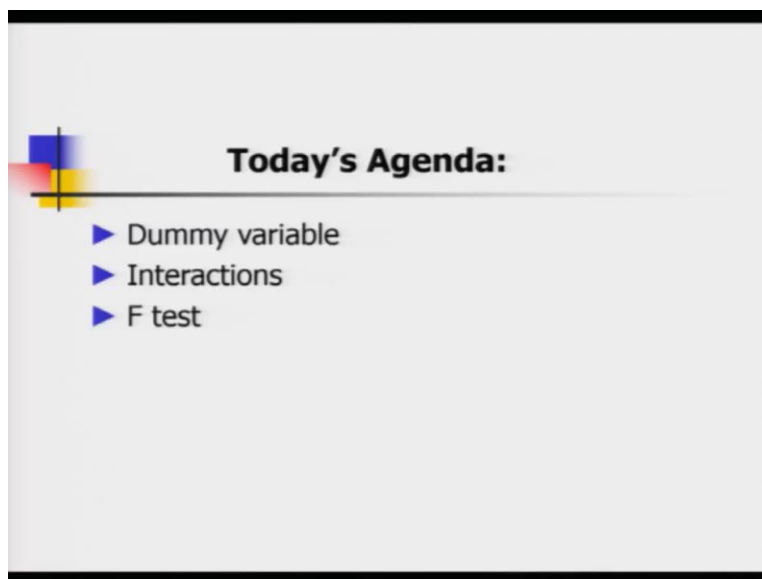


Applied Statistics and Econometrics
Professor. Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 34
Regression with Dummy Explanatory Variable

Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So, today we are going to start our discussion on an exciting area of econometrics, where we are going to deal with discrete variables. And in this week three consecutive lectures will be there, where we will be studying different cases of handling discrete variables in econometric analysis.

So first lecture in this, three lecture sequence is on Dummy Variables. And we have already seen dummy variables little bit in the regression context. In the ANOVA context also I have spoken about it briefly. But it is not a bad idea to devote an entire lecture on the dummies because dummies can do very interesting jobs for you, which we have not actually studied when we talked about the dummies in this course. So, this lecture is totally going to be devoted on dummy variables. Let us have a look at today's agenda items.

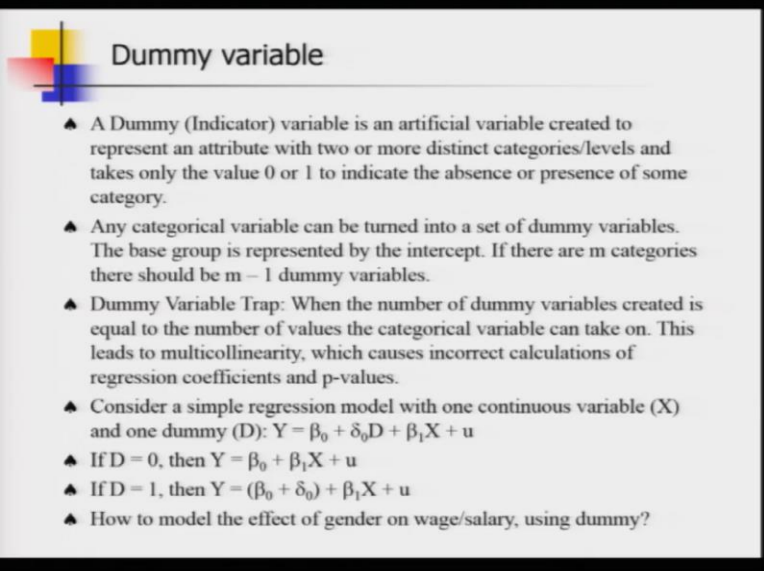
(Refer Slide Time: 01:21)



So, here we are going to first introduce the dummy variable once more in front of you with formal definition and all. Then I talk about interactions of dummy variables and then I am going

to show you how our good old friend F test is equally handy here also to deal with dummy variables.

(Refer Slide Time: 01:46)



Dummy variable

- ▲ A Dummy (Indicator) variable is an artificial variable created to represent an attribute with two or more distinct categories/levels and takes only the value 0 or 1 to indicate the absence or presence of some category.
- ▲ Any categorical variable can be turned into a set of dummy variables. The base group is represented by the intercept. If there are m categories there should be $m - 1$ dummy variables.
- ▲ Dummy Variable Trap: When the number of dummy variables created is equal to the number of values the categorical variable can take on. This leads to multicollinearity, which causes incorrect calculations of regression coefficients and p-values.
- ▲ Consider a simple regression model with one continuous variable (X) and one dummy (D): $Y = \beta_0 + \delta_0 D + \beta_1 X + u$
- ▲ If $D = 0$, then $Y = \beta_0 + \beta_1 X + u$
- ▲ If $D = 1$, then $Y = (\beta_0 + \delta_0) + \beta_1 X + u$
- ▲ How to model the effect of gender on wage/salary, using dummy?

So, let us start with a formal definition as usual. So, dummy or as it is also called an indicator variable is an artificial variable created to represent an attribute with two or more distinct categories or levels and it takes only values 0 and 1 to indicate the absence or presence of some category. So, basically you can see from this definition we are talking about couple of features of dummy variables. First of all it is representing a qualitative variable or attribute. And it is itself a binary or dichotomous variable taking two values 0 and 1.

Now you may ask can I call it, say 3 and 4 or 6 and 7. Yes, you can do it. Absolutely no problem. Here what we are doing in terms of dummy variable? We are basically assigning some number, some quantitative value to some qualitative features of an attribute variable. So, you can assign actually any value for that matter. But for simplicity, for easy handling in terms of interpretation of dummy variables generally in econometrics and statistics people prefer 0 and 1 coding.

So, now if you remember the first time we introduce dummy is in the context of time series analysis where we were trying to model seasonality. And therein we talked about something called dummy variable trap. So, what is dummy variable trap? So, before I go to the formal definition of dummy variable trap let me first set the rule and dummy variable trap can be presented as a violation of that rule. So, if the rule is broken by you then what will happen? That

is what basically the dummy variable trap. That is the consequences of violation of the rule that you must follow when you are dealing with dummy variables.

So, what is that rule? So, the rule says that if there is a qualitative variable which has some number of levels then, say the number of levels is m , so basically you cannot have more than $m - 1$ dummy variables in the linear regression equation if you also want to keep the intercept term. So, that is what is mentioned in the second bullet point. But if you are desperate to keep all the m levels in regression equation for some matter, then you have to exclude the intercept term from the regression equation. So, otherwise you are going to fall in the dummy variable trap.

Well, what is dummy variable trap? So, here in the third bullet I am giving you formal definition kind of thing for dummy variable trap. So, when the number of dummy variables created is equal to number of values of categorical variables can take on, so these are basically number of levels we are talking about then actually we are caught in a dummy variable trap. What happens exactly in this trap? Why it is called a bad thing?

So, basically if you are in the dummy variable trap, if you are violating this rule that I have said in the second bullet point then that leads to multicollinearity. Because one dummy variable can be expressed as a linear combination of the other dummy variable. So, there will be multicollinearity. And what will happen if there is multicollinearity? So, it will have the impact on the estimates of the coefficients, it will also have an impact on the p -values of the estimated coefficients. So, in a nutshell if you are caught in a dummy variable trap statistical inference is going to be misleading and that is why you should avoid it.

So, now let us represent dummy variable in linear regression context in terms of equations and symbols. So, let us make the story simple as usual. So, we can actually have the same model in terms of multiple X s or explanatory variables, but actually we are not doing that because we want to simply present the philosophy and the basic work of a dummy variable. For that one continuous regressor X is good enough and one dummy is also good enough. So, we are going to have a look at a very simple model.

So, here is the model. You have explanatory variable Y which is a continuous variable. Then you have an intercept term on β_0 and then D is the dummy variable and the associated regression coefficient is δ_0 . So, now we have one explanatory variable X

and the corresponding regression coefficient is B_1 . So, of course there is the stochastic disturbance or random error term u there.

Now you note that we can take two different values only and we are assuming as per the norm that they are going to take either 0 value or 1 value. So, if I now plug these two values in the regression equation in place of D , then I am going to get back 2 equations. One is basically for the base category. By the way I should also say that in defining dummies when you are representing a particular category or level of attribute or qualitative variable by 0 then you are calling that attribute level or the value of the qualitative variable as base category.

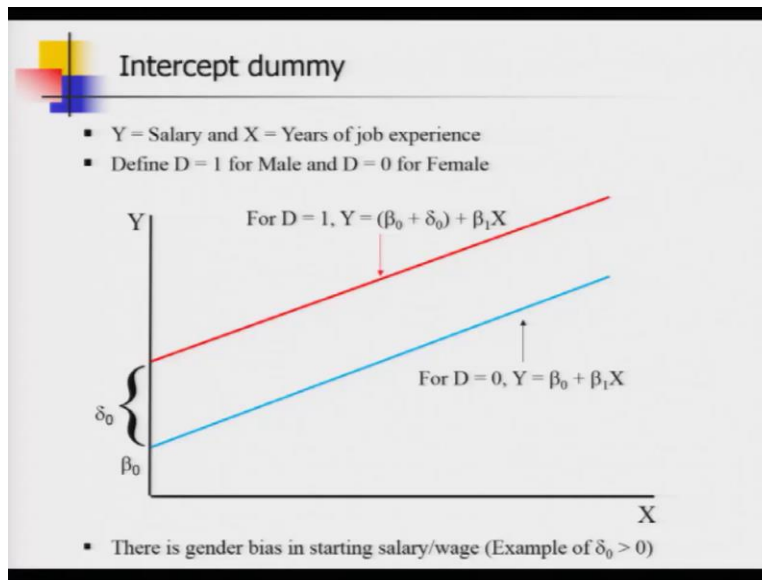
So, let us first have a look at the regression equation for the base category. So, for the base category dummy will take on value 0. So, you are much simplified equation which you are quite familiar with. So, you have made Y equals to β_0 plus β_1 times X plus u . What if I say that my dummy variable takes value 1 so that means there is presence of one particular attribute value or level for the qualitative variable and then actually you plug the value of D equals to 1 in the mother equation and then you get β_0 plus δ_0 plus β_1 times X plus u . Now you note one interesting thing. So, now you have got β_0 plus δ_0 as the common intercept term in this equation for D equal to 1 case.

Now with this dummy variable set up, it is not a bad idea to explore the extensions of dummy variable applications and we are going to fetch our application from the field of labor economics. So, in labor economics scholars are interested in many research questions and one of these research questions is discrimination in the labor market with respect to gender.

So, some labor economists are saying that there is substantial bias against the women or the female participants in the labor force and on an average the wage or salary for female worker in labor market it is much less compared to the male counterpart. So, this is kind of hypothesis that they are framing by looking at observations around them. Now hypothesis of course needs to be tested with help of real life data.

So, basically they conduct regression analysis to find answer to this query whether indeed in labor market there is some gender bias or not, whether there is discrimination in terms of wage or salary or not. So, we are going to look at this problem and we are going to make use of dummy variable to analyze this problem.

(Refer Slide Time: 10:25)



So, in this particular slide we have a very simple model. So, Y my explanatory variable is now salary or wage paid to a labor and X is the sole explanatory variable in this model which is years of job experience. So, of course it is a continuous variable. So, X equal to 0 means that the candidate or the participant in labor force has zero job experience. So, basically whatever he or she will get in terms of salary will be called starting salary, fine.

So, now we define dummy variable D which will take value 1 for male and it will take value 0 for female participant. So, now here is the diagram. So, here you see that along the horizontal axis I am measuring the years of job experience and along the vertical axis I am measuring the salary or wage obtained by the individual. And here you see I have two lines. And now I am going to explain these population regression lines one by one.

So, let us first assume that we are talking about male. So, for that the dummy variable takes value 1 and if you plug D equals to 1 into the equation that I have shown you in the previous slide then actually I get back the reduced regression equation saying Y equals to beta naught plus delta naught plus beta 1 times X. And that regression equation for the male group is shown by the red color upward-sloping straight line.

Now we concentrate on the relationship between years of job experience and the salary for the female group of labors. And there we have to then assume that our dummy variable D will take on value 0. So, if you plug D equals to 0 in that regression equation that I have shown you in the

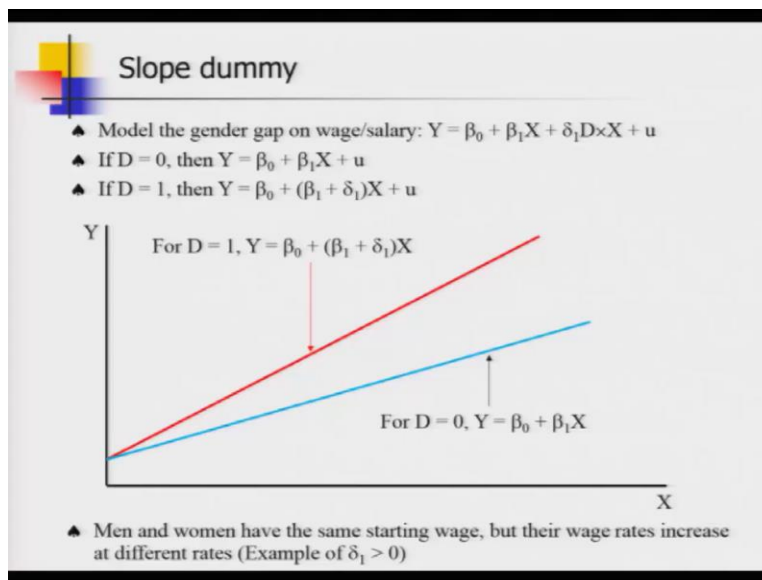
previous slide than you get another reduced equation. And you get to the simplest possible regression equation that is Y equals to β_0 plus β_1 times X . So, that upward-sloping straight line is represented by the light blue color, and of course here we are assuming that there is positive association between wage and salary and years of job experience. That is pretty normal to assume.

Now note that the way we have drawn these two lines mean something. And what is the gap between these two straight lines? So, the gap is basically in terms of the symbol δ . So, the δ is basically the regression parameter associated with the dummy variable D . So, here are the way I have drawn this diagram it implies that I am working with the positive δ . So, that is why you see that the blue straight line has an intercept of β_0 . And then I add δ on the top of β_0 and that will now become... So, that β_0 plus δ will now become the intercept for the regression line for the male group.

So, as these two lines are parallel, this gap δ is constant for all values of X . So, basically looking at this diagram what do we learn? So, we learn that there is some discrimination, there is some gender bias. So, it is not only the starting salary for group workers is higher than the female workers but this gap maintains for different values of X s. So, that is what we learn from the diagram.

But this is of course indeed a very simplified picture and there could be many more interesting things. The researcher can also be interested to study these questions, I mean, is there is a gender gap for salary and is that widening up as X increases. So, when you move up the ladder in corporate or any job place so of course you built on X , so that your number of years in job increases. So, is there any positive relationship between this gender bias and X . So, in other words I want to say is this gap widening up with increase in X ? So, how can we study this particular research question that is going to be the subject matter for the next slide.

(Refer Slide Time: 15:48)



So in this slide we are going to talk about slope dummy. So, the previous one we talked about intercept dummy. Here we have a new model but with the same old story. So, now look at the new model which says Y equal to β_0 plus β_1 times X . Now note that $\delta_1 D$ that disappeared. Now we have a new entity in place of that variable and that is $\delta_1 D$ times X . So, D times X is basically the interaction between the continuous variable X and the discrete dummy variable D . And δ_1 is basically the associated regression coefficient for the new variable that we have just created by interacting a dummy and a continuous regressor.

So, now as usual we can look at the reduced form equations by placing different values of D . So, if D equals to 0 then we get Y equals to β_0 plus β_1 times X plus u as usual. And if D equals to 1, so D equals to 1 means we are talking about the male group. So, we get reduced equation with the same intercept term β_0 but now we have a different slope coefficient. Then that is β_1 plus δ_1 .

So, again to graphically represent what is happening let us assume that δ_1 is positive. So, if we now assume δ_1 is positive what does that mean? So, it means that the value of the slope coefficient which is measuring the marginal effect of X the continuous regressor is actually higher for the male group compared to their female counterparts.

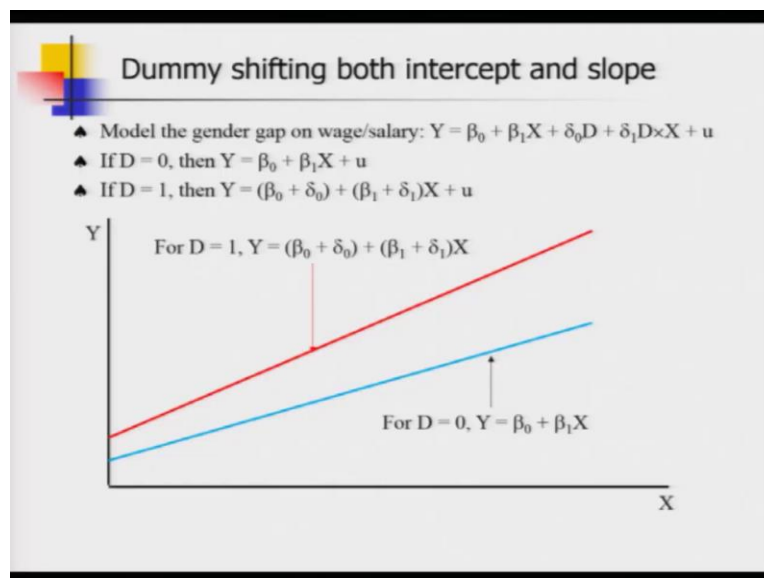
So, let us first talk about the female group. So, the regression equation, population regression line is represented by the light blue color as we did in the last slide also. So, here you see there is

one intercept β_0 which we are not of course showing here. Now we want move on to the male counterpart and the regression equation for them.

So, there you see the regression, population regression line is presented by red color and it starts on the same intercept β_0 but now as δ_1 is positive the slope coefficient $\beta_1 + \delta_1$ is greater than β_1 , of course. So, this upward-sloping straight line as it has higher slope, this gap between the red straight line and the light blue straight line is increasing with increase in X .

So, you can say that it may be the case that men and women have the same starting wage but their wage rates increase different rates. So, as number of years in job increases this gap in wage or salary due to this gender effect increases quite a bit. So, this is one way of looking at this problem. But there could be other type of interaction cases which could be seen as extension to this problem and that is what we are going to study next.

(Refer Slide Time: 19:24)



Now we are going to talk about the most general model involving one continuous regressor and one discrete dummy variable. So, here we are in the same setup. So, we are trying to explain the gender gap on wage and salary. So, now you have this mother equation which says that Y is equal to, Y of course is salary or wage. So, this is equal to β_0 plus β_1 times X plus δ_0 times D plus δ_1 times D times X plus u .

So, now you see that I brought back that intercept dummy from my model 1 and I have added that to the slope dummy model. So, I have both intercept and slope dummies working in same regression equation.

So, now as usual I want to look at the base category equation. So, that is basically for the female. And we have as usual the simple straight line $\beta_0 + \beta_1 X + u$. And then basically I have the regression equation, the new regression equation that is actually a reduced version of mother regression equation and this is for the male counterpart. So, for that the D will take value equals to 1.

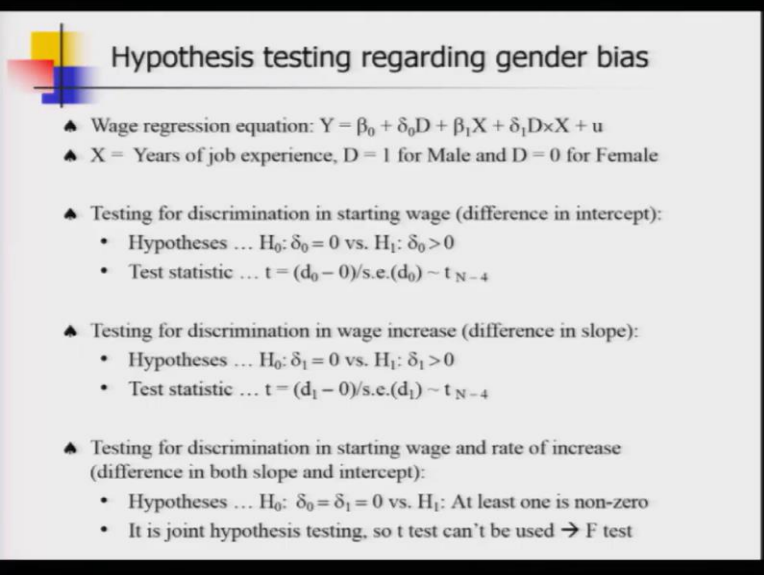
So, you plug D equal to 1 in the mother equation and you get an equation where you see now you have two components for the intercept parameter. And they are β_0 and δ_0 . And there are two components for the slope parameter which are β_1 and δ_1 .

So, here let us assume that δ_0 and δ_1 are both positive numbers and this is for graphical illustration of course. So, if I assume these two parameters δ_0 and δ_1 will take positive values only, then I actually I have this new diagram. So, as usual the light blue color straight line is depicting the relation between wage or salary and the number of years in job. And if you assume positive values for δ_0 and δ_1 you actually get a red straight line for the male counterpart in the labor market. And you see that not only the starting salary is higher for the male, the salary gap is also widening up as X increases.

So, this initial gap in the starting salary is measured by this coefficient δ_0 and the difference in the rate in increase is measured by this parameter δ_1 . So, we have studied this case of gender bias or wage discrimination in labor market through diagrams.

But diagram is diagram. So, you are assuming certain values δ_0 and δ_1 , actually not assuming certain values you are assuming that but they are both positive. But that is all hypothetical assumption to show you simplified cases in terms of the diagram. But whether in reality your data gives you enough sample evidence or not for assuming δ_0 and δ_1 are actually positive numbers, for that your to conduct hypothesis testing.

(Refer Slide Time: 23:18)



Hypothesis testing regarding gender bias

- ▲ Wage regression equation: $Y = \beta_0 + \delta_0 D + \beta_1 X + \delta_1 D \times X + u$
- ▲ X = Years of job experience, $D = 1$ for Male and $D = 0$ for Female
- ▲ Testing for discrimination in starting wage (difference in intercept):
 - Hypotheses ... $H_0: \delta_0 = 0$ vs. $H_1: \delta_0 > 0$
 - Test statistic ... $t = (d_0 - 0) / s.e.(d_0) - t_{N-4}$
- ▲ Testing for discrimination in wage increase (difference in slope):
 - Hypotheses ... $H_0: \delta_1 = 0$ vs. $H_1: \delta_1 > 0$
 - Test statistic ... $t = (d_1 - 0) / s.e.(d_1) - t_{N-4}$
- ▲ Testing for discrimination in starting wage and rate of increase (difference in both slope and intercept):
 - Hypotheses ... $H_0: \delta_0 = \delta_1 = 0$ vs. $H_1: \text{At least one is non-zero}$
 - It is joint hypothesis testing, so t test can't be used \rightarrow F test

So, now we are back to the slide and let us start with that wage regression equation, the most general case that we are studied in the last slide, so where we both have the slope dummy and the intercept dummies. So, here as usual X is number of years of job experience and D is the dummy variable where D equals to 1 means male and D equals to 0 means female.

So, now there could be three types of hypothesis testing problem given this setup. So, in the first stage we may be interested to go for testing whether there is a difference in intercept term or not. So, I am talking about now the first dummy variable diagram that I have shown you in this labor market story's context. So, how to set your hypothesis?

So, you set your null hypothesis as $\delta_0 = 0$ and that you challenge by placing an alternative hypothesis which says that δ_0 is actually positive. Note that here I am going for one tail test. I am not going for a two tail test. If I have written $\delta_0 \neq 0$ then actually that is the regular two tail test.

But here we are interested in challenging this view that δ_0 is actually not 0 and we claim that it is positive. So, that is why we are this time going for a one tail test by placing this inequality there in the alternative hypothesis. So, we are saying that in our alternative hypothesis says that this initial gap in the starting salary is indeed positive. It is not 0.

What could be the test statistic? As we are dealing with only one single parameter in the regression model we can make use of our friend t test. And then the test statistic is defined as $D_{naught} - 0$. So, what is D_{naught} ? So, suppose you run your regression via OLS method so the coefficient estimate for δ_{naught} is denoted by D_{naught} .

So, D_{naught} is basically the OLS estimator for δ_{naught} . And why we are deducting 0? Because if the null hypothesis is true then the unknown population parameter value is 0. And you remember that while conducting T test when we are framing the T statistic we have to take the difference between the sample statistic value and the hypothesized population parameter value. So, that is why we are subtracting 0 from D_{naught} .

So, this difference or D_{naught} has to be divided by the standard error of the estimated regression coefficient. So, we are going actually going to divide by standard error of D_{naught} . So, now this test statistic will follow a t distribution. Now t distribution comes with degrees of freedom. So, how many degrees of freedom we are expecting in this case?

So, note that here we have 4 parameters to be estimated, one intercept term and one each for the slope dummy and the intercept dummy and then there is one for the continuous variable. So, the degrees of freedom will be $N - 4$ where N is the number of observations of course. And I believe that you remember your lessons how to make decision rules and how to conclude a T test. So, I am skipping those discussions there.

And now I move on to the second case where I am interested to conduct a statistical testing regarding the difference in slope. So, basically here I set my null hypothesis as $\delta_1 = 0$ and my alternative hypothesis is $\delta_1 > 0$. I have explained in somewhat detail why we are not going for a two tail test in the first case. So, I am skipping that discussion and I believe that you can understand why I am placing strict inequality here in place of not equal to sign.

So, the test statistic will be the usual t-statistic, nothing new. So, you can follow it quite easily. And now I move on to the third case, where I am going to test for discrimination in starting wage and rate of increase. So, in other words I am interested to talk about testing the difference in both slope and intercept dummies. So, here the hypotheses are going to be somewhat different because here we are talking about two parameters at a time. So, of course this simple hypothesis testing will not work. We have to adopt the joint hypothesis testing.

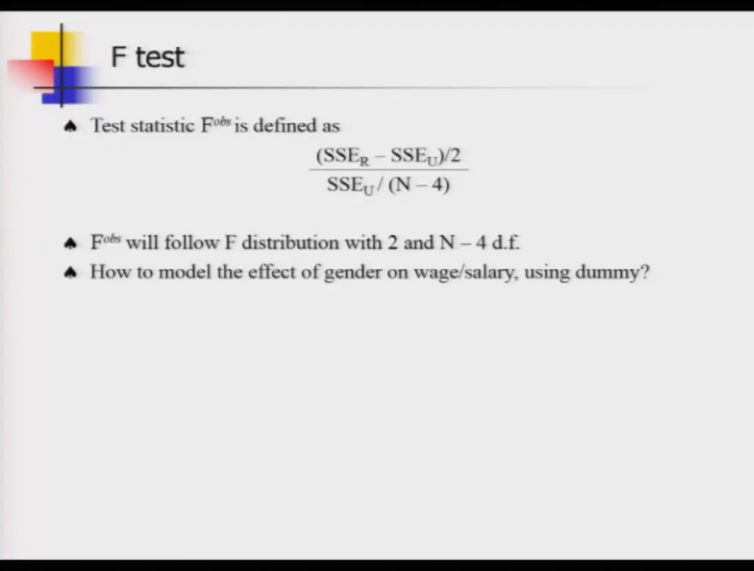
So, here in this context how to frame a joint hypothesis? So if there is no difference in the slope and the intercept then basically we are saying that both δ_0 and δ_1 they are equal to 0. So, you can set that as the null hypothesis. And alternative hypothesis we will say that at least one of them is non-zero. So, as this is a joint hypothesis testing we cannot use T test. So, we can go for F test.

So, you see F test comes back again to help us in joint hypothesis testing. So, how you are going to conduct an F test? I am going to talk about the steps in a short while but that is too mechanical. You have seen F test previously. So, you probably also remember the steps.

But I am going to spend a minute here by pointing towards the philosophy of this F test. Why are we applying F test here? And if you are applying F test, what is actually happening behind the curtain? So actually here you note down when I am writing that joint hypothesis testing problem, the null hypothesis the way I have written it, that is basically imposing exclusion on the mother regression equation.

And if you are imposing some kind of external exclusion restrictions then actually you are talking about the restricted versus unrestricted models. So, we already have seen that how F test is useful to resolve the debate between whether to go for a restricted model or whether to adopt an unrestricted model. So, that same old philosophy or rationale that we have used earlier that is coming back again to help us. So, in the next slide we are going to show you the F test which going to help us in this context of dummy variables. But that is going to follow the same philosophy of restricted versus unrestricted models.

(Refer Slide Time: 31:06)



F test

- ▲ Test statistic F^{obs} is defined as
$$\frac{(SSE_R - SSE_U)/2}{SSE_U / (N - 4)}$$
- ▲ F^{obs} will follow F distribution with 2 and $N - 4$ d.f.
- ▲ How to model the effect of gender on wage/salary, using dummy?

So, now in this slide we are going to talk about the test statistic which is following the case of restricted versus unrestricted model F test. So, here SSE of course you remember that is sum of square of error residuals. And R subscript stands for the restricted model and the U subscript stands for the unrestricted model. So, let me remind you once more about the calculation of this SSE metric. How do you calculate the SSE metric?

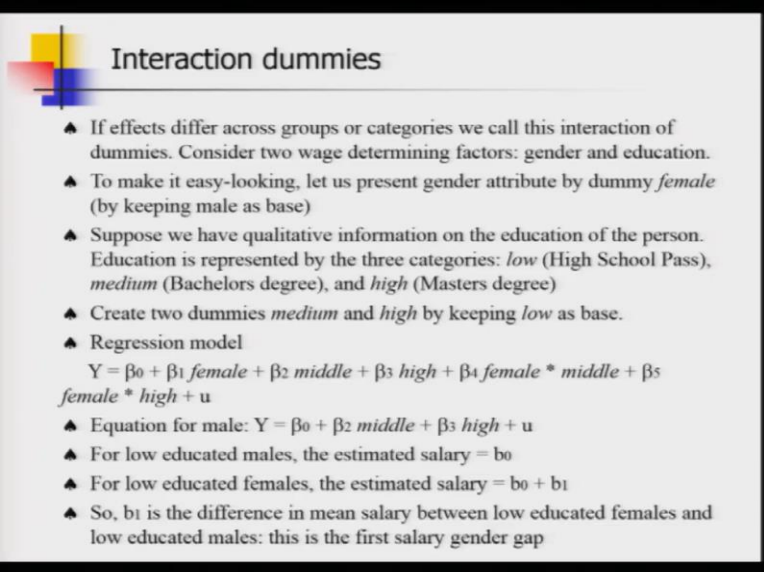
So, suppose we are talking about the unrestricted model. So, you estimate the unrestricted model with all 4 parameters and then you can actually have one regression equation which is the fitted line. Now you can generate the \hat{Y} values, the fitted values for Y by plugging different values of Xs, here only one X and only one dummy. So, you can plug the value of the dummy variable and the continuous explanatory variable in the fitted regression line to generate the fitted or predicted values of Y for all observations in your sample. So, if there are N data points in your sample you can generate N number of fitted values.

Now you take the difference between the actual value of Y and the fitted value of Y for all N individuals in the sample. Then you square them and then finally you sum all these squared residuals. So, this is the way you get sum of square errors from the unrestricted model. So, you can follow the same tactic to calculate the SSE R to calculate the square residuals from the restricted model.

So, now we know this difference between sum of square numbers between restricted and unrestricted model has to be divided by the degrees of freedom and here we are placing two exclusion criterion. So, that is why the degrees of freedom is 2. And the ratio has to be divided by sum of square error of unrestricted model. And that has to be divided by the degrees of freedom of the original or mother regression model so that is N minus 4.

So, of course needless to say if you remember this statistic will follow F distribution with two degrees of freedom 2 and N minus 4. So, this is the way basically you can model the effect of gender on wage or salary using dummy variable. So, basically I end this slide by reminding you about the question, original question with which we started our journey. So, that you can understand what have we are been doing for last 5 or 6 slides.

(Refer Slide Time: 34:06)



Interaction dummies

- ▲ If effects differ across groups or categories we call this interaction of dummies. Consider two wage determining factors: gender and education.
- ▲ To make it easy-looking, let us present gender attribute by dummy *female* (by keeping male as base)
- ▲ Suppose we have qualitative information on the education of the person. Education is represented by the three categories: *low* (High School Pass), *medium* (Bachelors degree), and *high* (Masters degree)
- ▲ Create two dummies *medium* and *high* by keeping *low* as base.
- ▲ Regression model

$$Y = \beta_0 + \beta_1 \text{female} + \beta_2 \text{middle} + \beta_3 \text{high} + \beta_4 \text{female} * \text{middle} + \beta_5 \text{female} * \text{high} + u$$
- ▲ Equation for male: $Y = \beta_0 + \beta_2 \text{middle} + \beta_3 \text{high} + u$
- ▲ For low educated males, the estimated salary = β_0
- ▲ For low educated females, the estimated salary = $\beta_0 + \beta_1$
- ▲ So, β_1 is the difference in mean salary between low educated females and low educated males: this is the first salary gender gap

But the story is not over. There could be further more complications in the labor market and how wages are determined. And we are going to talk about one such case here by introducing a new concept which is called interaction dummies. So, now I want you to think about that initial wage or salary regression equation with which we have started the discussion.

So, I said that there is one continuous regressor which is years of job experience and then there is one dummy talking about the gender qualitative variable. But that could be host of explanatory variables which actually are determining wages and salaries of individuals. So, basically let us

assume that education is of course one of the most important and relevant variables for wage or salary regression equation.

So, let us now introduce education in the story. And to make it more fitting for the topic of today which is dummy variables now I am going to assume that I am going to handle education as a qualitative variable. So, if you remember when we talked about education last time we measured education continuously by throwing a continuous variable in the regression and it was the years of schooling.

But now we are going to say that we are going to talk about levels of qualitative variable which is the degrees completed by an individual. So, now based upon the information we have that what degree this person has completed, we can actually talk about three different levels of education and they are low, medium and high.

So, there could be effects of gender and there could be effects of education on wage or salary. And primarily the way we have set up our research problem both of these explanatory variables are qualitative in nature. So, now as usual I assume that the gender qualitative variable has two levels, male or female. And to make it little bit easy-looking let us present the gender attribute by the dummy variable female.

So, basically male is the base category. So, dummy variable female will take value 1 if the worker is a female. Now I have already explained you couple of minutes before that we are dealing with education as a qualitative information and that qualitative variable is represented by three categories. And here I am showing you what do I mean by low, medium and high.

So, low is basically representing a person who has just passed the high school or maybe attended somewhat high school but high school dropout. And then medium says that there is a person with bachelor's degree. And then finally the high level denotes the person with master degree. So, here note that the qualitative variable education has 3 different categories or levels.

So, if we want to avoid the dummy variable trap then we have to define two dummy variables by keeping one base. So, here we can keep low as the base and throw two dummy variables medium and high. If we are interested to see whether with increase in educational qualification there is a positive impact on wage or salary or not, so then it is better to keep low as the base category

because then interpretation of the corresponding coefficients of these two dummy variables medium and high will be easy to handle.

So, now we have this regression model with all possible variables. So, Y equals to the overall intercept β_0 plus β_1 times female plus β_2 times middle plus β_3 times high plus β_4 times female cross middle. So, here is the interaction and β_5 finally is associated with another interaction variable female cross high. And finally of course there is this stochastic disturbance or end of error term u .

So, now as usual we are going first talk about the regression equation which is the simplest possible. So, basically we are going to talk about what is the regression equation for the base. So, as I said we are going to first look at the equation for the base. So, here is the equation for the male.

So, here you see that if we are talking about male worker then this female dummy will take value 0. So, β_1 times female will drop out from the equation. β_4 times female times middle that expression will also drop out from the equation. So, finally also β_5 times female times high will also drop out. So, we have a very simple equation which says that Y equals to β_0 plus β_2 middle plus β_3 high plus u .

So, now let us talk about low educated males versus low educated females. So, we are basically now fixing the level of education and we are going to see how gender is going to play a role in salary difference. So, I want you to be back to that equation again. And now let us assume that you have got data.

So, you have actually estimated this regression model involving all 5 explanatory variables. So, in total you have estimated 6 parameters from the regression model via OLS method. And let us assume that the estimated coefficients are denoted by English alphabet B . So, for β_0 the corresponding OLS estimate is B_0 .

So, with this now let us look at the story. So, here for low educated males what is the estimated salary. So, we have to plug the values for dummy variables. So, as we are talking about males we have to place the value equal to 0 for female and we are also talking about low education, so the values of these two dummy variables middle and high will also take 0 value. So, basically we

land up getting only the intercept term. So, for low educated males the estimated salary is basically the estimated value of the slope coefficient which is b_0 .

Now let us assume that we are talking about female. So, if we are talking about female then this female dummy will take value 1, but as this female worker is low educated the middle and high dummies in the regression equation will take value 0. So, even the interactions will be 0. So, everything will fall out of the regression equation and we will now have only 1 additional item that is basically the regression coefficient corresponding to the female dummy.

So, let us assume that OLS estimator for corresponding regression coefficient is b_1 . So, we are going to get this estimated salary for low educated female $b_0 + b_1$. So, b_1 is the difference in the mean salary between the low educated females and the low educated males. So, this is the first salary gender gap.

Now note that our regression model is pretty general in the sense that you can also talk about other types of gender gaps. So, now you can move on to a different level of education. So, you can move up to, say medium and then you can find the gender gap.

Then you can move on to, concentrate on the high education class only and then there also you can talk about the gender gap. So, you can talk about several gender gap numbers and this is in contrast to the first story where we had only one δ_0 which is basically the case of only intercept change in terms of gender. So, this model that I presented here, it is pretty general in nature and I leave this up to you to explore the other possible gender gap cases. So, wait for the next lecture. See you then, thank you.