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

Lecture - 47
Stepwise regression & Hierarchical regression

Welcome everyone to the class of marketing research and analysis and today we will be continue with our from last session, in the last session we discussing about regression and we have done with simple regression and then we started with multiple regression, we did with the multiple regression is the case when there is a more than one predictor variable and there is the single dependent variable.

So, when you have more than you know one independent or predicator variables, then it is case of multiple regression, but then the question there are few special cases are during multiple regression test which we need to check and today will be covering that so how to conduct that multiple regression which we are solve through some cases by hand and then did to SPSS also but then the one problem that arises is to sometimes we need to understand which predictors are most important during any for any during any study right already researched our study.

So, to do that we have 2 methods which will be conducting today will be doing today one is call the stepwise regression and other is call the hierarchical regression, so what stepwise regression means and what hierarchical regression means I will be explaining both of it systematically, sometimes people think it is a very confusing and some people they think it is very easy now it is none of it actually you need to understand what exactly stepwise means and what are other ways of entering the data in fact stepwise is nothing but and hierarchical is nothing but way you enter the data into the regression equation right in the case of multiple regression so let us see, so as it is says.

(Refer Slide Time: 02:15)

Introduction

- In multiple regression contexts, researchers are very often interested in determining the "best" predictors in the analysis.

- The researcher may simply be interested in explaining the most variability in the dependent variable with the fewest possible predictors(parsimony).

- Two approaches to determining the quality of predictors are: (1) stepwise regression and (2) hierarchical regression.

In multiple regression contexts, researchers are very often interested in determining the best predictors in the analysis. So which is the best predicator I have 3 predictors are 4 predictors let us say, predictors are my independent variables which one is the best out of it how to do we know. So, the researcher may simply be interested in explaining the most variability of the one which has the highest in the dependent variable with the fewest possible predictors.

So, if I can predict, I can create a model which can I explain me the highest variance right the variations with the lowest number of predictors or independent variables then it is more it is a good model and why because in research one of the conditions is that research says that it should be parsimonious in nature that means you should not be utilising unnecessarily excessive data or excessive variables ok. So, the idea is to have the best output with the minimum input ok.

(Refer Slide Time: 03:11)

Stepwise regression

- Stepwise methods are frequently used in psychological research to evaluate the order of importance of variables and to select useful sets of variables.

- It involves developing a sequence of linear models through variable entry as determined by computer algorithms that, can be viewed as a variation of the forward selection method since predictor variables are entered one at a time.

- However, true stepwise entry differs from forward entry in that at each step of a stepwise analysis the removal of each entered predictor is also considered.

- The entered predictors are deleted in subsequent steps if they no longer contribute appreciable unique predictive power to the regression when considered in combination with newly entered predictors

Two approaches to determine the quality of the predictors are as mentioned earlier stepwise regression and hierarchical regression let us see. What is the stepwise regression? Let us see, so this frequently used in psychological research to evaluate the order of the important of variables and to select useful set of variables. So, what is says in which order should I enter why is the order important? The question that I might be coming to your mind when I am saying this wise the order important.

If I place in any way the variables how does it matter it does matter it matters because when I am putting in the variables in haphazard manner then it what happens is suppose I put in the first variable and then the put in second variable so what happens is when I put in the second variable automatically there is the change in the first variable also in the coefficient in the first variable.

So, because of this I should always have a logical flow of entering the variables. So, if I am using a logic let us say the first variable is the one which has the highest explanation power then the second one has got the second best explanation power the third is the third best explanation power then what happens I am more in a logical way I am able to explain the things. It involves developing a sequence of linear models through variable entry as determined by computer algorithms that can be viewed as a variation of forward selection.

So, if you have done regression on SPSS or Minitab you will see there are 2 methods called the forward and the backward the selection method. So, the forward selection method is very very close to the stepwise you know regression method but then what is the different there

must be the some difference otherwise would not have been 2 names correct I will be explain that.

So, but understand in forward regression selection method what we do is we try to select the first predictor variable which has the highest R square of the highest explanation, since R square is my explanation, explanatory power. So, highest R square which shows I will use that as my first variable. Second I use my, the next one which uses the second highest explanation power. So, and I go on till the end right and the condition here that I am not omitting any variable I am just entering the variables during the forward selection.

In the case of backward selection what happens I take all the variables at one time and find out the R square and then I find out which one of them is weakest and I start eliminating one by one till I reach a model after which there is no improvement ok. However true stepwise entry differs from forward entry in that at each step of a stepwise analysis the removal of each entered predictor is also considered.

So, different between forward and stepwise is that although the look similar the approach is similar difference is that in the forward as I had said you do not think of removing anything. So, if are there are 3 variables the first, second, third I am including all the 3 in the study and I am trying to check what is change in the R square when I am introducing X 1 then I am introducing X 2 and finally X 3.

But in the stepwise the difference is that when I introduce X 1 nothing is there but when I introduce X 2 I will try to see what is the change in now X 1 whether X 1 is actually now contributing or it is no more contributing why will it happen why will such a situation occur you need to understand that this is sometimes you can understand as this is nothing but an interaction. So, when a new variable is entering because of the interaction are the may be some kind of correlation between these 2 variables because of the which there their own property changes.

So, now the property because of the change in the property it could so happen that this new the earlier variable which was the strongest of variable as some of became a weak variable because of the presence of this. In for example I can quote from the indian mythology there

was the Mahabharatha war in the Mahabharatha war Bhishma the character called Bhishma was the great warrior.

So, Bhishma was enter in to the you know war as the leader, he as the defence head. Then one time they found to in order to defeat Bhishma they introduce a new variable was introduce this character was a transgender. So, Bhishma head promised that I would not fight to anybody who is a lady, she was a half lady so I would not fight. So, with the presence of this X 2 new variable when this new variable came into play the X 1 which was the very powerful variable.

Which was the explaining highest or was the defence leader it automatically became a weak character and because of the presence of X 2, X 1 became a redundant character and it was removed, Bhishma was killed in the war he did not raises bow and arrow. So, sometimes a very powerful variable can also became weak due to the presence of a new variable. So, that is the stepwise regression considers the removal also during the process.

The entered predictors are deleted in subsequent steps if they no longer contribute appreciable unique predictive power to the regression when considered in combination with the newly entered predictors. So, if in combination how are the performing sometimes the performance may increase which is the symbiotic relationship sometime the performance may be decrease because of the coming of the new variable.

(Refer Slide Time: 09:03)

Step limitations what is the limitation of the stepwise method it is not gives the method is not without limitation you should not blindly get in to this method, stepwise regression will typically not result in the best set of predictors and could even result in selecting none of the best predictors, how will see I will give an example. The order of variable entry can be important how do you choose the order of variable you can choose it through the correlation highest correlation value or just take the t value now how do you do that. So, simply you take Y you take X 1 ok, find the regression suppose there are X 1, X 2, X 3 right.

So, 3 independent equations I will form I will run it Y and X 1 one equation Y and X 2 another equation Y and X 3 another equation and I will see the t coefficient in all the 3 cases. The one which is the highest I will select it has my first entry variable suppose X t is having the highest t value then that one will be enter into the stepwise regression first ok X 2.

Second after doing this I will again run a regression equation, now how do I run I will run by a combination X 2 and X 1, X 2 and X 3 ok. Now so 2 equations one this and other is this, so when I am running now I will see which one is giving me larger contribution. So, accordingly suppose a X 2 and X 1 is giving my larger contributing t value then I will select X 1 is the next entry variable.

So, first I entered X 2 then I entered X 1 and the remaining is obviously X 3 ok this is how you do. Each time you do regression so there will be multiple regression equations, so sometimes it is very difficult do it by hand you can use the software ok. If any of the predictors are correlated with each other, if the predictors they independent variables are correlated the relative amount of variance in the criterion variable that dependent right explained by each of the predictors can change drastically.

If they are correlated then if it is the positive correlation may be it will improve it is the negative correlation it will decrease, when the order of entry is changed that is what it says. Let us take this example stepwise selection of a team of let us say basket ball players or cricket players whatever, first picks the best potential player, so which we did it right for example X 2 then in the context of the characteristics of this player in the context of the characteristics.

Because you have taken already X 2 now we have to take X 1 or X 3 right picks the second best which you did and then proceeds to pick the rest of the 5 players in this manner may be this is the basket ball team ok. So, basket ball team so I need 5 players so let us say any sports you can think of. So, I have select the 5 players ok alternatively there is another method is also 5 potential players which play together best as a team are selected.

So, one way is through the stepwise what I did was I can select one by one the best than the next best than the next best up to 5 people or I can also select the team of 5 players by using my logic, sometimes the best 5 may not may be best on paper but when it comes reality on the field there combined team game as a team they do not perform well. But they can be other 5 team players who when they join together that makes a very formidable strong team.

But if you go by you know the first method then they will not be selected, so the question is alternatively says 5 potential players which play together best as a team are selected. The team that is picked via this method might not have any of the players from the stepwise picked team, and could perform much better. So, what I mean to say here is although stepwise method is very good method and very powerful method.

But sometimes stepwise method does not heal the desired result right and rather it gives to very distorted result which people ignore they try to blindly use the methods and that is where they you know get into trouble, so how to do this stepwise regression and then I tell you what is the alternative so which here we said the 5 potential players which plays together is a team. So how do I select it as a team? So I will explain that will come through the hierarchical regression method.

(Refer Slide Time: 13:44)

Stepwise regression in SPSS

- **Example:** Stepwise regression was used to regress mother's education level (ma_ed), father's education level (fa_ed), parent's income (par_inc), and faculty interaction level (fac_int) on years to graduation (years_grad).

$$\frac{Y}{Grad} = X_1$$

So, let us first complete this stepwise regression, how to do this in SPPS. So, this is the case which have taken from a book I was reading so it was this is I know related with how much time a person takes to graduate how much year or time takes to graduate these dependent year to graduate right graduation its dependent on its says factors like the mother's educational level. Mother's educational level is this one X m let us say + father educational level X f + parents income X I + the faculty interaction so X F faculty interaction those 4 variables. It is says will effect the dependent variable the criterion variable graduation graduating time how much time takes to graduate.

(Refer Slide Time: 14:49)

Steps in conducting stepwise regression in SPSS (continued...)

- Transfer dependent variable to dependent box and independent variable to independent box.
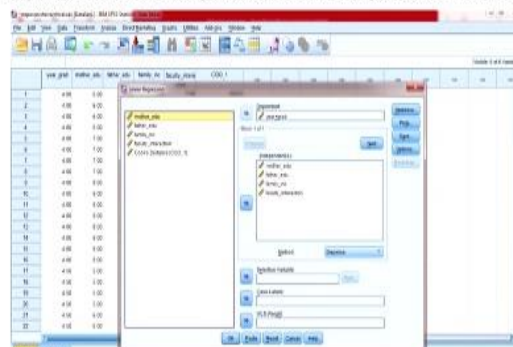
So how to do that so you can see I will just show you then I will get into SPPS. So, you go to linear regression and then you take all the variables the dependent variable to the dependent

side and the independent variable to the independent side and here you use the stepwise.

(Refer Slide Time: 15:03)
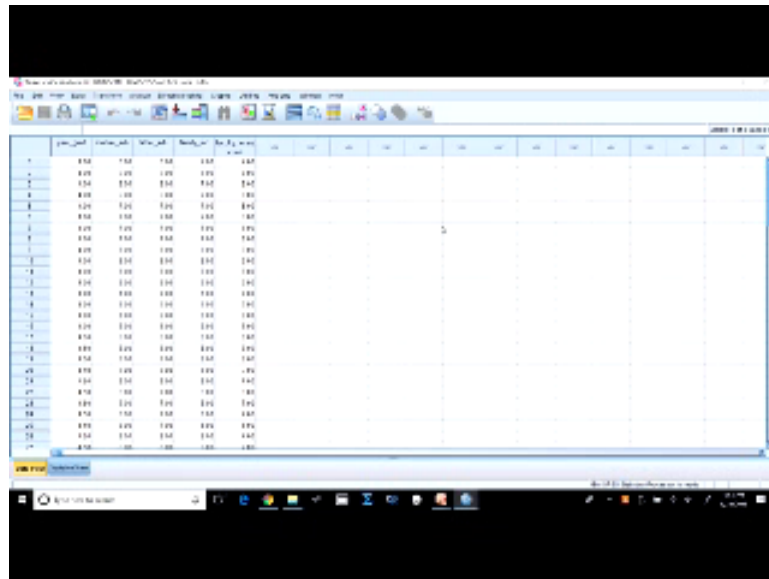
So, let me go set the data set, this is the data set I was talking about, so years to graduate has been given right this has sum of the years to graduate the education of the mother is given to you the fathers education is given to you. The family income is given to you and the faculty interaction level is given to you. So, what I can do is I can simply first understand the correlation among this variables I can just simply go to analyze and do a correlation and I will take let me first change the variable view.

So, this is also scale this is not ok. So analyse correlate by Bivariate and you take all the variables right 1 2 3 4 5 right and we will run a correlation and just see how this looks.

(Refer Slide Time: 15:50)
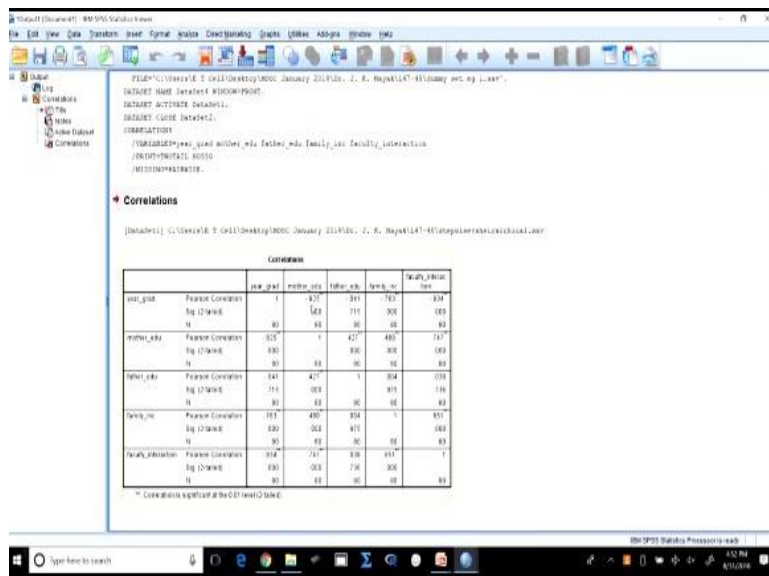
So, if we look at correlation you can see the years of graduation which the dependent variable is strongly correlated with the mothers education may be minus is the negative so do not worry because what does it means negative means 1 unit year increase that means with unit increasing in the mother's education let say there is a decrease in the years graduation.

So it as, it should be inversely related logically also you think if the education of mother is more than we would see can contribute more to the child and the years to graduate should come down so that say the negative relationship.

Similarly faculty interaction if it is more the time taken should be less so it is a negative relationship. So, -8 point if you look this looks to be the highest correlation, seconds it looks to be the mother education, which is highest correlated. Year of graduation and father's education is very, very poorly correlated family income is also correlated right. Now among the others you can see like for the mother's education father's but this is not for interest ok. So, lets us go back to the dataset, now how do I run my stepwise regression, so regression I go to analyze linear. So, this my dependent variable and I take all my independent variables here, so I have taken all my independent.

(Refer Slide Time: 17:16)
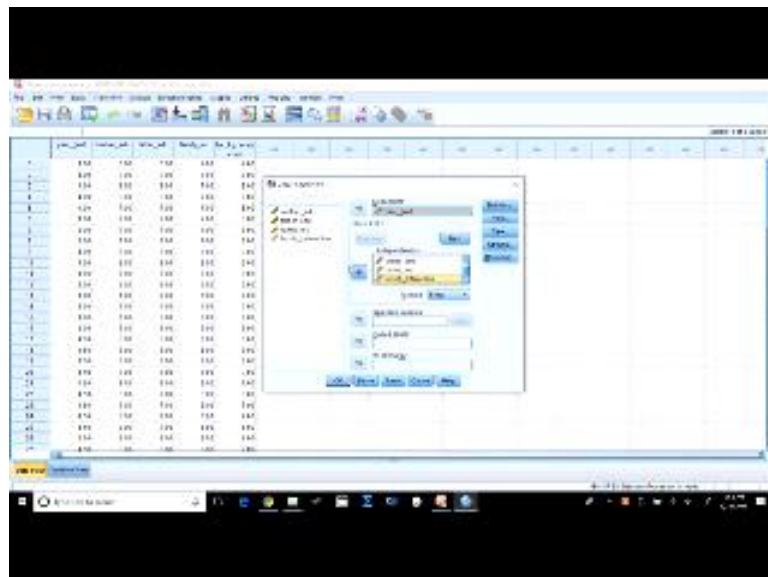
I will ask the computer to do it for me now, so statistics so I need to see here you can check for you know descriptive, partial correlation. If you want if you have doubt you can check for collided diagnostics also what will it give you it will tell you the multi-collinity problem. So, why let us to go ask for change also, so continue I would want the save anything so I just want to run it.

(Refer Slide Time: 17:42)
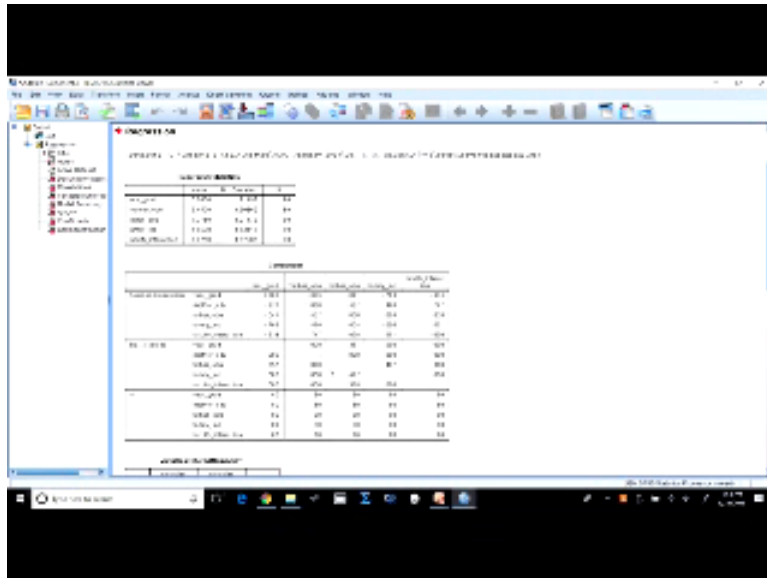
So, this is the same correlation think that you have seen I had shown you. Now let us go down this is the descriptive statistics look at the years graduation of the graduating time is 5 the mean is the 5 years standard device this much the education of mother is around 5.45, 1.5 so this is the sum of the descriptive statistics.

(Refer Slide Time: 18:06)
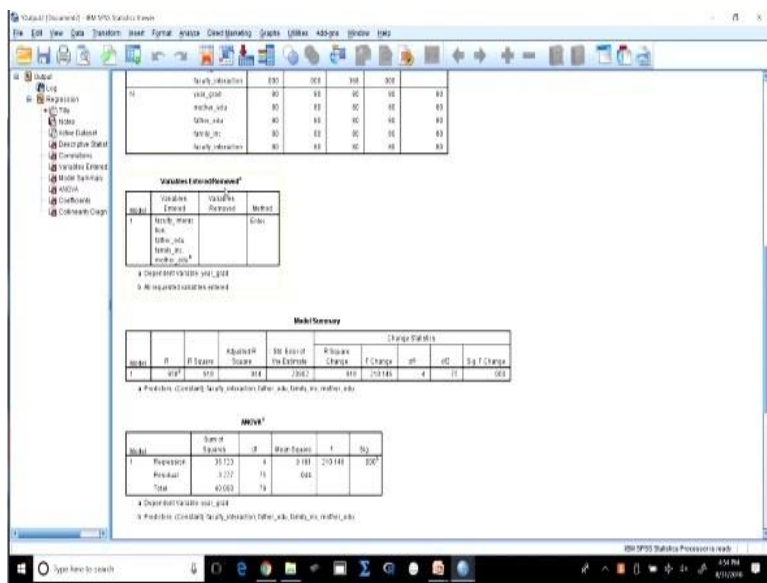
Now you see as I said the difference between a forward selection method and a stepwise regression method was what? That in the forward selection method you do not omit or remove any variable all are entered. But in the stepwise you may remove the variable after you take the combined effect and see which one is not now contributing ok. So, let us see all my dependent variable is time to graduate and my all other variables are entered. So, I did it only a stepwise enter I would have directly enter all them all of them into one so but I am interested to do it stepwise so I should have done this.
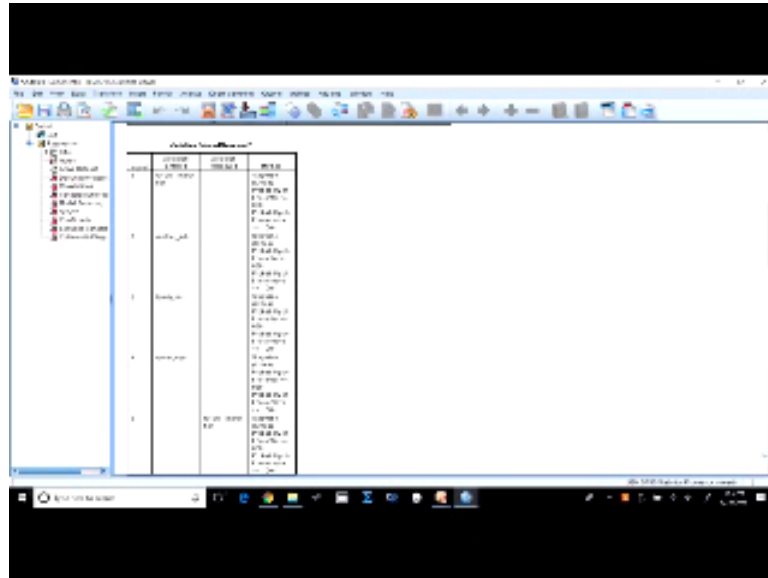
So, this will not change right let us go down so you see it was not showing anything but here you see what is says first faculty interaction was taken. So, you can do it the way from the correlation also you can see that the highest relationship was between years of graduation and faculty interaction, so you can do that second mother's education, third family income, fourth father education and this was the way the data was entered.

So, now model was created like this so the first case faculty interaction was taken, second case faculty interaction and mother education was taken, third faculty interaction mothers education and family income was taken, fourth all the four were taken but in the fifth case faculty interaction was removed why? Because faculty interaction is because of the presence of this now in the combined effect is coming down but if you remember this was the first model this was first variable that to was entered that means it was the most powerful case but still it is removed.

Now let us look at the R square values so when I am using in the first model case you see R square this much and R square change will remain the same. So look at the change in first model to second model 788, 696, 788 improved, 875 4th model 918, and 5th one 916 so the differences is hardly in difference and this is non-significant. So, the stepwise model this is the problem although it as told you some information that have given you.

The problem is that the variable which was the highest predictor or the best predictor, now itself has been removed from the model the final model. So, this is what the problem with the stepwise regression.

(Refer Slide Time: 20:43)

So you see this is what I want to show you. So look at this model, so, this is the coefficients, this is the t values, so faculty interaction the t -13, then you can look at it as it is improving, so the significance value you see the fourth model come to the fourth you can see the faculty

interaction is becoming non significant. So that means it got no effect. So, in the fifth one it has been removed.

But when you removed it, what has happened? Has it done any good we cannot see, so but at least I know one thing that variable which was contributing my highest is no more a part of my final model, so how can I it be considered as a good model. So, that is the problem which is happening in the stepwise regression, now what I will do is, will go back to the PPT.

(Refer Slide Time: 22:41)

So this is how you do, so this is how, so this was the table which would I have come from the same table I have copied here and pasted.

(Refer Slide Time: 21:45)

So, Interpretation, the inspection of correlation between the variables in the correlation table show that mother's education, parent's income, faculty interaction are all highly correlated. But father's education is only slightly correlated .04 something it was there. Also most of the predictor variables are correlated with each other with one correlation coefficient as high as .747. You can see .747, so which is .4747 faculty interaction with mother education this one.

But if you look at the years of graduation, so this is high this is very poor father's and this is also high and this is highest. But ultimately in your final model it has been removed and this is how the final model looks like.

(Refer Slide Time: 22:29)

Table 3

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.834[a]	.696	.692	.39501	.696	178.356	1	78	.000
2	.888[b]	.788	.783	.33157	.093	33.705	1	77	.000
3	.935[c]	.875	.870	.25669	.086	52.475	1	76	.000
4	.958[d]	.918	.914	.20902	.043	39.622	1	75	.000
5	.957[e]	.916	.913	.20967	-.002	1.479	1	75	.228

a. Predictors: (Constant), faculty_interaction

b. Predictors: (Constant), faculty_interaction, mother_edu

c. Predictors: (Constant), faculty_interaction, mother_edu, family_inc

d. Predictors: (Constant), faculty_interaction, mother_edu, family_inc, father_edu

e. Predictors: (Constant), mother_edu, family_inc, father_edu

f. Dependent Variable: year_grad

And interestingly you this is what I was saying is, so in the last model, in the last the e 1, 2, 3, 4, 5 a, b, c, d, e last same. Here this faculty interaction variable now as is not present here, it has been removed. But it was my highest predictor best predictor earlier.

(Refer Slide Time: 22:49)

So this is what it says variables removed.

(Refer Slide Time: 22:52)

Results and interpretation (continued...)

- the predictor variable that has the highest R with the criterion variable, faculty interaction (fac_int), is the first variable entered into the analysis. (Table 1)

- However, the final model of the analysis (model 5/e) does not include the faculty interaction variable. (Table 3) and it has become insignificant

- Thus, stepwise regression results in a bad model that does not include the predictor variable that has the highest correlation with the criterion variable.

The predictor variable that is the highest R with the criterion variable dependent variable faculty interaction is the first variable entered, however the final model does not include and has becoming significant. Thus stepwise regression results in a bad model or awkward model. You can say not bad I would not say awkward model that does not include the predictor variable that has the highest correlation. So, sometimes it is can create a trouble.

But this problem can be eradicated by using a hierarchical multiple regression method. Now what is the difference between this 2, I will just maybe described because I am running short of time and I will run it in next class.

(Refer Slide Time: 23:33)

Heirarchical regression

- Similar to stepwise regression, hierarchical regression is a sequential process involving the entry of predictor variables into the analysis in steps but the order of variable entry into the analysis is based on theory. The researcher chooses the order to enter the variables based on theory and past research.
- Hierarchical regression is an appropriate tool for analysis when variance on a criterion variable is being explained by predictor variables that are correlated with each other.
- Hierarchical regression is a popular method used to analyze the effect of a predictor variable after controlling for other variables. This "control" is achieved by calculating the change in the adjusted R^2 at each step of the analysis

So, similar to the stepwise regression the hierarchical regression is also a sequential process involving the entry of predictor variables in steps. But the order of variable entry is based on theory. See this is the difference, earlier you have doing on bases of some the variance explained or the correlation or something, right the t value. But here the computer is not doing anything, here you are thinking which should be the logically the one which should be taken together. So, the researcher chooses the order to enter the variables based on the theory and past research.

This is the basic different between stepwise and hierarchical regression. So, hierarchical regression is an appropriate tool for analysis, when variance on a criterion variable is being explained by predictor variables that are correlated with each other. Suppose there are variables which are correlated with each other. For example x1, x2, x3, x2 and x3 are correlated.

So then you have to consider that while doing the analysis. It is a very popular method used to analyze the effect of a predictor variable after controlling for the other variables, so how do we control I will show you. This control is nothing but it is achieved by calculating the change in the adjusted R square at each step of the analysis. So, at each step we will measure the R square and see whether there is an improvement or no improvement.

(Refer Slide time: 25:02)

Hierarchical (Sequential)

- Researcher defines order of entry for the variables, based on theory.
- Sets are IVs are entered in blocks or stages.
- R^2 change - additional variance in Y explained at each stage of the regression.
- F test of R^2 change.
- May enter 'nuisance' variables first to 'control' for them, then test 'purer' effect of next block of important variables.
- Manual technique & commonly used.

So, the researcher defines order of entry for the variables based on theory. Sets are independent variables are entered in blocks, so maybe a block can be only a one variable also or could be a combination of 2 to 3 variables which theoretically it say that go together and R square change method. May enter sometimes nuisance variables first to control for them, then test the purer effect of next block of important variables. Use your; it is manually you can use your logic, so theory is always most important thing here.

(Refer Slide Time: 25:38)

Hierarchical Multiple Regression

- Hierarchical regression is a way to show if variables of your interest explain a statistically significant amount of variance in your Dependent Variable (DV) after accounting for all other variables.
- This is a framework for model comparison rather than a statistical method.
- In this framework, you build several regression models by adding variables to a previous model at each step.
- In many cases, our interest is to determine whether newly added variables show a significant improvement in R^2 (the proportion of explained variance in DV by the model).

So, it is a way to show if variables of your interest explain a statistically significant amount of the variance in your dependent variable after accounting for all other variables. So, it takes to you an account how much of variances has been explained. This is the framework of model comparison now this is very important term you should understand. It helps you not only statistically test but tells you which one is the best model.

In this framework you build several regression models by adding variables to a previous model at each step, same like stepwise only here up to this much, but only difference is that we are using the logic here theory. In many cases, our interest is to determine whether newly added variables show a significant improvement in R square.

So, here I would tell you something before I crossover stop it now this improvement in R square which I am thinking of is there you must hard of this word R square and adjusted R square, so I will try to explain the difference between these 2, so what is the R square? I do not get a space, how do I measure R square? R square is equal to what now 1 - sum of square of the error variable divided by sum of square of total, correct.

So, when in a regression equation let say $Y = a + b_1 X_1 + b_2 X_2$ and I go on increasing, as I increase my number of independent variable up to X_n, see some are the other correlation will be there, so as you improve the increase the number of variables the R square will go on increasing. Because it is not possible that there is completely no relation in social science it is next to impossible. But that is the very dangerous thing.

So, if I am using that means if it is contributing .0001 still it is been taken and I had been added as a new variable R square increasing. So R square can sometimes be a very confusing variable. So, instead of doing that we use another which is called adjusted R square which you must of seen in all the result section. So what is the formula for adjusted R square let me tell you $1 - (1 - R \text{ square}) * (n-1) / n - p - 1$.

Now you see, now what is this p? The p is the number of the independent predictors or independent variables, so number of predictors, so as I said to you if I am increasing a variable suppose there are earlier 5 variables, now I am using let us the 6 variables, so what will happen, so automatically my R square will tend to increase. So if my R square is increasing so 1 - R square this value and $* n - 1 / n - p - 1$.

So, what is happen one side my R square is increasing but the other side since this is the denominator if it is increasing this will reduce the entire value. So, the adjusted R square gets somewhere normalized or adjusted with this one side increase of the R square and that increase in the denominator which pulls down the entire value. So, in the process if the value is not contributing much this becomes more dominating.

And as it becomes more dominating obviously what happen's the R square value does not increase rather it may start falling if the new variable is not contributing significantly to the study. So, therefore it is always advised to use the adjusted R square instead of R square doing any research study, well what I will do is, I will carry on this session in the next lecture where I will explain hierarchical regression little bit more and I will show you how to conducted and the SPSS, ok thank you very much.