

Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 53

Relaxing the assumptions of CLRM-Autocorrelation and Heteroscedasticity Part - 4
(Refer Slide Time: 0:40)

On presence of heteroscedasticity

$\sigma_i^2 = f(x_i)$ H_0 : no heteroscedasticity

$= \sigma^2 x_i^2$

Goldfeld-Quandt Test:

$$RSS_1/df = \frac{n-c}{2} - K = \frac{n-c-2K}{2}$$

$$RSS_2/df = \frac{n-c}{2} - K = \frac{n-c-2K}{2}$$


$\frac{RSS_1/df}{RSS_2/df} \sim F_{\frac{n-c-2K}{2}, \frac{n-c-2K}{2}}$


H_0 is rejected if $F_{obs} > F_{tab}$

Steps:

- Step 1: Delete c number of central observations. $c = 4$ when $n = 50$, $c = 8$ when $n = 60$.
- Step 2: Run two separate regressions, one for the sample consisting of smaller x 's and the other with the higher x 's. $RSS_1 \rightarrow$ small x , $RSS_2 \rightarrow$ higher x .

Arrange the x values from small to big





And the first measure that we are going to talk about is Goldfeld-Quandt test for detecting heteroscedasticity. And before we discuss about different tests one thing you have to keep in mind, in presence of heteroscedasticity what we assume actually the sigma square σ_i is actually a function of x instead of being constant it is a function of x , either increasing or decreasing I do not know but it is a function of x .

Now, what type of specific functional form you will assume depending on that there is a separate test that the econometrician have developed to detect heteroscedasticity. For example, let us say that I have assumed this type of functional form, sigma square into x_i square, this is a specific type of functional form I have assumed between sigma square σ_i and x .

So, sigma square a constant factor multiplied by x_i square. So, as x increases you can understand how the sigma square σ_i will behave. Let us say that this is our assumption about this functional form and if this is the assumption then the test, corresponding test actually is called Goldfeld-Quandt test developed by Goldfeld and Quandt. So, they have given different step for this Goldfeld and Quandt test. What they say if you have your observation on y and x let us say

this is your observation, so the first step they say that arrange the x value from small to big. What is the first step? Arrange the x values from small to big, this is the first, small to be, you have to arrange.

Then what is step 2? In step 2 what they say after rearranging the x from small to big you delete, you delete c number of central observations and this c equals to 4 when n equals to 40, n equals to 40 and c equals to let not 40 let this is 30 and c equals to 8 when n equals to 60 something like that, something like that this is a rule of thumb basically, you delete c number of observations.

Then in step 3 what you do, run 2 separate regressions, one for the sample consisting of small x values or smaller xi's and the other with higher x values. So, that means once you arrange the sample from small to big and then you delete c number of central observation, you will get two different samples, in one sample you have small x values and in the other sample you will have higher x values.

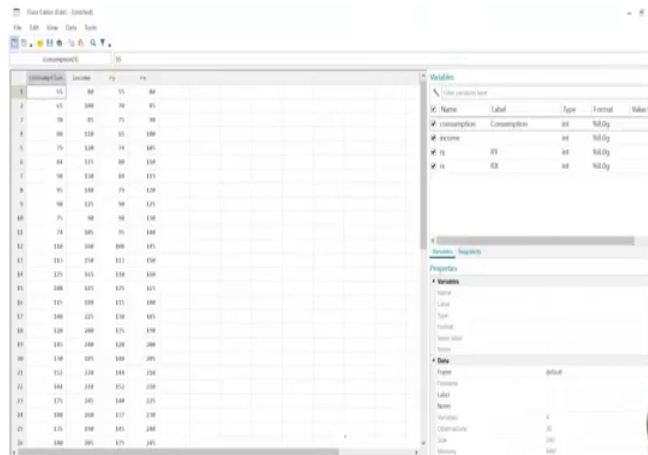
And collect RSS1 and RSS2, this is from the small x, from higher x. If you have two sample, if you run two separate regression you will get two different RSS. Now, what would be the degrees of freedom for RSS1 and RSS2? I have total n number of observations, I have deleted c number of observations. So, what would be the degrees of freedom for RSS1 and RSS2?

It is very simple, if you have n number of observation and you deleted c that means now you have n minus c by 2 for the first sample and second sample will also have n minus c by 2 and then degrees of freedom would be n minus k that means here it is n minus c by 2 minus k. So, that means this would become n minus c minus 2k divided by 2, this is also n minus c minus 2k divided by 2. This is the degrees of freedom for the RSS1 and RSS2.

And now, if you construct a test statistic in this way RSS2 by RSS1 and their corresponding degrees of freedom then that basically will follow an F statistic with degrees of freedom n minus c minus 2k by 2 for the numerator and n minus c minus 2k by 2 for the denominator. So, this would become your F statistic. And if that calculated F statistic is greater than the tabulated F statistic at a specific level of significance then what you will get, you will say that if calculated F, if calculated is greater than F tabulated then we will say that reject H_0 .

And what is our H naught here? The H naught is no heteroscedasticity. That is the decision that you can take. Now, what we will do? We will take one data set and then we will estimate a model and after that we will try to detect the heteroscedasticity.

(Refer Slide Time: 10:36)



Consumption	Income	Y1	Y2
53	58	55	58
63	68	70	65
78	85	75	90
88	118	85	100
75	108	74	105
88	115	90	110
98	138	88	115
95	138	75	120
98	125	98	125
75	98	98	138
74	105	95	148
118	158	108	145
113	158	113	150
113	158	113	150
125	165	118	160
108	165	125	165
115	168	115	168
148	215	138	165
138	208	135	158
145	208	138	208
138	195	148	205
152	218	148	218
148	218	152	218
175	265	168	215
188	268	157	218
175	198	145	248
188	265	175	245



And this is the data set if you look at, we have a data on income and consumption. For how many observations we have? We have total 30 observations. First we will estimate the model and we will get some initial impression using the graphical measure.

(Refer Slide Time: 11:12)

Stata 16.0

979-696-6661 (fax)

Single-user Stata perpetual license:
Serial number: 00166221948
Licensed to: NPTEL

Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.

*(4 variables, 30 observations pasted into data editor)

```
. reg consumption income
```

	Source	SS	df	MS	Number of obs =
	Model	41886.7134	1	41886.7134	496.72
	Residual	2361.15325	29	81.329018	
	Total	44247.8667	29	1525.78951	

	Source	SS	df	MS	F(1, 28)	Prob > F	R-squared	Adj R-squared	Root MSE
	Model	41886.7134	1	41886.7134	496.72	0.0000	0.9466	0.9467	9.183
	Residual	2361.15325	29	81.329018					
	Total	44247.8667	29	1525.78951					

	consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	income	.577846	.0260567	22.29	0.000	-.579166	.6964011
	_cons	9.290307	5.231386	1.78	0.087	-1.4257	20.00632



So, here what you are doing, you are estimating this model. what is the model? Reg consumption which is actually a function of income. Now, look at the result as usual the income is highly significant at 1 percent level because t statistics is 12.29 and corresponding p value is 0.000.

(Refer Slide Time: 12:10)

Stata 16.0

979-696-6661 (fax)

Single-user Stata perpetual license:
Serial number: 00166221948
Licensed to: NPTEL

Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.

*(4 variables, 30 observations pasted into data editor)

```
. reg consumption income
```

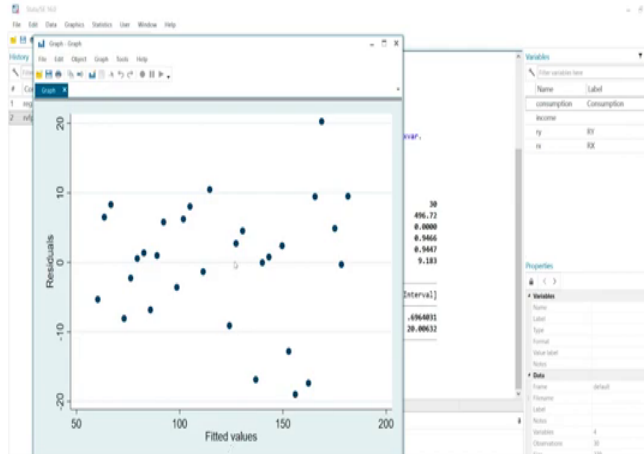
	Source	SS	df	MS	Number of obs =
	Model	41886.7134	1	41886.7134	496.72
	Residual	2361.15325	29	81.329018	
	Total	44247.8667	29	1525.78951	

	Source	SS	df	MS	F(1, 28)	Prob > F	R-squared	Adj R-squared	Root MSE
	Model	41886.7134	1	41886.7134	496.72	0.0000	0.9466	0.9467	9.183
	Residual	2361.15325	29	81.329018					
	Total	44247.8667	29	1525.78951					

	consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	income	.577846	.0260567	22.29	0.000	-.579166	.6964011
	_cons	9.290307	5.231386	1.78	0.087	-1.4257	20.00632

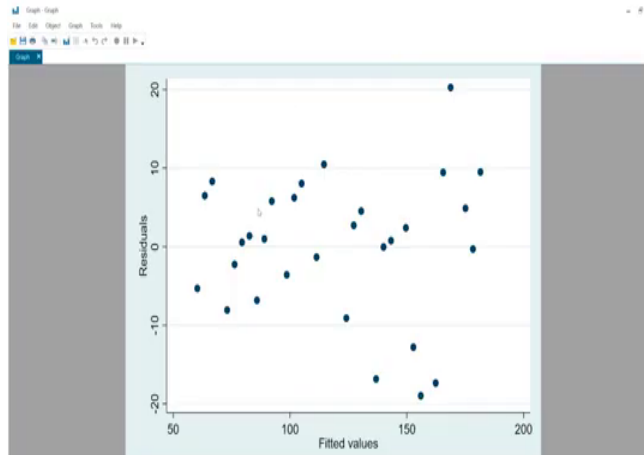
```
. rvfplot
```





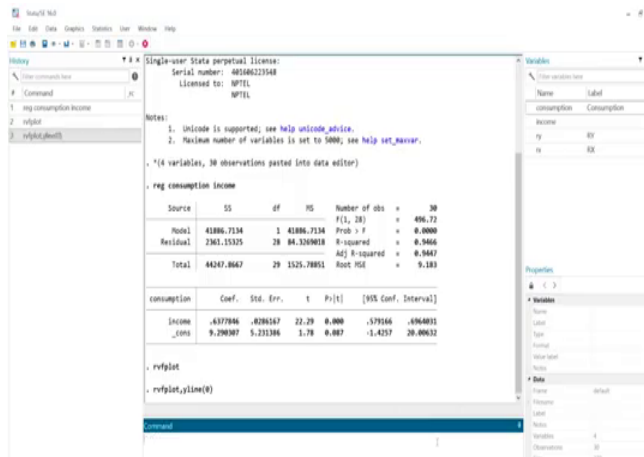
Now, if you try to put your error u_i hat square along with your y_i hat then there is a specific command in stata and that will tell you, this command is basically `rvf plot`. So, from this what we can say that though it is not very clear, the command is `rvf plot` and what is the meaning? `r` for residual and then `v` for versus, residual versus fitted, `f` for fitted, fitted value of y that is why the name `r v f` residual versus fitted plot. In the y axis I have residuals, in the x axis I have basic sorry, I have fitted so this is basically that means u_i hat square and the fitted that is how you are getting `rvf plot`.

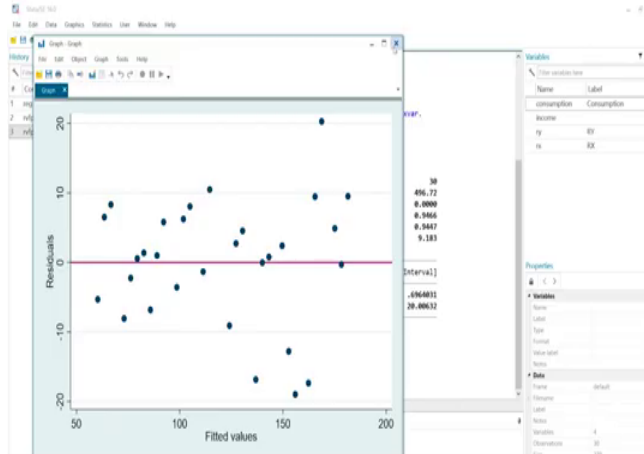
(Refer Slide Time: 13:12)



So, there is no clear pattern but we can say that it is not constant, see this is 0 line I can say that there is some kind of increase in the error variance.

(Refer Slide Time: 13:44)





If you put another comma let us say rvf plot, y line 0 then you will see this is a 0 line and this is how your errors are behaving, it is not constant, some kind of initial impression, but it is not very clear, that is why we need to go for Goldfeld and Quandt test.

And what is the first step? You need to arrange the x data from small to big and then it is not only about arranging the x you have to match those values with y also for running regression. So, you cannot simply arrange the x data and then run because you have that mapping between y and x that is very time consuming but in stata there is a specific command and if you put that command then stata will make your life very simple.

(Refer Slide Time: 14:57)

Stata 16.0

History

```

1. reg consumption income
2. rvfplot
3. rvfplot, yline(0)

```

Notes:

- Unicode is supported; see help unicode_advice.
- Maximum number of variables is set to 5000; see help set_maxvar.

.*(4 variables, 30 observations pasted into data editor)

```

. reg consumption income

```

Source	SS	df	MS	Number of obs	F(1, 28)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	4186.7134	1	4186.7134	30	496.72	0.0000	0.9466	0.9447	9.183
Residual	2361.15325	28	84.3268018						
Total	44347.8647	29	1525.78953						

consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.4377846	.0286167	22.29	0.0000	-.579166	.6964811
_cons	9.290307	5.231386	1.78	0.087	-1.4257	20.00632

```

. rvfplot
. rvfplot, yline(0)

```

Command

```

reg consumption income

```

Variables

Name	Label
consumption	Consumption
income	Income
ry	RY
rx	RX

Properties

Variables

Name	Label	Type	Format	Position	Width	Alignment
consumption	Consumption	float	%12.0f	1	12	right
income	Income	float	%12.0f	2	12	right
ry	RY	float	%12.0f	3	12	right
rx	RX	float	%12.0f	4	12	right



How will you do that? You put this command sort x, here it is sort x means income, sort income. So, I will first sort income and then what I will do, I will first run a regression for the first total number of observation is 30 and I remove 4 observations, central 4 observations and then that means the first sample consists of 13 observation and second sample consists of 13 observations.

(Refer Slide Time: 15:53)

Stata 16.0

History

```

1. reg consumption income
2. rvfplot
3. rvfplot, yline(0)
4. sort income
5. reg consumption income in 1/13

```

consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.4377846	.0286167	22.29	0.0000	-.579166	.6964811
_cons	9.290307	5.231386	1.78	0.087	-1.4257	20.00632

```

. rvfplot
. rvfplot, yline(0)

```

```

. sort income
. reg consumption income in 1/13

```

Source	SS	df	MS	Number of obs	F(1, 11)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	3818.06412	1	3818.06412	13	87.79	0.0000	0.8887	0.8785	5.856
Residual	377.16423	11	34.2857842						
Total	3387.23077	12	282.269231						

consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.4967742	.074366	9.37	0.0000	-.5338958	.6666326
_cons	3.480429	0.788254	0.39	0.703	-15.74998	11.56884

Command

```

reg consumption income in 1/13

```

Variables

Name	Label
consumption	Consumption
income	Income
ry	RY
rx	RX

Properties

Variables

Name	Label	Type	Format	Position	Width	Alignment
consumption	Consumption	float	%12.0f	1	12	right
income	Income	float	%12.0f	2	12	right
ry	RY	float	%12.0f	3	12	right
rx	RX	float	%12.0f	4	12	right



So, that means first I have to run a regression `reg consumption income` for the first 13 observations and if you recall once again how will you put that command? The command is in 1 by 13 that means from 1 to 13, this is your first regression.

(Refer Slide Time: 16:26)

The screenshot shows the Stata 16.0 interface. The command window contains the following commands:

```

1. reg consumption income
2. nlfit
3. nlfit,plot
4. out income
5. reg consumption income in 1/13
6. reg consumption income in 18/30

```

The results window displays the following statistics for the first regression:

Model	SS	df	MS	F(1, 11)	Prob > F
Model	3818.86452	1	3818.86452	87.79	0.0000
Residual	377.164251	11	34.2879412		0.8887
Total	4196.02877	12	349.670731		0.8785

The second regression results are as follows:

Model	SS	df	MS	F(1, 11)	Prob > F
Model	5088.89274	1	5088.89274	36.42	0.0001
Residual	1516.79957	11	137.89052		0.7681
Total	6605.69231	12	550.47436		0.7679



And then for the second regression `reg consumption` then `income in`, so up to 13 after that 14, 15, 16 and 17, these central four observations are deleted so in second regression you have to start from 18 to 30. So, this is how you have constructed. Now, what you should do? You should actually construct your F statistic and how will you do that?

(Refer Slide Time: 17:26)

Stata 16.0

```

reg consumption income in 18/30
-----+-----
Model      3018.06412   1   3018.06412   F(1, 11)   =   87.79
Residual   377.162753   11  34.2878412   Prob > F   =   0.0000
Total      3387.23077   12  282.249331   R-squared  =   0.8887
                                           Adj R-squared =   0.8785
                                           Root MSE   =   5.8566

      consumption   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----
      income        .6967742   .074366   9.37   0.000   .538958   .8606526
      _cons         3.409429   0.780924   4.37   0.001   1.847006   5.011852

. reg consumption income in 18/30
-----+-----
Source      SS           df   MS       Number of obs =   13
-----+-----
Model      3018.06274   1   3018.06274   Prob > F       =   0.0001
Residual   377.16275   11  34.2878412   R-squared      =   0.7481
Total      3387.23077   12  282.249331   Adj R-squared  =   0.7470
                                           Root MSE     =   5.8566

      consumption   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----
      income        .7981373   .1313879   6.04   0.000   .5043274   1.091947
      _cons         -28.07717   38.64214   -0.91   0.380   -95.47006   39.415173
    
```



Stata 16.0

```

reg consumption income in 18/30
-----+-----
Model      3018.06412   1   3018.06412   F(1, 11)   =   87.79
Residual   377.162753   11  34.2878412   Prob > F   =   0.0000
Total      3387.23077   12  282.249331   R-squared  =   0.8887
                                           Adj R-squared =   0.8785
                                           Root MSE   =   5.8566

      consumption   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----
      income        .6967742   .074366   9.37   0.000   .538958   .8606526
      _cons         3.409429   0.780924   4.37   0.001   1.847006   5.011852

. reg consumption income in 18/30
-----+-----
Source      SS           df   MS       Number of obs =   13
-----+-----
Model      3018.06274   1   3018.06274   Prob > F       =   0.0001
Residual   377.16275   11  34.2878412   R-squared      =   0.7481
Total      3387.23077   12  282.249331   Adj R-squared  =   0.7470
                                           Root MSE     =   5.8566

      consumption   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----
      income        .7981373   .1313879   6.04   0.000   .5043274   1.091947
      _cons         -28.07717   38.64214   -0.91   0.380   -95.47006   39.415173
    
```



So, that means you have to remember your RSS2 and RSS1, so RSS2 is 1536.79, so I will put on excel sheet 1536, so this is what is the value 1536.79 that is RSS2, 1536.79 and then what is RSS1, RSS1 is 377.16, 377.16. If you divide this by this, this is your F statistic. So, basically what you have to do, RSS2 by RSS1 which is 4.79. And what is the degrees of freedom for the numerator and denominator for this? The degrees of freedom would be 11 and 11.

(Refer Slide Time: 18:59)

On presence of heteroskedasticity
 $\sigma^2 = f(x)$
 $= \sigma^2 x_i^2$
 H₀: no heteroskedasticity
 Reject H₀ at 5% sig

Goldfeld-Quandt Test

$$RSS_1/df = \frac{n-c}{2} - R = \frac{n-c-2R}{2}$$

$$RSS_2/df = \frac{n-c}{2} - R = \frac{n-c-2R}{2}$$

Step 1: Arrange the X values from small to big

Step 2: Delete c number of central observations
 c = 4 when n = 30, c = 8 when n = 60

Step 3: Run two separate regressions, one for the smaller X's and the other with the higher X's
 RSS₁ → small X
 RSS₂ → higher X

Step 4: $F_{cal} = \frac{RSS_2/df}{RSS_1/df}$
 $F_{cal} = 4.07$
 $F_{tab} = 2.82 (5\%)$
 $F_{tab} = 4.46 (1\%)$
 If $F_{cal} > F_{tab} \Rightarrow$ Reject H₀

Goldfeld-Quandt Test

$$RSS_1/df = \frac{n-c}{2} - R = 11 - 11 = 0$$

$$RSS_2/df = \frac{n-c}{2} - R = 11 - 11 = 0$$

$$F_{cal} = \frac{RSS_2/df}{RSS_1/df} = 4.07$$

Excel Spreadsheet Data:

Y	X
1536.79	377.55
40216.57	40216.57

Step 1: Arrange the X values from small to big

Step 2: Delete c number of central observations
 c = 8 when n = 60

Step 3: Run two separate regressions, one for the smaller X's and the other with the higher X's
 X
 etc

So, that means while calculating I will go back to our formula. This is our formula RSS2 by RSS1 and here what you are doing basically this calc, this if you apply this I am not dividing by degrees of freedom because that is 11 and 11 and this is 4.07. which is the calculated value.

F calculated equals to 4.07, you have to now find out F tabulated at 1 percent and 5 percent level of significance with 11 and 11 degrees of freedom. If you do so, if you go to your F table and mention your F statistic and then you will see that this is 2.82 at 5 percent and 4.46 at 1 percent.

So, that means at least you can say that you can reject your null hypothesis at 5 percent level of significance, you can reject your null hypothesis at 5 percent reject null level of significance.

So, reject H_0 at 5 percent significance. But if you determine that no I will reject the null only at 1 percent then actually you cannot reject your null that is why from the graphical representation also it was not a very clear indication about heteroscedasticity, but somehow, we have some kind of indication because rejecting null at 5 percent is also sometimes acceptable.

So, this would be your decision making, this is about your Goldfeld and Quandt test. But one thing is not clear, why we are deleting the central observations? see we have arranged the data and then we have deleted the four observations so that the residual RSS of the second sample becomes quite different from RSS of the first sample based on that logic only we are deleting the central observations. Are you getting my point?

First you have arranged the data from small to big and then we are deleting the central observations also, so that my RSS_2 becomes significantly different from RSS_1 , higher the difference between RSS_2 and RSS_1 higher would be the probability of getting evidence for heteroscedasticity, because what we assume that error variance is not constant rather it is increasing as x increases.

That is why when I arrange and run the regression for the other sample, sample 2, where all the large x values are there, we assume that RSS_2 would be quite large than the RSS_1 and to get a very stark difference between RSS_2 and RSS_1 the central observations are deleted.

But there are certain limitations about this Goldfeld-Quandt test. First of all, the c , how many number of observations I will delete there is no clear cut statistical rule which will say this is the value corresponding to this particular n . This is kind of rule of thumb, c equals to 4 when n equals to 30; c equals to 8 when n equals to 60, c equals to 10 when n equals to 120, something like that.

(Refer Slide Time: 23:56)

Limitations of Goldfeld-Quandt test

- (i) there is no clear rule for detecting c
- (ii) if there are more number of explanatory variable, then the same procedure needs to be repeated.

So, Limitations of Goldfeld-Quandt test. First, there is no clear rule for detecting c . Secondly, if there are more number of explanatory variables then the same procedure needs to be repeated, here I have only one variable, one explanatory variable x , that is why it is very simple I arrange x , delete the central c observations and run two different regressions.

But suppose I have now x_1, x_2, x_3, x_4 and x_6 so I have to arrange all these x variables x_1, x_2, x_3, x_4, x_6 , delete central observations and run separate regression, so the process would be very time consuming, these are the limitations. So, in our next class what we will do, we will discuss about another test which is more general in nature, you do not need to delete these observations that means at one go we can implement the test without repeating the procedure. So, we will go for more advanced test for detecting heteroscedasticity in our next class.