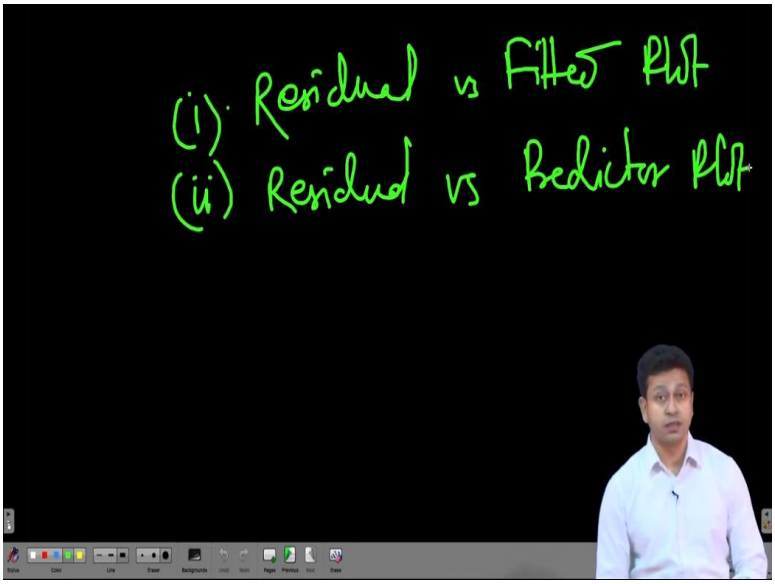
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 8
Lecture - 65
Heteroscedasticity (Contd.)

Hello and welcome back to the lecture on Applied Econometrics, and we have been talking about heteroscedasticity. Now, in the previous lecture we have introduced the concept and we have shown how graphically it looks like and what are the situations it may arise. Now, in this lecture, we are actually going to see, when we have data in our hand, how do we understand if the dataset really has a problem of heteroscedasticity?

(Refer Slide Time: 00:46)



And for that, we need to understand these two kind of plots that we will talk about. One is residual versus fitted plot and another is residual versus predictor plot. So, what is residual versus fitted plot? So, we have already seen that we can actually identify the presence of heteroscedasticity by plotting Y hat and error; so, the error term that you already have and the predicted Y, Y hat basically. So, that is what is essentially residual versus fitted.

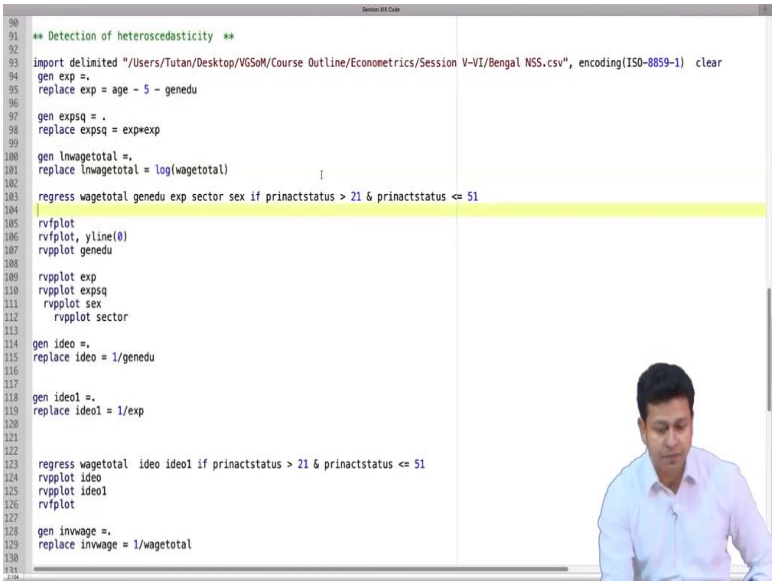
Fitted is nothing but the fitted Y. Once we actually conduct the regression, perform the regression, you get the residual versus fitted plot. And residual versus predictor is that, once you actually conduct the regression, once you already have the residual; basically, you get the

residual after you conduct the regression. So, whatever is left out as a residual, that is what you actually take and plot it.

And in this case, in the second case, you get the residual and you plot it with the X variable. So, whatever X variable is your concern, you may think that in your regression equation, it is like one particular X variable is actually influencing the error term. So, if you plot these two, then, what you will get is, you get a plot where it will show if there is any relationship. It is basically, simply looking like a scatter plot.

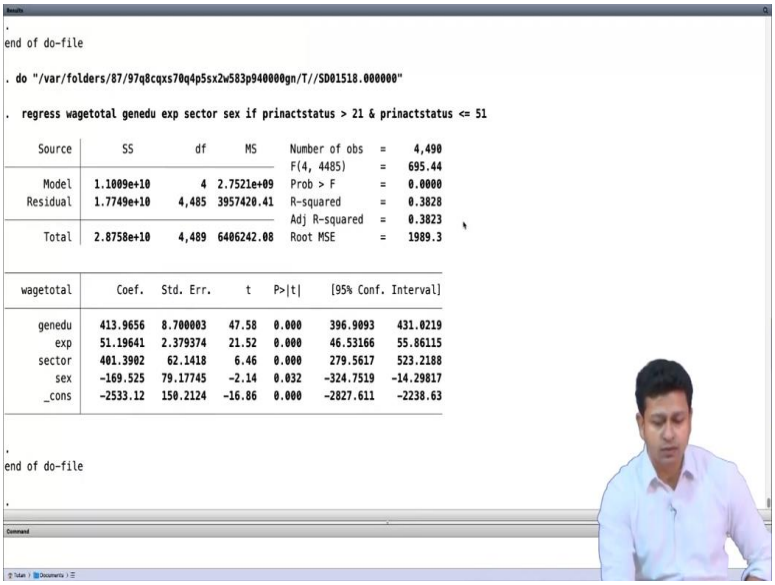
In a scatter plot, you simply see if you see some sort of association between the 2 variables. It is as simple as that. So, let us first try to actually run a regression line, a regression equation, and then we will see how the residual versus fitted is looking like or how the residual versus predictor plot is looking like.

(Refer Slide Time: 02:23)



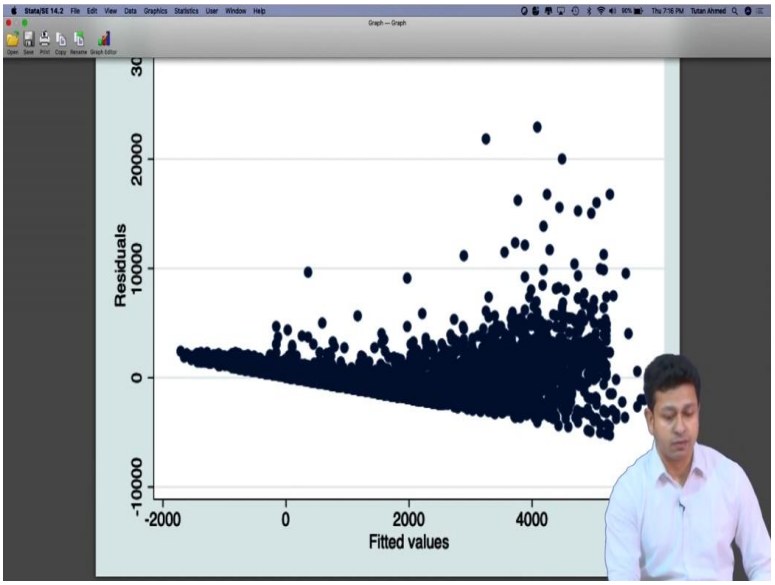
So, we will just go back to our data set that we have been using over the time. So, this is basically National Sample Survey data. And in this National Sample Survey data, I have already defined all these variables; I am not going to define it again. I will just run the regression line and let us see if the regression is running fine.

(Refer Slide Time: 02:47)



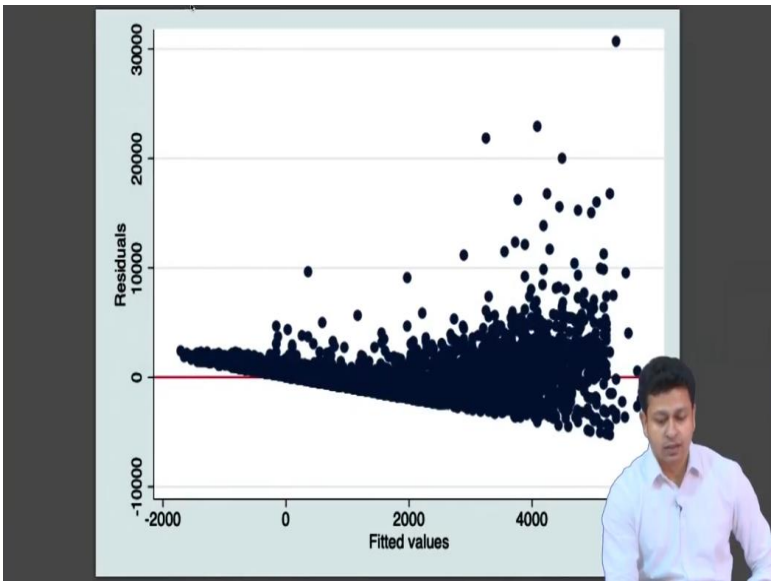
The regression is running fine. It is a reasonably good model with R square 0.38. Now, in this model, I want to see if there is any; what we can see if I do a residual versus fitted plot. So, this is residual versus fitted plot. The command is very simple in Stata, rvf, residual versus fitted, r for residual, versus fitted. And if I do rvfplot, let us see what we get here. So, it is basically giving us like a scatter diagram between your residual and the Y hat.

(Refer Slide Time: 03:28)



And if we do that, this is what we get. And you see that, in this case, what we see is, with the increasing value of fitted Y, your residual is actually also increasing. So, this is something we have seen as the usual case of heteroscedasticity, where with the increasing value of X or increasing value of Y, your error term is also increasing. So, this is one example, what we see that probably in our data set, we actually have the problem of heteroscedasticity. You can use this command of yline. So, what you will get is basically, you will get the 0 line.

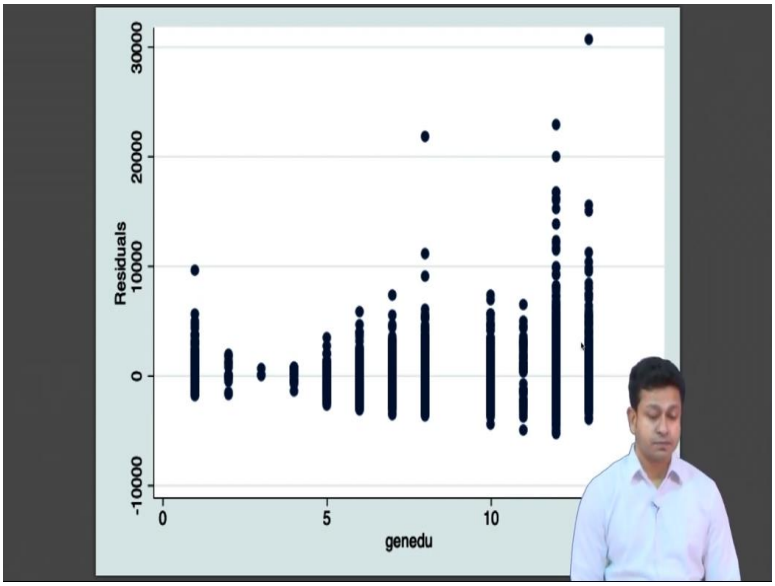
(Refer Slide Time: 04:03)



It will give a prominent sort of the 0 line, where the 0 line is there. Now, the other plot is rvpplot. Now, rvpplot is nothing but the residual versus predictor plot. Now, predictor means basically your X variables. In this case, you note that, if your model has like 3 X variables, you can actually plot your residual with all these 3 X variables. So, you have to decide, you have to make a sense of which variable might have some relationship with the error term.

Or you can actually plot all the different possibilities, like all the error terms with the different predictors. So, let us actually do that. And here, rvp, residual versus predictor plot; and I am assuming that, well, general education might have something to do with the heteroscedasticity problem. So, let us just try to run it.

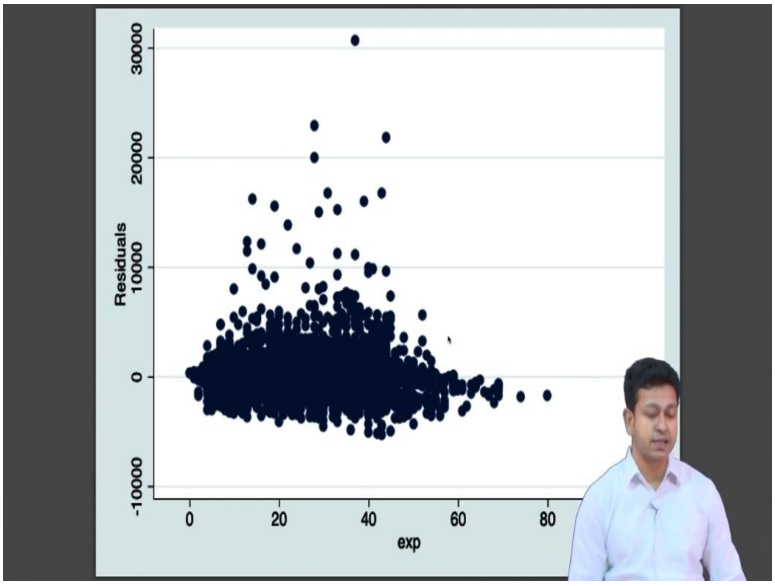
(Refer Slide Time: 05:36)



And I see, my idea was correct. So, we actually see that with different levels of general education; so, these are levels of education, so, like class 4, class 6, class 8 and different ways NSS actually collects data. And graduation level is probably the highest, the dispersion is very high. And probably after that, it is post-graduation. So, as you increase the education, the dispersion is also increasing.

So, that is something we have to; we can kind of say that, well, there is a problem, this particular X variable might have something to do with the heteroscedasticity problem. You can also do for other continuous variables or other numerical variable like experience, let us say.

(Refer Slide Time: 06:26)

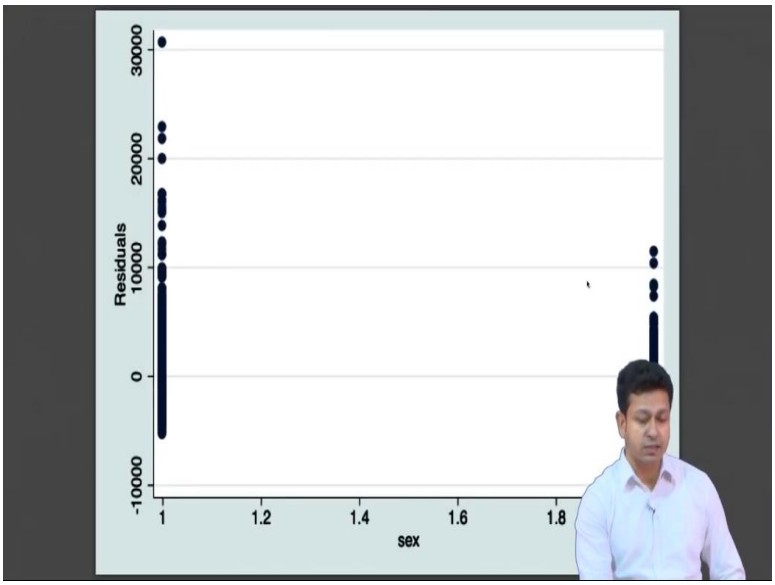


If I do with experience, the number of years, so, it is something also making sense. It is of course, not strictly like diverging. It might have some bell shaped sort of nature. And that is actually how the experience influences the wage. As your experience increases till a certain age, maybe 50, 55, 60, you actually keep on earning more; but then, after certain age, even if your experience is increasing, as per our definition of experience, you may not be earning that much; your income is actually decreasing.

So, that actually makes sense. So, essentially, it shows that the error term is actually correlated with the experience, and that is not absolutely a random; so, there is a correlation, it is not absolutely random. You can also plot with experience square. And let us say if we plot with the dummy variable; so, here, sex or gender is basically; we have 2 dummy, loop 2 values, 0 and 1, male and female.

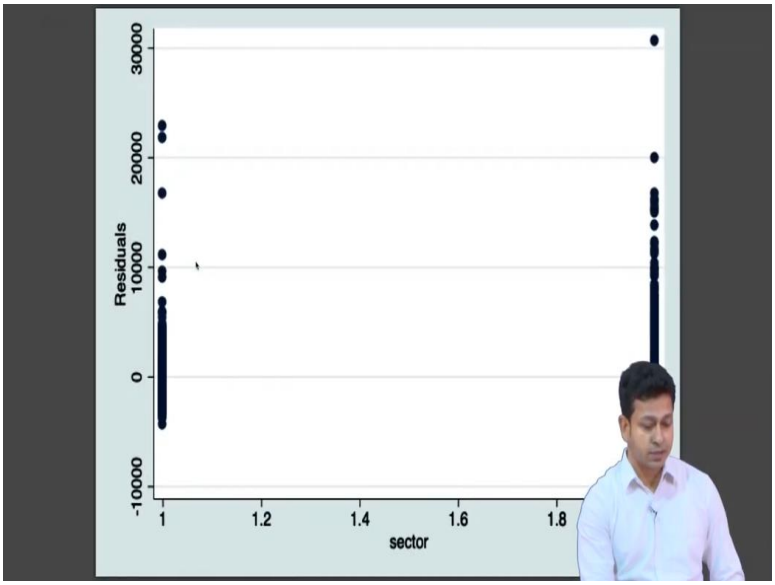
And for sector also, this is basically for rural and urban. So, we can expect that for females; well, so, since they might not have lot of opportunities, so, the dispersion or the observations are less in number; so, the dispersion might be low.

(Refer Slide Time: 07:39)



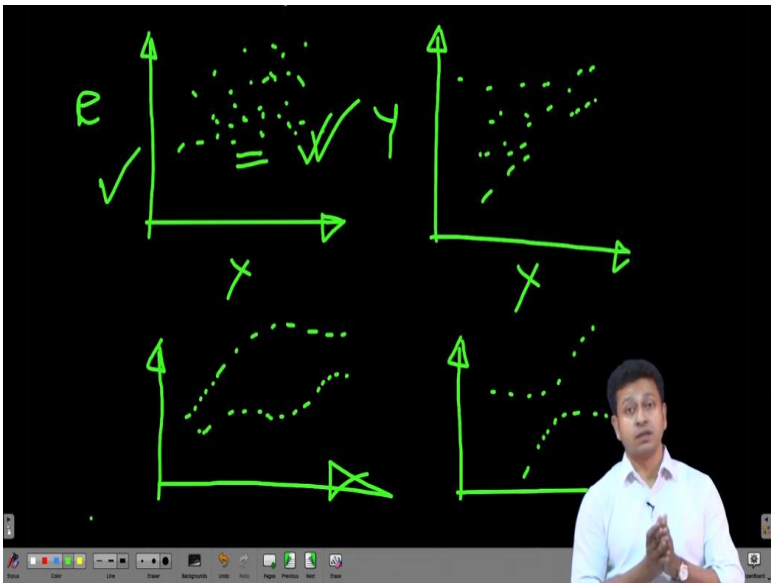
And exactly, that is what we see here. So, for males, the dispersion is very high, but whereas for females, the dispersion is actually low. And you can just pause the video for a moment and think what will happen if I plot for rural and urban. And if I plot it, we will see that for urban, because the opportunity has a different type, like the chances of getting into different occupations are like plenty; so, you are likely to have more dispersion in the urban area. Whereas, for rural area, since the opportunities are limited, the dispersion is also going to be low.

(Refer Slide Time: 08:15)



And that is exactly what we see. So, in our data set, 1 is the rural and 2 is the urban. And you see, the dispersion in urban area is much higher than the dispersion in the rural area. So, this is the beauty of rvpplot. Now, I can ask a question that, well, we have seen in all the cases, the heteroscedasticity looks like our case 1.

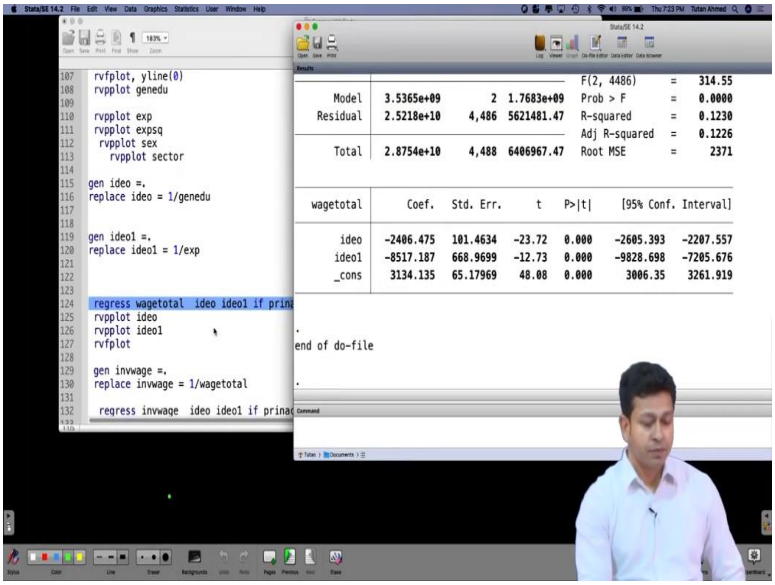
(Refer Slide Time: 08:46)



This, our case 1, where we said that usually the error term would show a sort of a diverging nature. But will it always be the case? And that is what we are going to see now. And I will claim that it will mostly depend on how you specify your regression equation, how the functions are, what are the variables. So, that is where the nature of the heteroskedasticity gets determined.

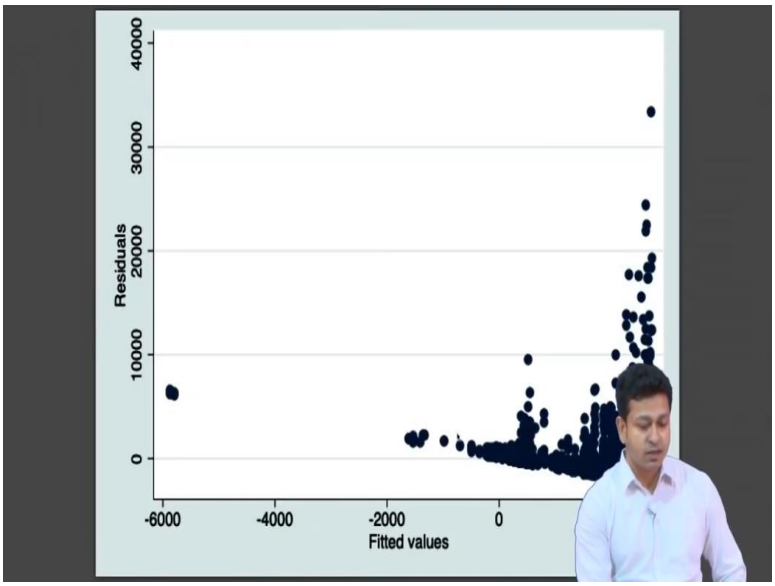
So, let us say, instead of general education, I create an idiosyncratic variable; I call it ideo equal to 1 by general education. So, whatever is the value of general education, it will be now 1 by that. And then I create another variable ideo1, and that is 1 by experience. And now, if I run my regression with ideo and ideo1; let me actually do that. So, I already have generated. So, I will not run it again; Stata will give me an error. So, if I run it with ideo and ideo1, just these 2 variables, let us see what we are going to get.

(Refer Slide Time: 10:00)



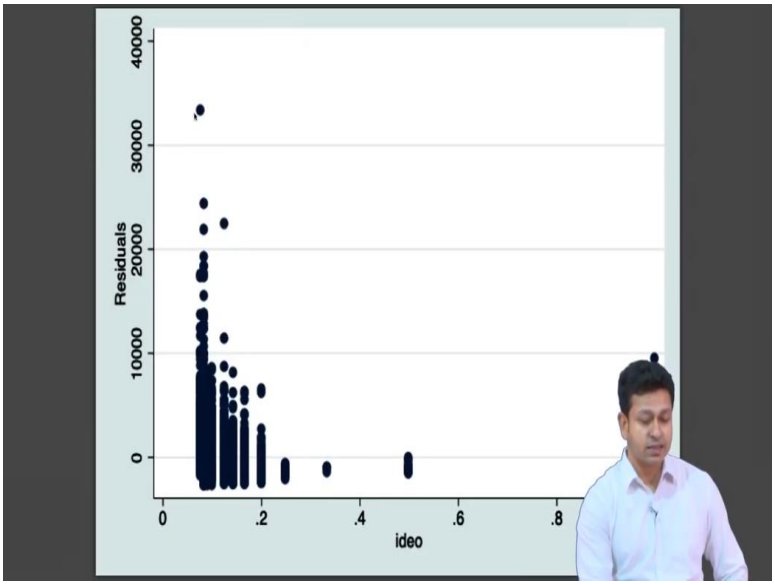
So, my regression has run fine. Now, make a guess; if I plot the rvfplot and if I plot the rvpplot, where exactly the differences will be actually visible? So, will it be any different if I do the rvfplot? You can pause for a moment and think about it. Actually, there will be no difference; because in rvfplot, you are plotting the fitted Y with the error term. So, there might be a little bit of change in the fitted value, but the nature of the curve is going to remain same, because the Y, you did not change any functional form of the Y. The Y remains Y; you did not do it log Y or you did not do 1 by Y.

(Refer Slide Time: 10:44)



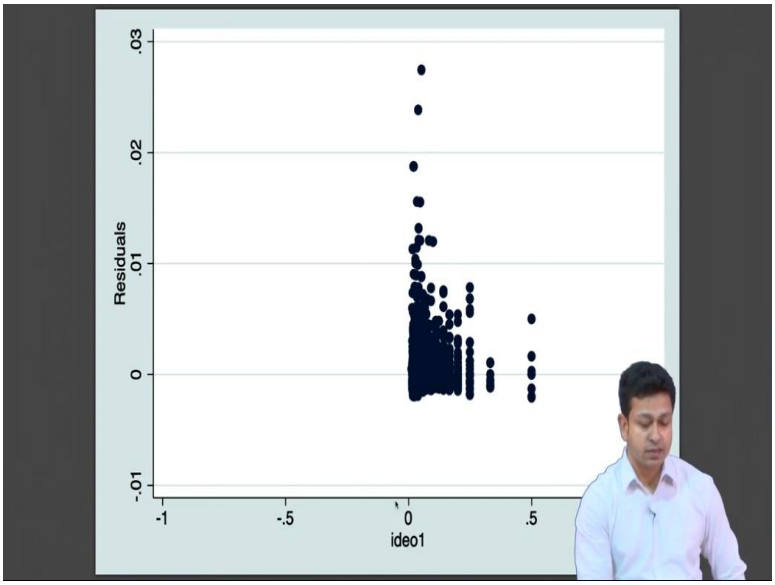
So, if I do the rvfplot, we will see, we will get the same residual and fitted value sort of relationship. But if I do rvpplot; because your explanatory variables have changed, the functional form has changed, you will see the relationship to be a little different.

(Refer Slide Time: 11:06)



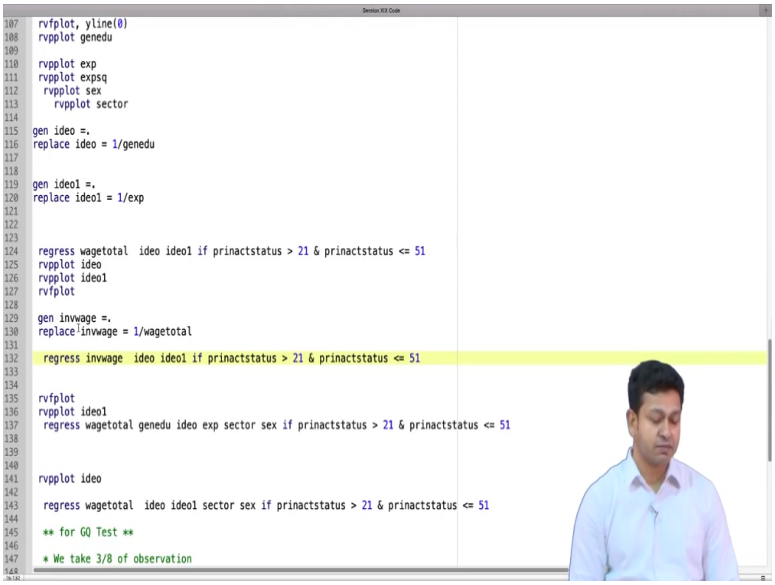
So, you see the dispersion is actually higher when your values are actually lower. And it is actually lower when your values are; and that is totally reasonable, because you have done 1 by general education instead of general education. And similarly, for experience, you will also see the same thing.

(Refer Slide Time: 11:28)



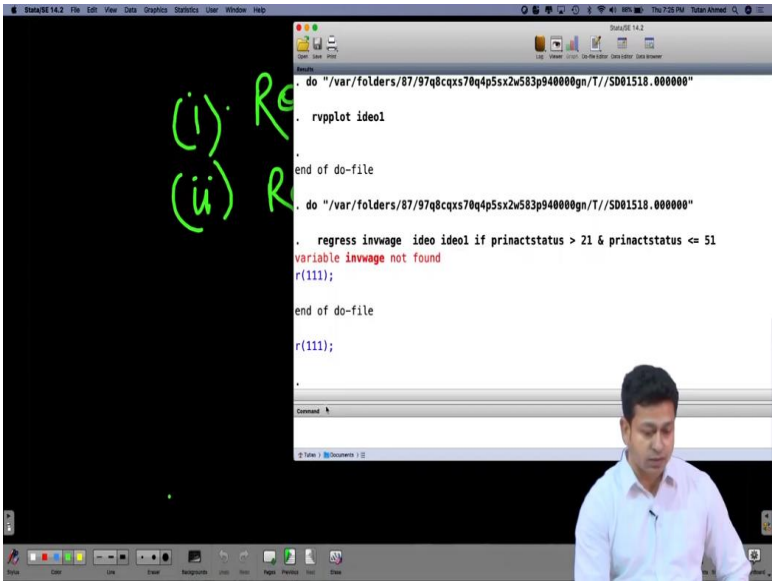
So, you will have the highest level of dispersion at the value of 0, when you are like 0 or close to 0, because again, here the new variable ideo1 is nothing but 1 by experience. So, as your experience value is increasing, your 1 by experience is actually decreasing. So, it is essentially showing the same thing. But this is how, just by changing the functional nature, you can actually have different types of heteroskedasticity plotted.

(Refer Slide Time: 11:59)



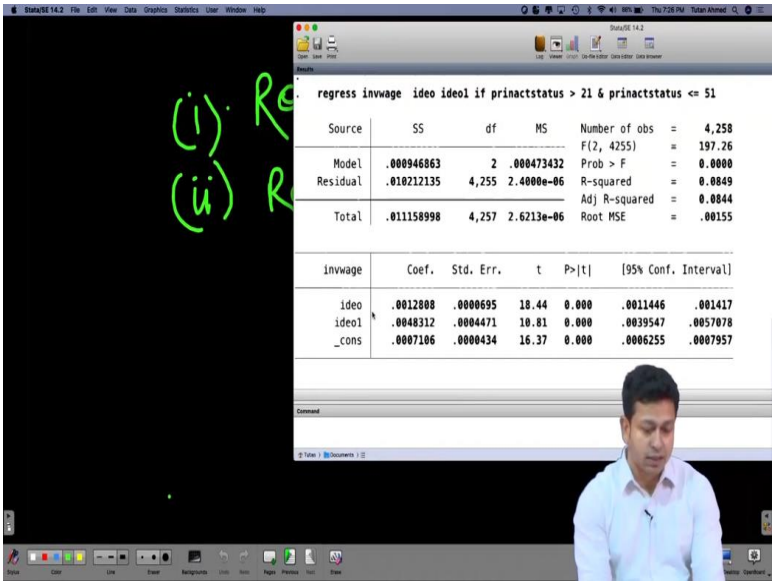
Now, if I suppose change the Y variable. Now, my Y variable, instead of log wage, I create a variable inverse of wage. My dependent variable is wagetotal, so, my new variable is going to be invwage, which is 1 by wagetotal. So, if I run the regression here with invwage as the Y variable, so, I will get the result; invwage is not found.

(Refer Slide Time: 12:34)



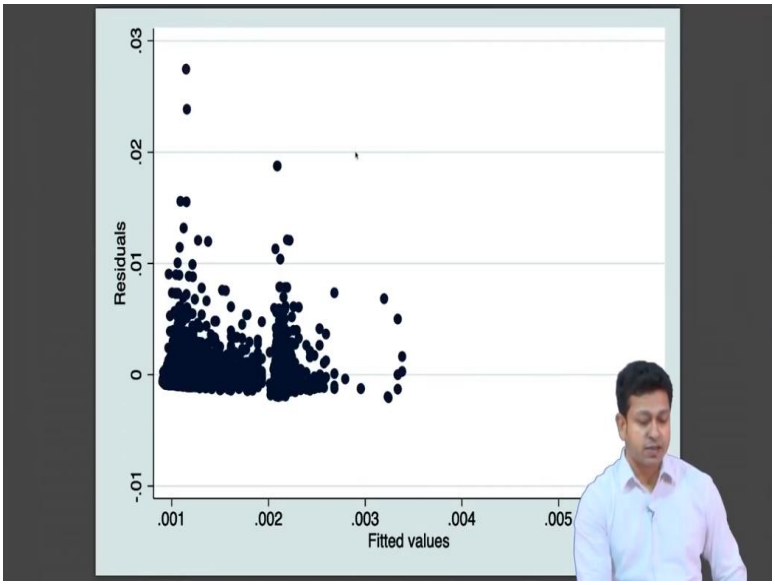
So, that means, I need to define it. So, let me define it. So, essentially, I have to run it. Now, let us say it should be working fine.

(Refer Slide Time: 13:09)



So, now I got the regression with my dependent variable as invwage. So, now I have to see what happens to the rvfplot. So, what will happen to the rvfplot? Now, I have changed the Y variable. So, essentially, we will see something similar to what you have seen previously. Here, in this case, since we have changed your Y variable, your rvfplot is going to change, because your Y is now 1 by Y.

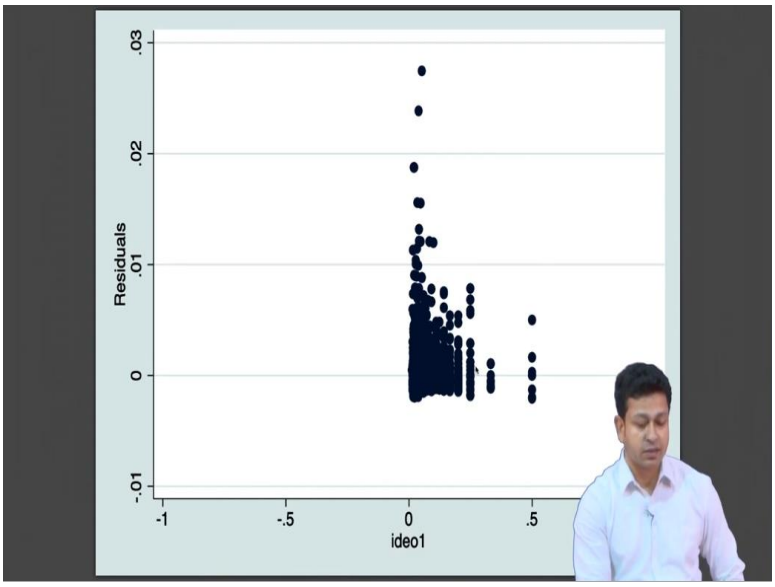
(Refer Slide Time: 13:38)



So, essentially, you see that. So, as your fitted values are lower, your dispersion in the error term is much higher, because now your Y variable is nothing but 1 by previous Y variable. So, that is why your dispersion is much higher when this new Y variable has a lower value. So, this is how by just by changing the functional form, you can actually get different representation of the heteroskedasticity. Now, you may also want to actually see what

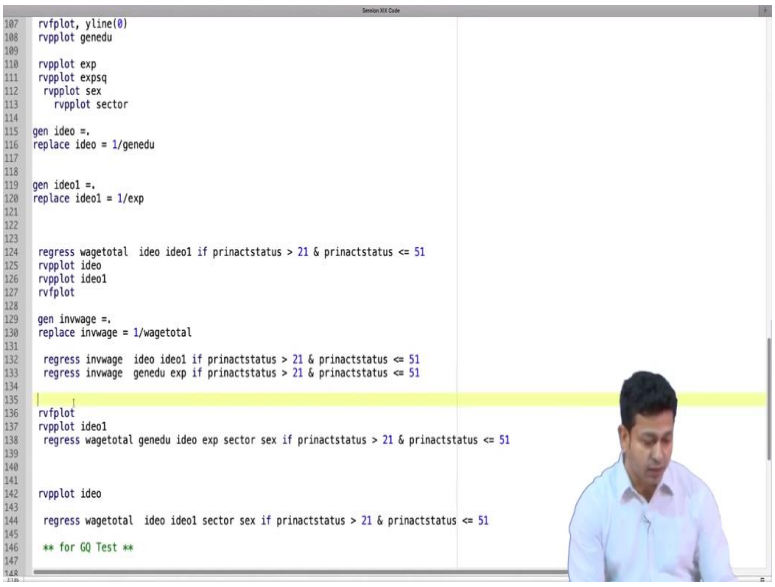
happens with the rvpplot here. So, of course, your rvpplot is not going to change, because the X variable has remained same.

(Refer Slide Time: 14:20)



Oh, I have actually kept ideo1.

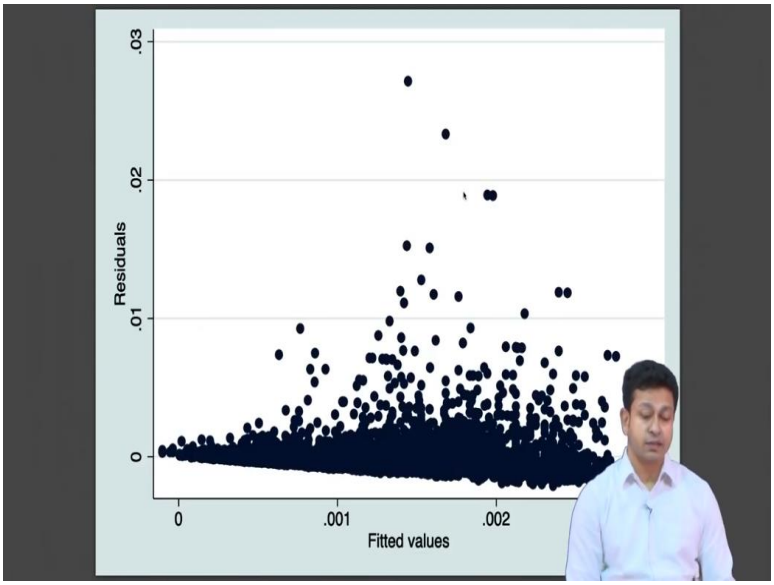
(Refer Slide Time: 14:36)



```
rvfplot, yline(0)
rvpplot genedu

rvpplot exp
rvpplot expsq
 rvpplot sex
   rvpplot sector

gen ideo =.
replace ideo = 1/genedu


gen ideo1 =.
replace ideo1 = 1/exp



regress wagetotal  ideo ideo1 if prinactstatus > 21 & prinactstatus <= 51
rvpplot ideo
rvpplot ideo1
rvfplot

gen invwage =.
replace invwage = 1/wagetotal

 regress invwage  ideo ideo1 if prinactstatus > 21 & prinactstatus <= 51
 regress invwage  genedu exp if prinactstatus > 21 & prinactstatus <= 51


rvfplot
rvpplot ideo1
 regress wagetotal genedu ideo exp sector sex if prinactstatus > 21 & prinactstatus <= 51



rvpplot ideo

 regress wagetotal  ideo ideo1 sector sex if prinactstatus > 21 & prinactstatus <= 51

 ** for GQ Test **
```

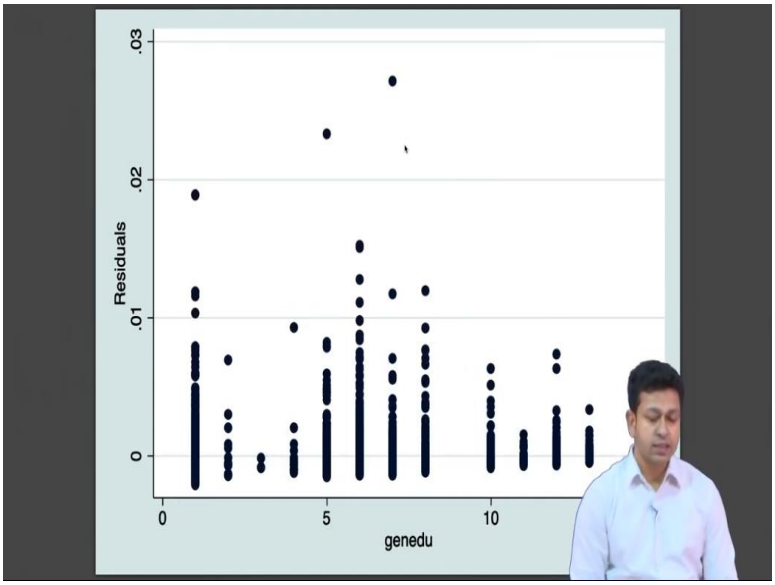
Instead, if I would have kept, say X genedu and experience; so, essentially, without changing the form of the X variable; and then, if I do rvpplot, then I will see;

(Refer Slide Time: 14:56)



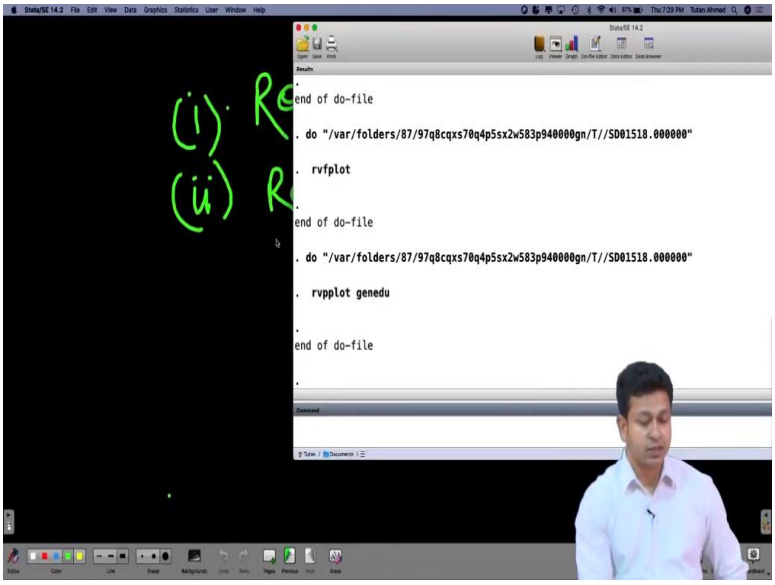
It will remain, it did not change; the residual now has a, again diverging nature of;

(Refer Slide Time: 15:04)



So, essentially, we will see that the residuals, it will show some changes because the residuals have changed, because the model is different; but it will not be like the converging error term that we have just seen previously.

(Refer Slide Time: 15:19)



So, that is how we need to understand if I do this residual versus fitted, residual versus predictor plot, and if we use different functional forms. So, with this, we will end the lecture here. And in the next lecture, we are going to talk about why this heteroskedasticity is a problem. So, we see what is a heteroskedasticity; we have seen how to understand heteroskedasticity; but in the next lecture, we are going to actually talk about why it is a problem. Thank you.