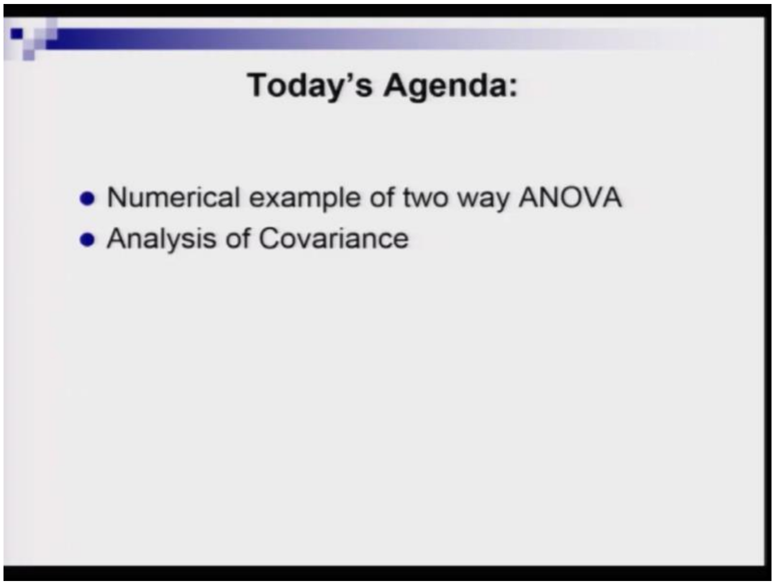
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-20
Analysis of Covariance

Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. And today, we are going to finish our discussion on analysis of variance. So, let us look at today's agenda items.

(Refer Slide Time: 00:29)

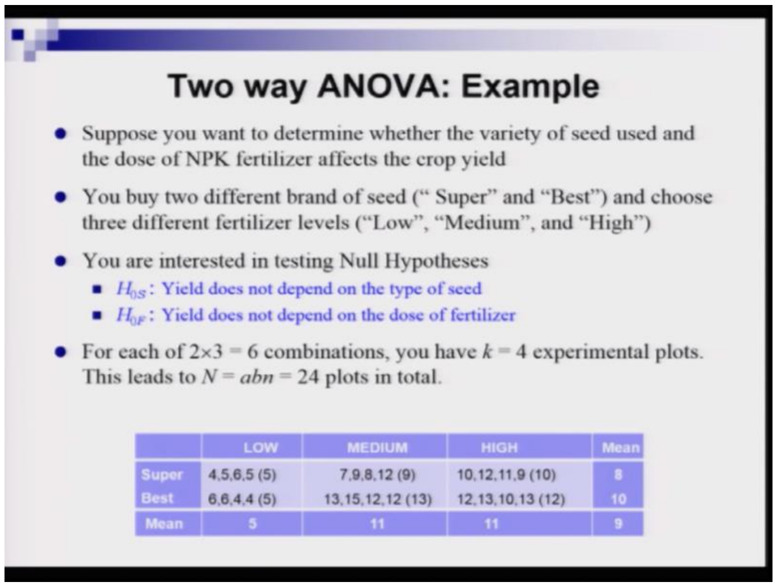


So, as I promised you last time, I will show you a numerical example of two-way ANOVA and I am going to focus on the calculation of sum of squares, because that is the most important and complicated part of ANOVA, be it one-way or two-way, and rest just by following some simple steps. Now, I am going to end today's lecture with a very brief discussion on analysis of covariance. I am going to show you a simplest possible ANCOVA model and with that our discussion on analysis of variance will end.

So, let us go back to the example from last class. So, we have discussed a simple two cross two factorial design model in the context of two-way ANOVA, and that was the example of an agricultural experiment. So, let us assume that you are an agricultural scientist or there is an agricultural scientist who wants to study the impact of variety of seed and those of fertilizer on crop yield and here is the experimental setup.

So, the scientist knows that there are two types of seed varieties available for a particular crop and there are three possible doses that are permissible to cultivate that crop. And this is basically the case of two factors. So, one factor seed has two varieties and we can name them something. And then fertilizer dosage has three levels, so that is another factor and it has got three levels, we can call them low medium and high.

(Refer Slide Time: 01:10)



So, let us assume that you buy two different brand or variety of seeds and their names are super and best. And then as I told you that there are 3 different fertilizer levels to choose from, namely low, medium and high. So, there are 2 factors here, one is seed and it has got two levels, and they are these the other factor which is fertilizer and it has got three levels. So, how many potential combinations are possible? 2 times 3 equals to 6.

So, now, let us frame our ANOVA exercise. So, we are interested in testing some null hypothesis. So, here there are two factors. So, of course, we want to see whether these different factors or groups have impact on the population mean or not. So, in other words, whether they impact crop yield or not. So, let us start with two null hypothesis, one for seed factor and one for the fertilizer factor. But note that the type of the null hypothesis are exactly the same, it does not matter which factor you are choosing for.

So, here, let us focus on the seed factor because that is the first one I have written in this slide. So, here the null hypothesis could be framed as, yield does not depend on the type of the seed. And alternative could be that, yes, yield does depend on the type of seed and mean crop yield for different types of seed, here are two types of seeds are actually different.

So, now, let us come back to the experiment again. So, I have already explained you that there are 6 possible combinations of levels from two factors, and let us now assume that the agricultural scientist or you have decided to have some number of plots and there you are going to basically apply a particular combination out of these 6 combinations and rest are going to be keep that same level. So basically, if you are applying these combinations to a half acre plot, then for all experimental plots, the land size or the cultivated land area will be half acres only.

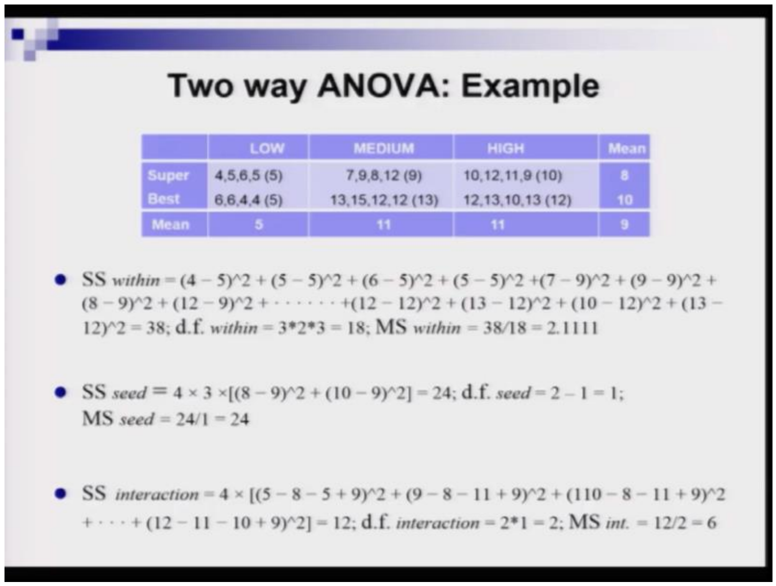
And then you are going to apply the same number of irrigation, you are going to apply same amount of monitoring and labour input and other inputs. And we are also assuming that as these experimental plots are going to be located very nearby, so agro climatic conditions are also going to be more or less similar for these experimental plots. Now, with this experimental setup, we got some crop yields from our experimental plots, so these are basically our sample observations. So, let us now see what we can do with this crop yields data from experimental plots.

So, if we have decided four experimental plots per combination, then we are going to have a total of 24 plots. So, here you see A is the number of categories for factor A which is seed, so that is two. So, B is basically the number of categories for factor B which is fertilizer here, and that is 3 in our example. And N is basically 4 that we have decided. So here it is a typo, so please note that. So, k will be N, or you can replace in that formula abn, N by k.

So, in total, we have 24 plots. Now, we have collected data on crop yield in some units and they are reported in the table that you see at the bottom of the slide. So, here, you see, for each cell I have observed 4 crop yields. So, here, suppose, let us focus on the cell super and low. Here, I find there are four cropping numbers, 4, 5, 6, and 5, so the cell mean is 5. So, you can get the cell mean by calculating the arithmetic mean of these four reported numbers for each cell. And note that you can also calculate the marginal mean for any of these levels of a particular factor. And finally, you can also calculate the grand mean.

So, here, I am going to show you some final calculation, some outcomes of final calculations. And I expect you to go back to the previous lectures and then match these numbers with the formula that I have shown previously. And then you can get the better understanding how we derive at a particular number.

(Refer Slide Time: 07:34)



So, next slide, we are also going to continue with the same table. But here in this slide, I am going to show you the same table, but here in this slide I am going to show you the sum of squares calculation, because that is the most complicated part of ANOVA. And we have not discussed these with a numerical example.

So, here, first, I am going to show you the case of the sum of squares within groups. And you see that here I am showing you the working of the formula, of course, due to limitation of space I could not show you all the components of that complicated sum. But I am showing you enough so that you can get hint how to compute sum of squares within.

So, after you match the numbers with the formula from the previous lecture, you see that you can calculate the value of SS within to be 38. Now, how to find the degrees of freedom? Again, in the last lecture I have shown you the formula. So, by applying that formula, the degrees of freedom for within groups will be 18. And if you divide sum of squares within by degrees of freedom within, you get the mean square within. So, that is basically 38 by 18, and that is equal to 2.11.

Now, we move to one factor. We have two factors, seed and fertilizer, so let us focus on the seed only. I am showing you the steps for seed, and I hope that you can calculate this quantity for the fertilizer factor easily by following the steps, because the steps are same. Now the sum of square for seed, here I am showing you the expression, it is a simple formula compared to the SS within. So, if you take that formula from previous lectures and apply it on the numbers that you see in the table, you get the sum of squares for seed factor equals to 24. And in this case, degrees of freedom will be 1 only. So, the mean square for the seed factor will be 24.

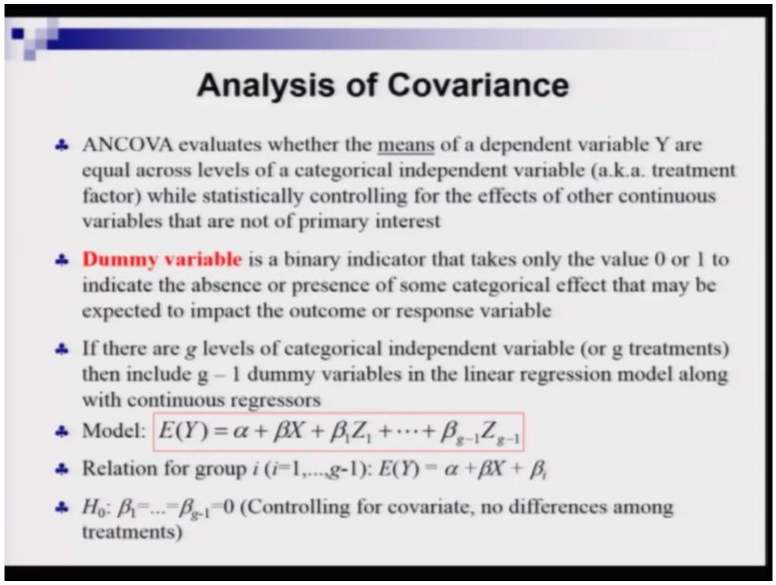
Now, we move on to another interesting sum of square component in the ANOVA identity, and that is the sum of square for interaction. And, again, I am showing you some parts of the formula. And it is not possible to show all the components of the sum, because it's quite lengthy. If you divide SS by degrees of freedom, you are going to get the mean square for the interaction term to be equal to 6.

So, we are going to now start our discussion on analysis of covariance. So, before I go to slides, let me talk about ANCOVA in a very simple language. So far, we have done what, we started with one dependent variable or outcome or response variable, whatever you want to call it, and let us denote that by Y. And then we were statistically testing what impact some qualitative factors will have on this continuous dependent variable Y.

And if we start with one qualitative variable or categorical variable or factor, whatever you want to call it, then we are going to have one-way ANOVA, and if you have two such qualitative variables or factors, then we will have two-way ANOVA. But in reality, there are many other variables which are of continuous in nature that can also have impact on Y. So, if you are truly interested to measure the group differences or the interaction effects between two qualitative variables or levels of qualitative variables, then you also have to control for those confounding factors or the covariates or you can also call them nuisance variables.

Now, in reality, it is difficult to collect the data on all possible nuisance variables or confounding factors, but certainly you may be able to get at least data on few variables. So, a complete model will not only have some categorical explanatory variables, it shall have some continuous explanatory variables as well. So, in that case, how do you proceed?

(Refer Slide Time: 12:28)



So, ANCOVA or analysis of covariance actually helps us to solve these kinds of problems involving one continuous explanatory variable together with categorical explanatory variables. So, if you want a definition then we can express it in simple language by saying that ANCOVA evaluates whether the means of a dependent variable Y are equal across levels of a categorical independent variable, while statistically controlling for the effects of other continuous variables which are not of primary interest to the researcher.

So, to have ANCOVA in operation, we need to learn about one new concept and that is the concept of dummy variable. So, a dummy variable is a binary indicator that takes only the value 0 or 1 to indicate or represent the absence or presence of some categorical effect that may be expected to impact the outcome or response variable.

So, if there are g levels of categorical independent variables or sometimes in textbooks, they are also called g treatments, then you have to include the g minus 1 dummy variables in the linear regression model, and that too with the continuous regressors. So, here, I am showing you our model, which is basically the simplest possible ANCOVA model. So, here I am assuming that we have only one continuous explanatory variable x to incorporate in our model. So, here, I am writing the mean regressions or what is mean regression.

So, mean regression is basically the expected relationship between the dependent variable Y and all other regressors. So, you can say that it is basically the population relation. As it is a population relation, there is no random noise or stochastic disturbance term associated with it. So, here, you note that, as I was saying that if there are g levels of categorical independent variable, we have to throw g minus 1 dummy variables In the regression model, and that is what we are doing in this model that you see in the red box. So, you see, I have added Z 1, Z 2 to Z g minus 1 dummy variables, and the corresponding coefficients are beta 1, beta 2 to beta g minus 1.

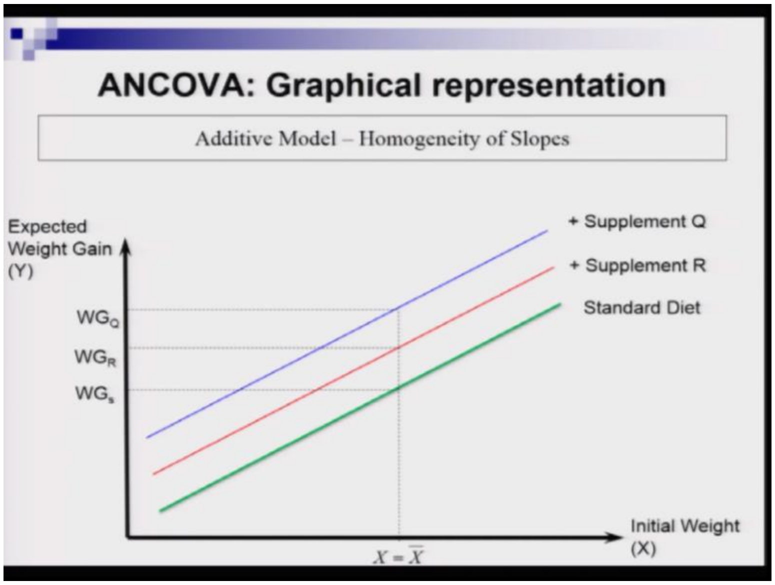
So, from this regression model, we can comment on the relation for group i. So, there are now g minus one such groups, because we have these g minus 1 dummy variables in the regression model. So, for a particular group i, the regression equation will be written as expected value of Y equals to alpha plus beta times x, plus beta i.

Now, what happened to the other regression coefficients? So, note that if your observation is from one particular group i, and you know about that, then this particular observation does not belong to any other group or category. So, for all these dummy variables, the value taken will be 0. So, only for this particular dummy variable which is representing the group you have chosen, it implies the group i, only for that dummy variable the indicator value will be 1. So, only that explanatory variable and the corresponding or associated regression coefficient will stay in the regression equation, rest will fall from the regression, as the values of these dummy variables will be 0.

So, once we write our model and we see the relationship for different groups, now we have to figure out whether there are some differences among these groups or treatments or not. So, how to statistically test that? Here we can write a joint null hypothesis and we can statistically test whether jointly beta 1, beta 2 and all other groups specific coefficients, up to beta g minus 1 are equal to 0 or not.

Now, I am going to take you through two diagrams, where I am going to explain that in this ANCOVA setup, how do I present two different interesting cases, the first case is the modelling of the main effects in the absence of interactions, and the second is modelling of main effects in the presence of interactions.

(Refer Slide Time: 17:43)



So, here in this slide, I am going to present the first case, which is the case of additive model. And it says that there is homogeneity of slopes, so that means that there is no interaction. Anyway, so this will be much clearer as I start explaining the diagram. So, before we explain the diagram, let us again start with a story and if or understand the story, then it will be better for you because it will be helpful to understand what is happening in that diagram.

So, let us assume that there is a poultry farm and there is a manager, who gives food to all these birds and he monitors the gain in the weight, and there could be a staple diet, a standard diet that you have to give to the birds, but then some veterinary doctors and some other knowledgeable person are also saying that, if you give supplement to the birds, then the gain in weight will be much higher for these birds.

So, basically, the manager was told that he can mix some supplements with the standard diet and that will basically help the birds to grow faster. So, with this hope, he can now try to conduct an experiment, he can have a group of birds to which he can give only the standard diet, and there could be another group to which this manager can give a mixture of standard diet and some supplement, and then there could be another group where the birds are fed the standard diet plus other supplements.

So, if that is the kind of experimental setup, then we may be interested to know that is there any difference in the mean weight gain, because we are talking about three groups here, right. So, in this context, what could be your categorical variable and what could be your continuous variable, if you want to conduct ANCOCA analysis?

So here, you note that it is not only the treatment, which is standard diet or some supplement that will have impact on birds' weight gain, but the initial weight or age of the Bard will also play a big role, and that is beyond control of this manager. So, this is truly something exogenous the manager cannot control. So, basically, ideally speaking, you can add this age or the initial weight of the bird as an explanatory variable and that could be the continuous x that is required for ANCOVA analysis.

So, with this set up, now, let us go back to the diagram. So, now, here, along the x axis or the horizontal axis, I am measuring the values of this continuous regressor, which is the initial weight of bird. And then along the vertical axis, I have measured the dependent variable, which is again continuous, and that is the expected weight gain.

So, now, let us assume that there is some mean for the continuous regressor variable x, and that is x bar. And here, at the sample mean of the data, which is x equal to x bar, we are interested to measure the impact of different diets. So, these are basically three different treatments you can say. So, the first line at the bottom is basically the standard diet curve, so that is marked in green. So, you see that if no supplement is given to the birds, and they are only fed the standard diet, then there is a linear relationship expected between the initial weight of the bird and the expected weight gain.

Now, if you give birds a mixture of standard diet and some supplement, say R, then there is a shift in this positive relationship, and this is a parallel shift. And that shifted line is marked by colour red. So, that is basically, you can say, the relationship between expected weight gain Y and initial weight X when treatment number 2 is provided. So, that is basically the standard diet plus supplement R.

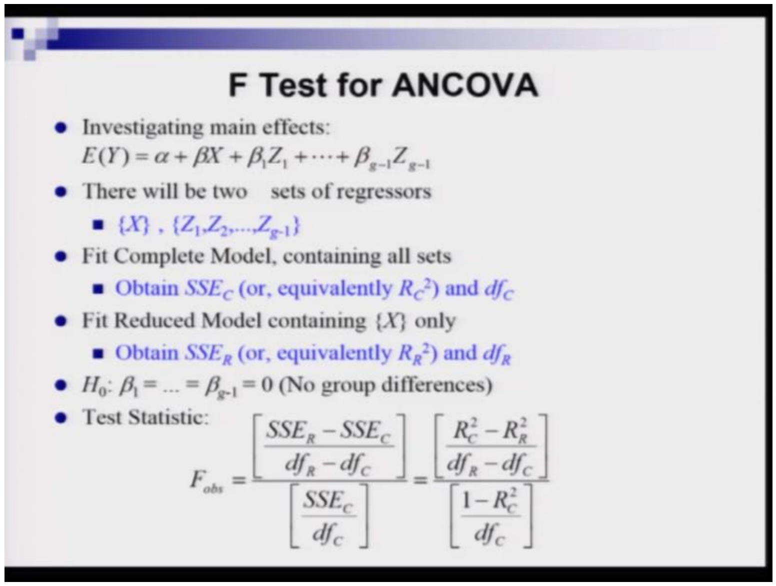
Now, there could be another treatment, treatment 3, where the standard diet is mixed with supplement Q. And here, you see that the curve has shifted up. These are all hypothetical shifts. I am not saying that supplement Q is better than supplement R, so this is a re completely hypothetical story. So, here you see that, again, the functional shift is parallel to each other, so the blue line is also parallel to the green line. So, you can say that at the sample mean of the data, for both supplements, the manager could see higher weight gains, and for supplement Q the weight gain is the maximum.

But, these treatment group 1, 2 and 3, they could also interact with the continuous variable. So, if there is an interaction, then how my story is going to change? So, that is going to be the subject matter of the next slide. So, here in this slide, I am going to show you the working of an interaction model. Here you see that my slopes are not uniform, so, they are not homogeneous in nature. So, the story goes like this.

So, in group 1, when the birds are fed with standard diet, then we see that green line is telling you the linear relationship between the Y and X's. So, in group 2, when birds are fed with supplement R, plus the standard diet, then we see that there is no change in the intercept of the green straight line that holds for the standard diet case, but there is a change in the slope and the slope is higher.

Now, we move to the third category or population group where birds are fed with supplement Q and standard diet. So, the supplement Q has not only impacted the slope of the curve or the relationship between Y and X, it has also impacted the intercept of the equation. So, you see that for population group 3, there is a change in both slope and the intercept parameter. But, of course, you can see that at the sample mean X equal to X bar, in treatment groups 2 and 3 you see the weight gain is higher.

(Refer Slide Time: 25:07)



Now, its time to look at the ANCOVA through some symbols. And I am going to now very quickly take you through the formal modelling and corresponding hypothesis testing for two cases, which are basically homogeneous slope and the heterogeneous slope. So, I will first start with the case of investigating main effects.

So, here I am going to assume that interactions do not exist. So, in that case, my population regression model will be expectation of Y or E of Y equals to the intercept parameter alpha plus beta times X, the continuous regressor, plus g minus 1 dummy variables, with their corresponding coefficients.

Now, there will be two sets of regressors, one is basically continuous X, then the other one will be a set of indicator, a dummy variable, g minus 1 number of them. And then you feed the complete model and then you opt in the sum of squares error or within sum of square, and that basically can be replaced by the R square also.

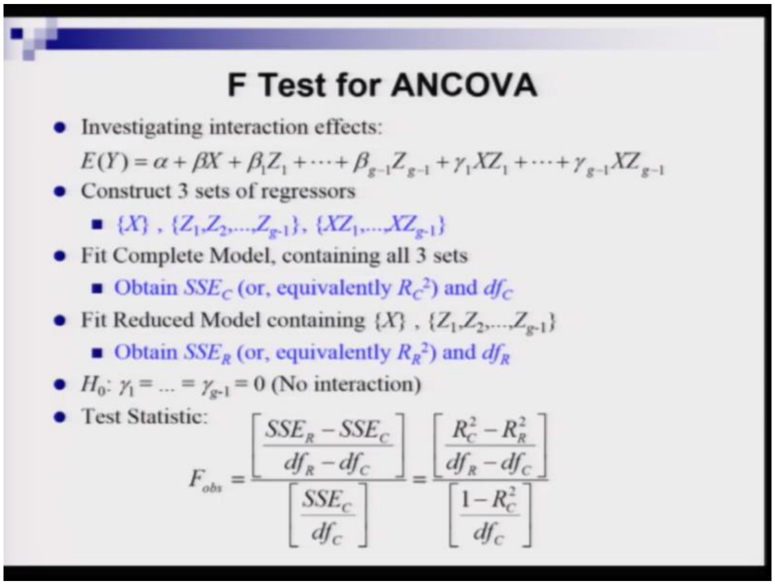
So, what is this big R square? If you remember our discussion on the regression analysis, you see that there was one component or one entity called coefficient of determination. So here in this case, we are computing that same coefficient of determination, but in the case of multiple regressor, so you can call this multiple coefficient of determination. So, later in the econometrics component of the course also I am going to discuss about this big R square.

And of course, the sum of square number will come with degrees of freedom. So, that is denoted by DFC. And then in the next step, we are going to feed the reduced model where we keep only the continuous regressor. So, in that case also we can generate the sum of square error or some square within, and the R square from the regression model and these two numbers will have degrees of freedom R.

Now, note that we are going to test a null hypothesis where we are assuming that there is no group difference. So, that means that beta 1 equals to beta 2 equals to beta g minus 1, equals to zero. So, that means that all together, jointly, all the regression coefficients for all g minus 1 dummy variables are jointly 0.

So now, we have to talk about the test statistic. So, here we have the F test statistic, and you will see there is a complicated formula. You need to remember this and you need to calculate this from the regression. And then of course, you will know how to conduct an F test, we have discussed about that several times in this course. So, you just have to compare the critical value of A from the F table with this calculated test statistic value and then you have to decide whether you want to reject the null hypothesis or not.

(Refer Slide Time: 28:50)



Now, we are going to talk about the case of interactions. So, note that if we assume that there is interaction effect, if we start with that assumption, then my regression equation becomes complicated, and then not only the continuous regressor X and g dummy variables will be part of my regression equation, but there will be interaction terms. So, the interaction between X and dummy variables Z 1, interaction of the continuous regressor X and the last of the dummy variables least Z minus, g minus 1, one all these will be there. So, there will be g minus 1 interaction variables also add the corresponding coefficients are denoted as gamma 1 to gamma g minus 1.

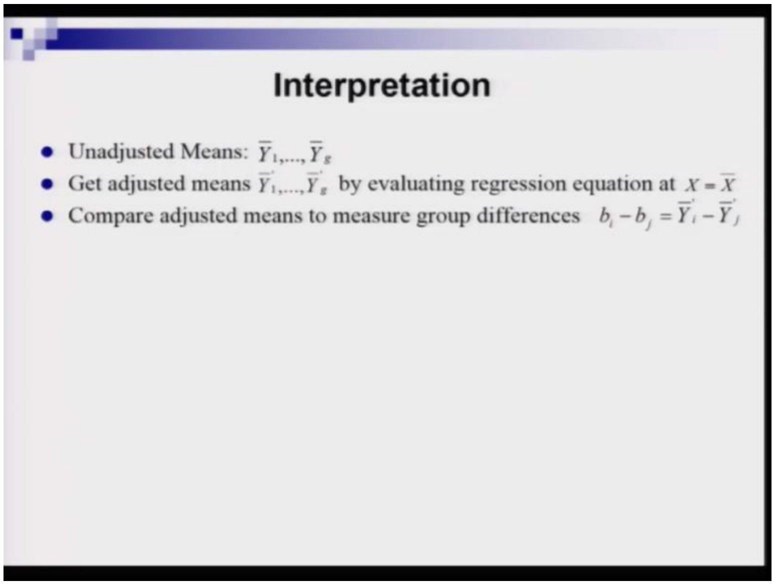
So, there will be three sets of regressors, X, the continuous regressor then g minus 1, dummy variables; and then again g minus 1, interaction variables, interaction between the continuous regressor x and the dummy variable Z. Now, you fit the complete model, so we are going to follow the exactly same steps as we did last time.

So, here we are going to feed the complete model and then we are going to get the sum of square error or sum of square within, or equivalently you can also work with the R square. And then you need to feed the reduced model and you just note down your sum of square error or within R square or R square, whatever you choose. And then finally, you have to frame your null hypothesis.

What would be the null hypothesis? So, you are going to test whether jointly the coefficients of these interaction variables are equal to 0 or not. So, that means, whether gamma 1 equals to gamma 2 equals to dot dot dot, gama g minus 1, equals to 0 or not? What could be the alternative hypothesis? The alternative is that at least one of these gammas is not equal to 0. Now, you have to talk about the test statistic. So, test statistic will remain the same, if you compare the test statistic formula with the previous slide, they are exactly the same. And in the same way you conduct F test.

So, what that ANCOVA is a very special kind of statistical technique which blends two different statistical techniques, namely analysis of variance and linear regression analysis. And ANCOVA, actually that is why is very useful when you conduct the dummy variable regressions, later we will revisit this case of dummy variable regression in the course and there you will see that these ANCOVA type of statistical analysis becomes very handy when you want to conduct hypothesis testing.

(Refer Slide Time: 31:48)



So, let us assume that the unadjusted means can be represented or denoted for G categories or groups as Y bar 1, dot dot dot, Y bar g. So, these are basically the sample mean of Y when these Y observations or raw observations on variable Y are bucketed or placed in g categories or g levels for the qualitative variable.

Let me talk about one more interesting thing. So, this is related to the interpretation. So, after you are done with your ANCOVA and F test and all, then what do you learn from all this? Suppose you find that your ANOVA is that, well, there is difference in groups and all so you reject your null hypothesis that there is no group difference or there is no interaction effect. So, then from there how do you actually interpret from the coefficients that you get for your dummy variables and all. So, that is what we are going to briefly discuss.

Now, you can also obtain the adjusted means, which are to be denoted by Y bar dash dot dot dot Y bar g dash, and that is basically by evaluating the regression equation at X equal to X bar. That is easy, so you take a fixed value of continuous regressor and that is basically the sample mean and then you assume that, now I am going to see the impact of my group 1.

So, what are you going to do here? So, suppose you are interested to find the impact of group 1 or treatment group 1. So, what will you do? So, you will assume that there is an observation which belongs to that particular group, so you are going to assume that the dummy variable associated with that group, which is Z1, will take value 1 and all other dummy variables in your regression will take values 0. So, then you can get the change in the mean Y, because this particular observation belongs to the group 1. Similarly, you can conduct this analysis for group 2 and other groups, and this is the way you are going to get the adjusted means.

Once these adjusted means are computed for various groups, then you can actually measure the group differences by subtracting two such groups. So, this can be done by taking difference of Y bar i prime and Y bar prime j. So, basically you calculate these adjusted means, and if you do take difference between two such means, then that will talk about the group difference. So, we come to an end of our discussion on analysis of variance. So, in the next lecture we are going to start a new topic and that is time series data analysis. See you then. Thank you.