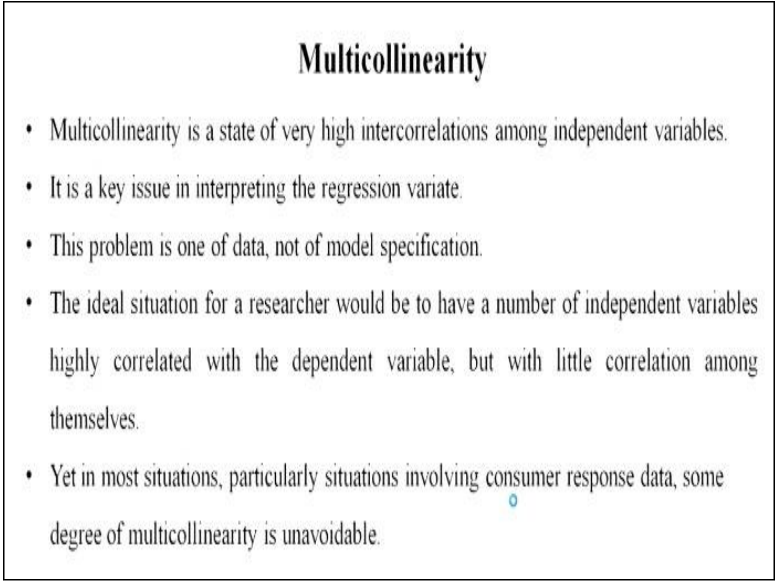
The data needs to be homoscedastic now this is also called as the homogeneity of variance, so there it should follow the assumption of homogeneity of variance, which is generally done through a Levene's test which I have mentioned here or you can just check through a scatter plot in regression or in the graphical charts, so you can just check homoscedastic means it should form a rectangle.

For example if the data if you plotted on the graph it should look like a rectangle and otherwise suppose it looks something like this the data plot has looks something like this it is like the cone or funnel then we say it is a case of a hetroscedastic so it is no more it is evaluation, similarly the 6th assumption is there should not be any significant outliers right, finally you need to check the residuals are approximately normally distributed or not this is also done through a plots in regression which is later on we will see.

So you can see by you know the taking the standardized residuals and the standardized prediction, predicated values if you take both of them and plot a graph you will see whether it is normally distributed or not otherwise you can also use the statistic by standard error where you go to analysis in SPSS and go to you know descriptive and explore then you put the values, then it gives a statistic by standard error this ratio should lie between - 2 to + 2, you can find this in my data preparation lecture.

Assumption 8 is the most one of the most vital assumption and I will be explaining in deep, the data should not show multi-collinearity, which occurs when you have 2 or more independent variables that are highly correlated each other .Now this is the very dangerous thing, right. The presences of multi-collinearity can really disturb the whole scenario, the whole research study.

(Refer Slide Time: 11:00)



Multicollinearity

- Multicollinearity is a state of very high intercorrelations among independent variables.
- It is a key issue in interpreting the regression variate.
- This problem is one of data, not of model specification.
- The ideal situation for a researcher would be to have a number of independent variables highly correlated with the dependent variable, but with little correlation among themselves.
- Yet in most situations, particularly situations involving consumer response data, some degree of multicollinearity is unavoidable.

So let us understand the little bit more about this concept of multi-collinearity what you should do and how you should cover up clear this problem. So what is this multicollinearity let us explain as the names comes from collinearity so multicollinearity is a state of very high inter correlations among the independent variables we are not bothered about relation between the dependent and independent if it is high it is good but we are concerned about the relation among the independent variables, ok.

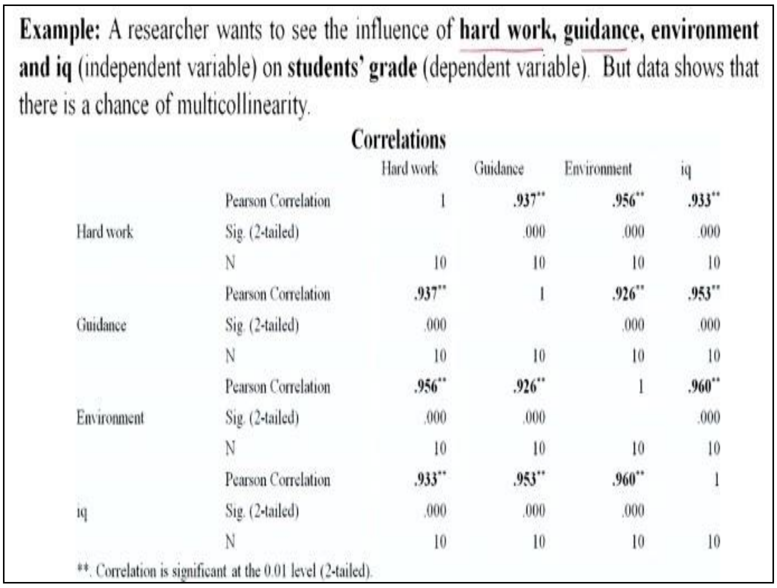
It is a key issue in interpreting the regression variate, ok. The problem is the one of data not of model specification now what it says it is basically the multicollinearity problem is the problem of the data, so if the data is so close to each other that means they are highly close to each other then it means more or less you can understand why should you have 2 independent variables.

They are if they are meaning the same they are having the very high correlation there should be used only one of them why 2? Because the research says that it should be patrimonial in nature you should not be using unnecessarily you know unnecessary variables you should not be using, you should be using minimum to explain the maximum.

The ideal situation for researches would have to be to have a number of independent variables highly correlated with the dependent variable, but with very little correlation or little correlation among themselves that means among the independent variables. Yet in many situations, particularly involving consumer response data, some degree of multicollinearity is unavoidable why? Because in social sciences we have seen the consumer response data is from costal science area.

So it has been seen that many of the question we ask in social sciences are very close to each other and sometime the respondent also does not understand and he feels one question to be like the other, so he confuses or there is very thin line of differences so that is why this problem of multicollinearity can occur ok.

(Refer Slide Time: 13:20)



Example: A researcher wants to see the influence of **hard work, guidance, environment and iq** (independent variable) on **students' grade** (dependent variable). But data shows that there is a chance of multicollinearity.

Correlations

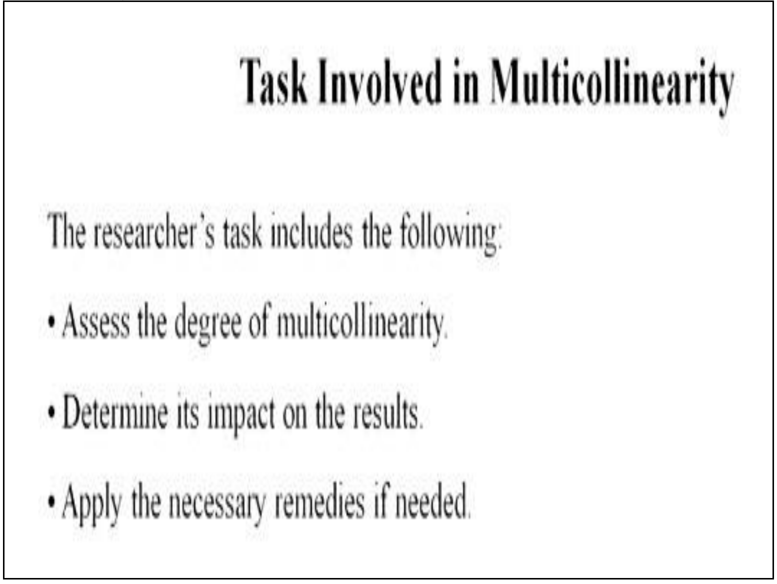
		Hard work	Guidance	Environment	iq
Hard work	Pearson Correlation	1	.937**	.956**	.933**
	Sig. (2-tailed)		.000	.000	.000
	N	10	10	10	10
Guidance	Pearson Correlation	.937**	1	.926**	.953**
	Sig. (2-tailed)	.000		.000	.000
	N	10	10	10	10
Environment	Pearson Correlation	.956**	.926**	1	.960**
	Sig. (2-tailed)	.000	.000		.000
	N	10	10	10	10
iq	Pearson Correlation	.933**	.953**	.960**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	10	10	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Now let us take this look at this, the researcher wants to see the influence of hard work, guidance, environment and iq on student's grade. So what is the dependent variable student's grade and these are the independent variables. Now let us see there is any; we have drawn the correlation chart and we see the hard work and its relationship with their type of guidance is .937, with environment .956 and with iq .933.

Similarly the guidance with hard work gone this is gone with environment .926 and with iq .953, environment with iq is .96 and iq is obviously done with all, so if we look at this right, if you look at this we can see that there is very high correlation, this is very high correlation among the independent variables.

(Refer Slide Time: 14:11)



Task Involved in Multicollinearity

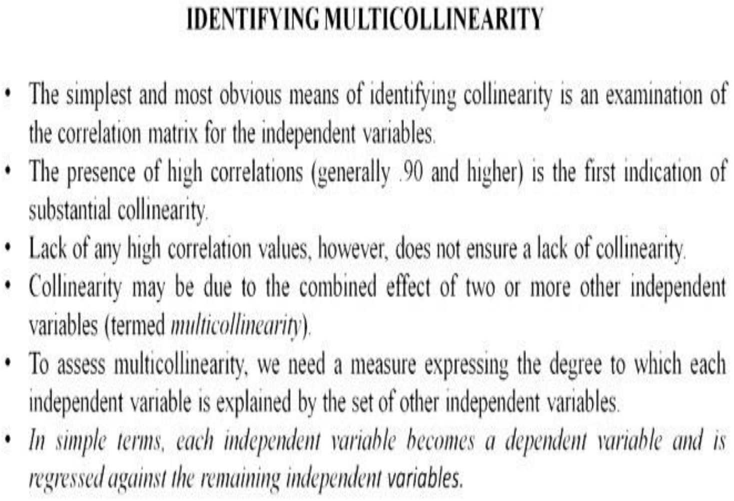
The researcher's task includes the following:

- Assess the degree of multicollinearity.
- Determine its impact on the results.
- Apply the necessary remedies if needed.

So how do you address this problem of what will happen I will explain you what will happen then multicollinearity what happens, so the researcher task includes the following first find out the degree what is the level of multicollinearity is it really dangerous or very bad state that means if it is very high then it will really have a adverse effect on the research outcomes. Determine it is impact on the results that means it is impact on the dependent variable.

Apply the necessary remedies if needed, if you required to eradicate the problem of multicollinearity if it is a series problem then you need to handle it and carefully eradicated. So we will see.

(Refer Slide time: 14:56)



IDENTIFYING MULTICOLLINEARITY

- The simplest and most obvious means of identifying collinearity is an examination of the correlation matrix for the independent variables.
- The presence of high correlations (generally .90 and higher) is the first indication of substantial collinearity.
- Lack of any high correlation values, however, does not ensure a lack of collinearity.
- Collinearity may be due to the combined effect of two or more other independent variables (termed *multicollinearity*).
- To assess multicollinearity, we need a measure expressing the degree to which each independent variable is explained by the set of other independent variables.
- *In simple terms, each independent variable becomes a dependent variable and is regressed against the remaining independent variables.*

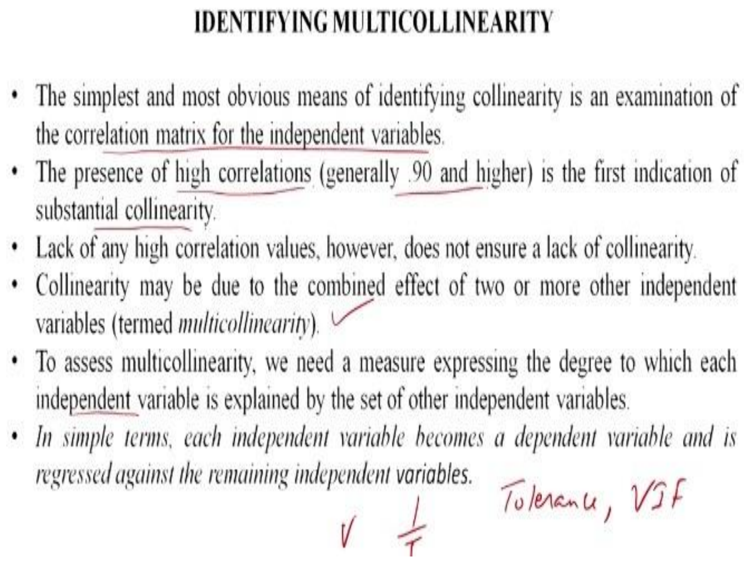
How do I identify? The simplest and most obvious means is an examination of correlation matrix for the independent variables, so what you do is just take the independent variables draw a correlation, now by know you understood correlation so a correlation chart. So, presences of high correlations generally .9 and higher is the first indication of some collinearity. So, I am not saying we are not saying here that it is bound to be but it is an indication that multicollinearity might exist, ok.

But please do not understand that there is a lack of any high correlation values that does not mean that cannot be any collinearity it is not that true, so it is not a very stringent test correlation is not a stringent test to measure collinearity, but it gives to some indication. Collinearity sometimes may be due to the combined effect of 2 or more other independent variables which is multicollinearity.

Now to assess multicollinearity we need a measure expressing the degree to which each independent variable is explained by the set of other independent variables. We look in to each all of them, In simple terms, each independent variable becomes a dependent variable and is regressed against the remaining independent variable I will explain how to do that how to check the multicollinearity.

So, multicollinearity is checked by 2 things one is called the tolerances and other is called a VIF variance influence factor, in fact both of them are related 1 / tolerance, VIF = 1 / tolerance, so this is what.

(Refer Slide Time: 16:30)



IDENTIFYING MULTICOLLINEARITY

- The simplest and most obvious means of identifying collinearity is an examination of the correlation matrix for the independent variables.
- The presence of high correlations (generally .90 and higher) is the first indication of substantial collinearity.
- Lack of any high correlation values, however, does not ensure a lack of collinearity.
- Collinearity may be due to the combined effect of two or more other independent variables (termed *multicollinearity*). ✓
- To assess multicollinearity, we need a measure expressing the degree to which each independent variable is explained by the set of other independent variables.
- *In simple terms, each independent variable becomes a dependent variable and is regressed against the remaining independent variables.*

Toleranu, VIF

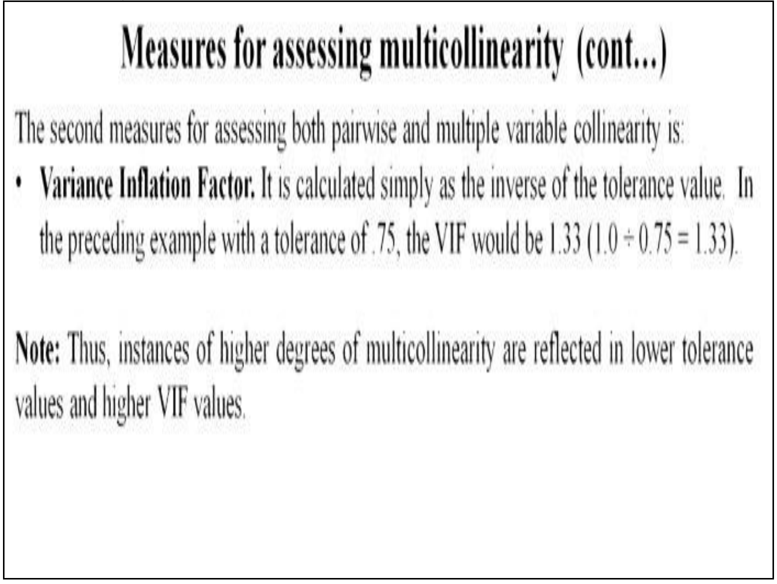
$V \quad \frac{1}{T}$

So how do we check the first measures for assessing both pair wise and multiple variable collinearity is first Tolerance, What is the tolerance? A direct measure of multicollinearity is tolerances, which is defined as the amount of variability of the selected independent variable not explained by the other independent variables. Now how it is calculated, now it is very simple take each independent variable, one at a time and calculate the R square, the amount of that independent variable that is explained by all of the other independent variables in the regression mode right.

In this process the selected independent variable is made a dependent variable predicted by all the other remaining independent variable. Tolerance is then calculated as 1 - R square. For example if the other independent variables explain 25 % of independent variable X 1 R square = .25, then the tolerance of X 1 is how much .75, so let us understand what it means so my; suppose I have X 1, X 2 and X 3 few independent variables. Now first this is my Y = this much but what I will do is to check my tolerance, to calculate tolerance first I will take X 1 as my dependent variable so what is the equation a + b 1 now X 2 + b 2 X 3 + error.

Second I will have X 2 let us say I will take X 2 so what is the equation a + b 1 X 1 + b 2 X 3 + e, the third equation so let us find it for X 3 = a + b 1 X 1 + b 2 X 2 + e so when we do this so when we are doing X 1 independently for X 2 and X 3 so you find the R square and when you subtract this R square from 1 this is what it tells to the tolerance.

(Refer Slide Time: 18:46)



Measures for assessing multicollinearity (cont...)

The second measures for assessing both pairwise and multiple variable collinearity is:

- **Variance Inflation Factor.** It is calculated simply as the inverse of the tolerance value. In the preceding example with a tolerance of .75, the VIF would be 1.33 (1.0 ÷ 0.75 = 1.33).

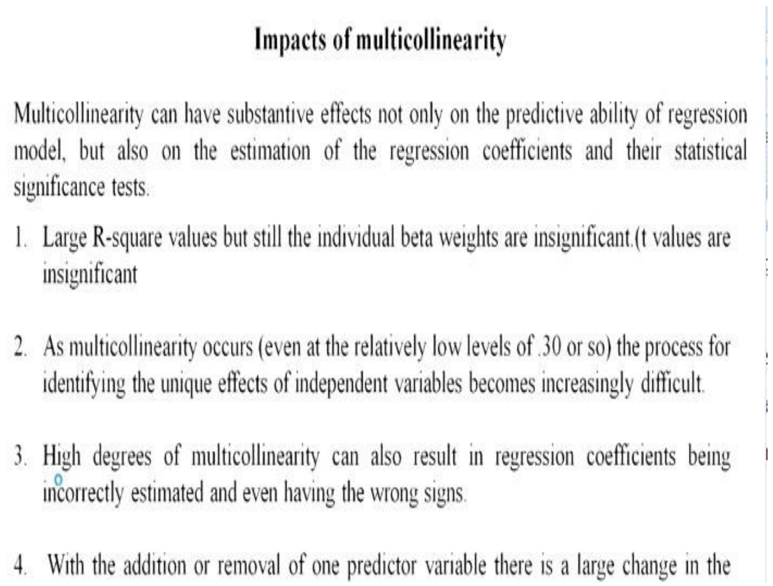
Note: Thus, instances of higher degrees of multicollinearity are reflected in lower tolerance values and higher VIF values.

The second lesson for assessing a collinearity is variance Inflation factor it is calculated simply as the inverse of the tolerance, I told you VIF = 1 / tolerance, in the preceding example a tolerances we had a tolerance of .75, so the VIF was how much now 1 / 0.75 so that is = 1.33. So remember a tolerance value a high tolerance value, tolerance value lies up between 0 and 1. So the closer the tolerance is to 1 the lower the multicollinearity problem exists. Similarly higher the tolerance values closer to 1 lower the multicollinearity.

Similarly lower the VIF that means if the VIF suppose if you say my VIF ranges between let us say you can go to any number but suppose we can say from in between 0 to 10 so the cut off value is little bit 10 actually so if it is 10 we say still is ok but generally researcher have ignored that and there now saying it has been said that scientific community that VIF value this value if it is more than 5 then there is a series case of multicollinearity or very high chance of multicollinearity

And just imagine to have this as 5 you should have the your tolerance is how much 0.2, so that makes it 5, so my, if my tolerance is 0.2 which is 1 - R square = 0.2 so that means you can understand what should be my R and R square.

(Refer Slide Time: 20:25)



Impacts of multicollinearity

Multicollinearity can have substantive effects not only on the predictive ability of regression model, but also on the estimation of the regression coefficients and their statistical significance tests.

1. Large R-square values but still the individual beta weights are insignificant.(t values are insignificant

2. As multicollinearity occurs (even at the relatively low levels of .30 or so) the process for identifying the unique effects of independent variables becomes increasingly difficult.

3. High degrees of multicollinearity can also result in regression coefficients being incorrectly estimated and even having the wrong signs.

4. With the addition or removal of one predictor variable there is a large change in the

Impact of multicollinearity this is very important to understand it has some substantive effects not only on the predictive ability of regression model, but also on the estimation of the regression coefficients right, and the statistical significance tests. This is the first problem, a large R square value, sometimes we are very happy as a researcher that we find when we find R square value to be high we fell that we have already done in high explanation is very high of the model.

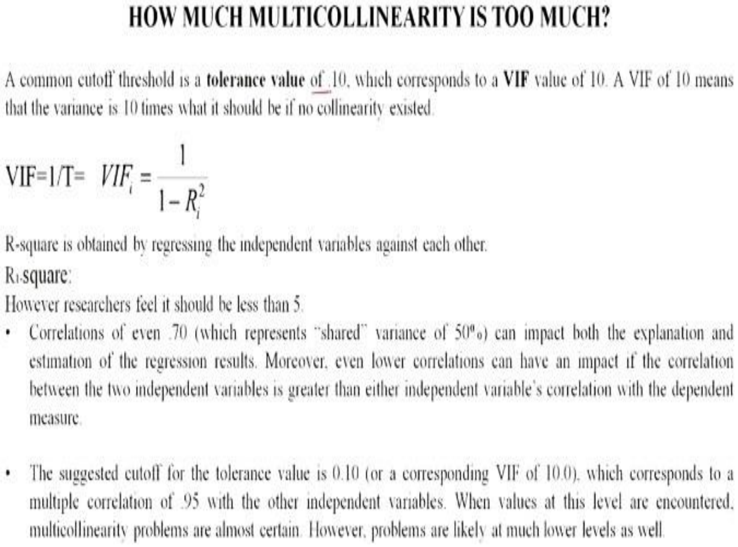
So a large R square values might come but if you find the individual beta weights are insignificant that means the t values are insignificant. Then there is a high chance of there is a presences multicollinearity, that means if multicollinearity persist or exists then you may have a very explanation of R square but your individual beta might comes insignificant, so which should not happen in fact.

Similarly as multicollinearity occurs even at the relatively low levels of .3 or so suppose, the process for identifying the unique effects that means which when X 1 is being checked by keeping X 2 constant, similarly X 2 is being checked by keeping X 1 constant. The effects of independent variables become increasingly difficult. So if there is multicollinearity then you cannot understand which independent variable is impacting the dependent variable in which way. So that is not done than the whole idea about regression is whole sprit of regression is lost ok.

Third high degree of multicollinearity can result in regression coefficient being incorrectly estimated and even having the wrong sign that means what? You from the correlation matrix suppose you understand or you estimate that there is a very positive correlation, so you except a positive effect of the independent variable on a dependent variable. But suppose the coefficient instead of coming positive it is coming negative then you can understand that there is a, this is a because of the presence of multicollinearity right. You might have seen sometimes ok well.

Suddenly one of your independent variable is beta coefficient is coming negative but then your surprised how did you come negative? Why it is coming negative? Please go back and check whether there is multicollinearity problem is existing or not and at very high chance it must be existing ok. And the last is with the addition or removal of one predictor variable that means one of the independent variable is removed there is a large change in the model, the model may become unstable. So, you see how difficult or dangerous it is to have a multicollinearity problem ok. So this is what I was explaining

(Refer Slide Time: 23:09)



HOW MUCH MULTICOLLINEARITY IS TOO MUCH?

A common cutoff threshold is a **tolerance value** of .10, which corresponds to a **VIF** value of 10. A VIF of 10 means that the variance is 10 times what it should be if no collinearity existed.

VIF=1/T= $VIF_i = \frac{1}{1 - R_i^2}$

R-square is obtained by regressing the independent variables against each other.

R_i square:

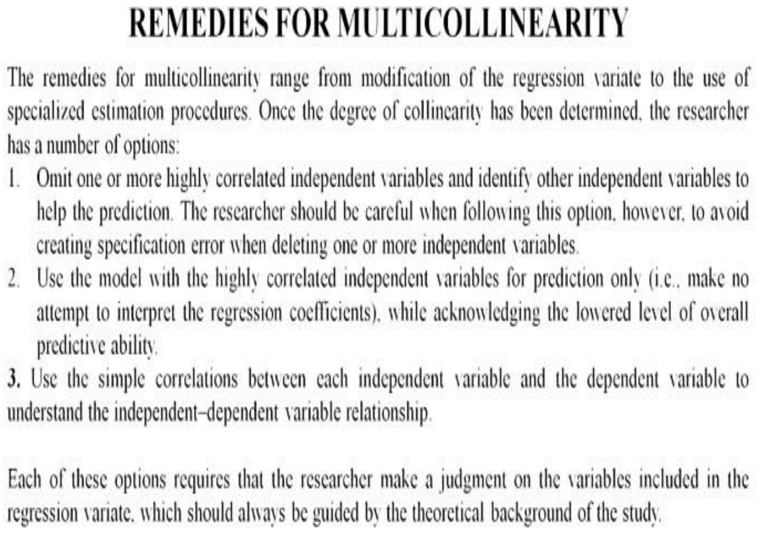
However researchers feel it should be less than 5.

- Correlations of even .70 (which represents "shared" variance of 50%) can impact both the explanation and estimation of the regression results. Moreover, even lower correlations can have an impact if the correlation between the two independent variables is greater than either independent variable's correlation with the dependent measure.

- The suggested cutoff for the tolerance value is 0.10 (or a corresponding VIF of 10.0), which corresponds to a multiple correlation of .95 with the other independent variables. When values at this level are encountered, multicollinearity problems are almost certain. However, problems are likely at much lower levels as well.

A common threshold tolerance value of taken as .1, so obviously the VIF is .10 but then this is not accepted in the scientific community, so you need to have at least I have written here it should be less than 5 right. Correlation of .7 can impact both the explanation and estimation of the regression results. Moreover even lower correlation can have an impact if the correlation of the 2 independent variables is greater than either independent variables correlation that means the combined effect becomes larger and that can create a problem, so this is what is about the multicollinearity. Now how do? What is the remedy? How do I clear this problem?

(Refer Slide Time: 23:53)



REMEDIES FOR MULTICOLLINEARITY

The remedies for multicollinearity range from modification of the regression variate to the use of specialized estimation procedures. Once the degree of collinearity has been determined, the researcher has a number of options:

1. Omit one or more highly correlated independent variables and identify other independent variables to help the prediction. The researcher should be careful when following this option, however, to avoid creating specification error when deleting one or more independent variables.
2. Use the model with the highly correlated independent variables for prediction only (i.e., make no attempt to interpret the regression coefficients), while acknowledging the lowered level of overall predictive ability.
3. Use the simple correlations between each independent variable and the dependent variable to understand the independent–dependent variable relationship.

Each of these options requires that the researcher make a judgment on the variables included in the regression variate, which should always be guided by the theoretical background of the study.

First is if you have a problem on multicollinearity omit one or more highly correlated independent variables. So, you can see which are the independent variables are highly correlated right. So check the once which are highly correlated and delete one of them if possible. Identify other independent variables to help the prediction instead so suppose you have taken the variable X 1 and you find X 1 and X 2 are highly related keep only one of them and bring in a new variable in case your falling short of variables.

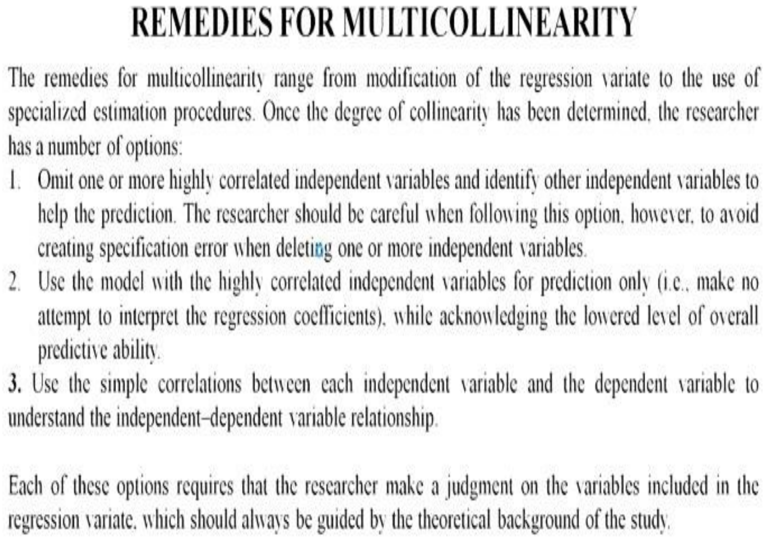
The researchers should be careful when following this option however to avoid creating specification error when deleting one or more independent variables, so you be careful while deleting the variables. Second use the model with the highly correlated independent variables for prediction only right, while acknowledging the lowered level of overall predictive ability that means what it is saying is use the model with highly correlated independent variables for prediction that is make no attempt to interpret the regression coefficients.

So, if there is a problem right, if there is a problem multicollinearity the regression coefficients do not explain you much, the explanation power is reducing right. Third the simple correlations between each independent variable and the dependent variables to understand the so you can simple use what did you say this just use the correlation chart and find out the relationship which I have been telling repeatedly ok.

Now coming to; once we have understood what is regression now we have understood what is multicollinearity problem? Now we will come into what is the types of multiple regression, I will just wind up the class here, in the next lecture we will continue with the both the types regression, the multiple standard multiple regression and the hierarchical multiple regression and how to conduct it on you know calculate by hand and then how to do it on SPSS I will explain.

But I hope today you must be clear by now what is multiple regression and how multiple regression can have help you in understanding the prediction power or the prediction ability of a model and what are the factors that you should be as in assumption you should be maintaining and finally what is the impact of multicollinearity on the multiple regression or the entire predictive modelling study of the researcher or the predictive model of the entire study that the researcher is interested ok. Now, since you have understood the entire thing about multicollinearity.

(Refer Slide Time: 26:20)



REMEDIES FOR MULTICOLLINEARITY

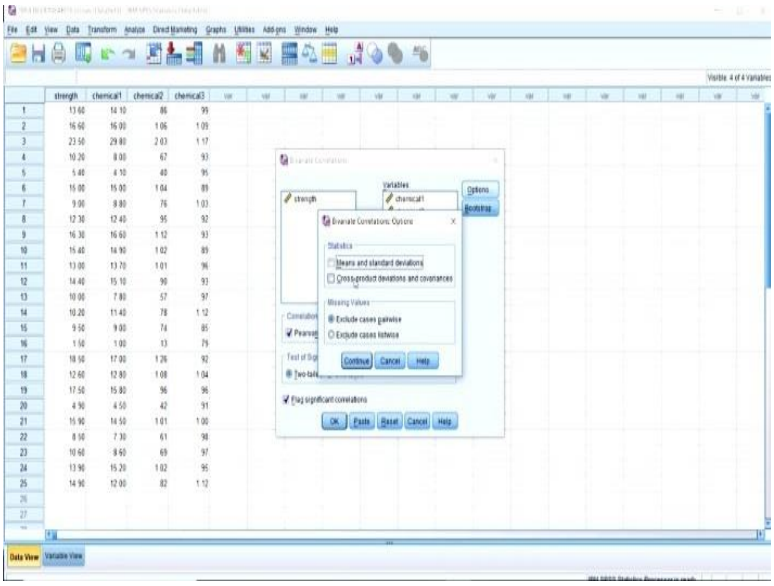
The remedies for multicollinearity range from modification of the regression variate to the use of specialized estimation procedures. Once the degree of collinearity has been determined, the researcher has a number of options:

1. Omit one or more highly correlated independent variables and identify other independent variables to help the prediction. The researcher should be careful when following this option, however, to avoid creating specification error when deleting one or more independent variables.
2. Use the model with the highly correlated independent variables for prediction only (i.e., make no attempt to interpret the regression coefficients), while acknowledging the lowered level of overall predictive ability.
3. Use the simple correlations between each independent variable and the dependent variable to understand the independent–dependent variable relationship.

Each of these options requires that the researcher make a judgment on the variables included in the regression variate, which should always be guided by the theoretical background of the study.

Let us see how to check multicollinearity, I will show you multicollinearity in the SPSS right.

(Refer Slide Time: 26:28)



	strength	chemical1	chemical2	chemical3
1	13.60	14.10	.86	.99
2	16.60	16.00	1.06	1.09
3	23.50	29.80	2.03	1.17
4	10.20	8.00	.67	.93
5	5.40	4.10	.40	.95
6	15.00	15.00	1.04	.89
7	9.00	8.80	.76	1.03
8	12.30	12.40	.95	.92
9	16.30	16.60	1.12	.93
10	15.40	14.90	1.02	.89
11	13.00	13.70	1.01	.96
12	14.40	15.10	.90	.93
13	10.00	7.80	.57	.97
14	10.20	11.40	.78	1.12
15	9.50	9.00	.74	.85
16	1.50	1.00	.13	.79
17	18.50	17.00	1.26	.92
18	12.60	12.80	1.08	1.04
19	17.50	15.80	.96	.96
20	4.90	4.50	.42	.91
21	15.90	14.50	1.01	1.00
22	8.50	7.30	.61	.98
23	10.60	9.60	.69	.97
24	13.90	15.20	1.02	.95
25	14.90	12.00	.82	1.12

Bivariate Correlations

strength

Variables: chemical1

Bivariate Correlations: Options

Statistics
- □ Means and standard deviations
- □ Cross-product deviations and covariances

Missing Values
- ● Exclude cases pairwise
- ○ Exclude cases listwise

☑ Pearson

● Two-tailed

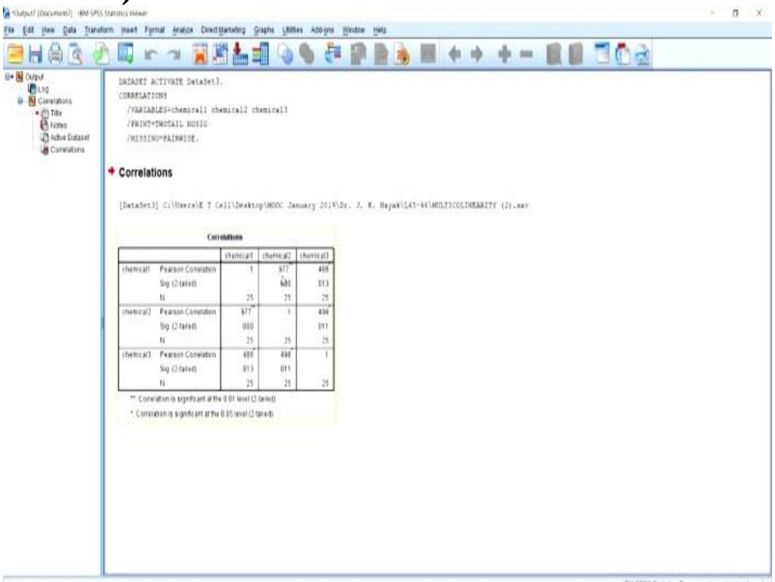
☑ Flag significant correlations

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 4 of 4 Variables
Options
Bootstrap
Continue Cancel Help
OK Paste Reset Cancel Help
Data View Variable View

So I have made a file so let us see this is a file, so this files talks about there are which in this we are having 25 cases and you see there are 3, 4 variables one dependent variable which is the you can assume although it is the collinearity case you can assume that the strength is the dependent variable and 3 types of chemicals are been used to measure the strength to have the change the strength or the impact the strength of the product.

With the 3 independent variables chemical 1, chemical 2, chemical 3 so now if I want to check the multicollinearity first what should I do is first lets us see the simple we will do a correlation, first we will go for a simple correlation. So, we want to see the correlation between the independent variables correct, so we do not want to do anything else so nothing simple. I just want to see a simple correlation.

(Refer Slide Time: 27:22)



```
DATASET ACTIVATE DataSet3.
CORRELATIONS
  /VARIABLES=chemical1 chemical2 chemical3
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

```
[DataSet3] C:\Users\E T Cell\Desktop\MOOC January 2019\Dr. J. K. Nayak\L45-46\MULTICOLINEARITY (2).sav
```

Correlations

		chemical1	chemical2	chemical3
chemical1	Pearson Correlation	1	.977**	.488*
	Sig. (2-tailed)		.000	.013
	N	25	25	25
chemical2	Pearson Correlation	.977**	1	.498*
	Sig. (2-tailed)	.000		.011
	N	25	25	25
chemical3	Pearson Correlation	.488*	.498*	1
	Sig. (2-tailed)	.013	.011	
	N	25	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Correlations
Title
Notes
Active Dataset
Correlations

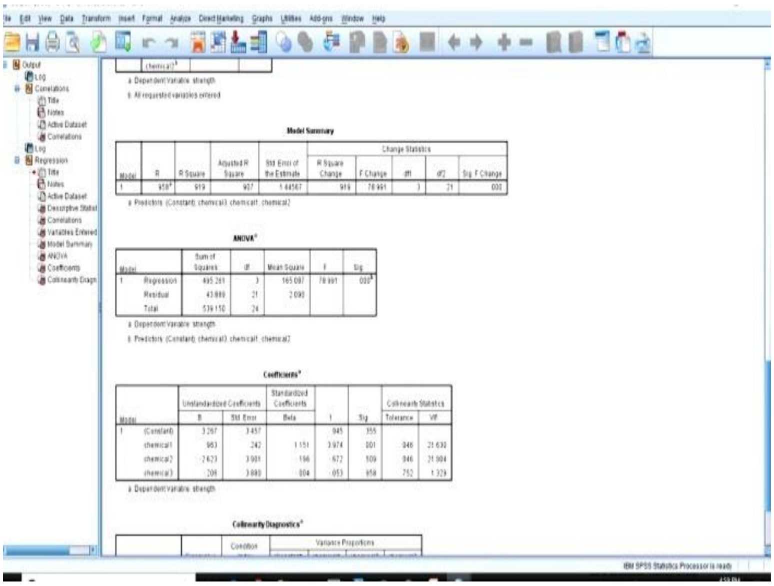
So, what I will do is, I will go to the correlation chart now, so look at it chemical 1 versus chemical 2, this is a very high correlation .977 remember it, in fact you can keep note of it. chemical 1 versus chemical 3 .488 and that is both are significant, this is significant at .01 level ** and this is significant at .05 level *. Now chemical 2 and chemical 3 right so, this is also significant right, but it is significant at .498 and at a .05 level.

So know we can understand that chemical 1 and chemical 2 are very highly connected or very strongly correlated, chemical 1 and chemical 3 are strongly correlated but not so high, chemical 2 and chemical 3 are also correlated but not as high as 1 and 2 ok. Now let us check this is only a just brief glance of the correlation chart. Now we will go to the model again and check through regression, we will see.

Now go to regression and take the strength, so this strength is our dependent variable and these 3 are my independent variables ok. So now what I am doing is I will go to statistics and here I want to check these things R square, descriptive, collinearity diagnostics, part and partial correlation if you want you can do otherwise it is ok, my main interest is this.

So, now look at it, know look at the model first, by the model says the predictors chemical 1, 2 and 3 these 3 the multiple correlation value, multiple correlation value which we have done in the last to last lecture, but there how to calculate is .958. So, what is my R square, the square of this .919. Now adjusted R square is something which says that you it automatically fall down when you introduce unnecessary variables. So, this case this model, let us says the oral model is significant and we have a very decent R square value, very high R square value in fact right.

(Refer Slide Time: 28:41)



	chemical2[b]		

a. Dependent Variable: strength

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	.958[a]	.919	.907	1.44567	.919	78.991	3	21	.000

a. Predictors: (Constant), chemical3, chemical1, chemical2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	495.261	3	165.087	78.991	.000[b]
	Residual	43.889	21	2.090		
	Total	539.150	24			

a. Dependent Variable: strength

b. Predictors: (Constant), chemical3, chemical1, chemical2

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	3.267	3.457		.945	.355		
	chemical1	.963	.242	1.151	3.974	.001	.046	21.630
	chemical2	-2.623	3.901	-.196	-.672	.509	.046	21.904
	chemical3	.209	3.880	.004	.053	.958	.752	1.329

a. Dependent Variable: strength

Collinearity Diagnostics[a]

		Condition Index	Variance Proportions			

Now let us go to the you know this is the area which we are interested in now look at the each chemicals right, now if you look at the chemical 1 so what did we say the t value is 3.974 and it is significant at 0.01 level so what is the tolerances value .046 right, so we said tolerance value should lie between 0 and 1 right and the variance replace factor is 21.63 so it is highly it is series multicollinearity problem.

Chemical 2 is again showing 21.3, chemical 3 is showing a 1.329. So that means what we can understand here that there is a problem of multicollinearity in this case, but chemical 3 which is the one, if you look at the chemical 3, this chemical 3 does not have a; show a problem. So if you look at the tolerances it is what it is saying that tolerances is what 1 - R square. So when we are taking chemical 3 as a dependent variable and we are measuring the tolerance we can see that it gives us a very descent tolerance value and says there is no problem of multicollinearity.

But multicollinearity is very high problem in between 1 and 2 and that to we are seen during the correlation also, if we just go back to the correlation you see 1 and 2 there was a very high correlation among them, so this is again a substantiated and justified by looking at the VIF right. So, once you have it then you can understand the chemical 1 and 2 out of these 2 I can only retain one and may be leave the other right. So this is what you do in the case of multicollinearity.

So, well I hope you have understood what is regression, what is simple regression, what is multiple regression and what are the different assumptions involved in it, what is multicollinearity and how multicollinearity is very dangerous thing and it can completely destroy the whole predictive power of a study and finally how to check for multicollinearity and then how you should remedy, what remedy you should you know have for removing this problem of multicollinearity.

Thank You for the day we will meet in the next lecture and we will talk about the different types of regression and I will show you how to calculate and then we will also see how to do it on case basis. Thank You So Much.