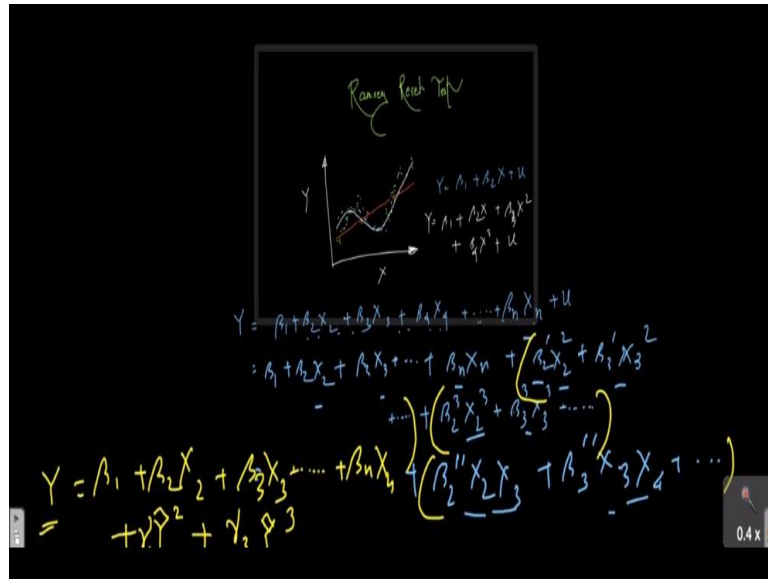


**Applied Econometrics**  
**Prof. Tutan Ahmed**  
**Vinod Gupta School of Management**  
**Indian Institute of Technology - Kharagpur**

**Lecture – 92**  
**Ramsey Reset Test**

(Refer Slide Time: 00:27)



Hello and welcome back to the lecture on applied econometrics. We have been talking about model specification. And we talked about the different problems, different sources of model misspecification problem. But we need to have some tests to identify if there is a model specification problem and one such test is Ramsey Reset Test or omitted variable bias test. So let us try to understand what this test is talking about.

Before I actually get into that actually conceptually explain what are the different problems that may happen when we have let us say some omitted variable. Let us say I have this plot here, not the grid X and Y coordinate, but let us say Y is here and X is here. And let us say my X's and Y actual distribution is something like this. But I actually let us say I have ended up getting some sort of linear model.

Let us say this is a very simple model so where I write  $Y = \beta_1 + \beta_2 X + u$  that is a linear model and I just fit the linear model, and I get perhaps something like this. But that is clearly not the best fit. The best fit perhaps would be if I have something like this, so that

would be a better fit. But the linear equation is not able to capture this pattern, this sort of ups and downs that we have in my actual fit, actual line that I should.

So perhaps it is a better idea if we have some sort of quadratic equation. Let us say if I have  $Y$  is equal to some sort of  $\beta_1 + \beta_2 X + \text{let us say } X \beta_3 x^2$ , then let us say some  $\beta_4 X^3$  and so forth and then then we have some error term, you can have a power of 4,  $X$  can have power of 4 and so forth. Now, this is something that we can have for a simple OLS. But let us say when you have many, many different independent variables, not just  $X_1$ ; but there are  $X_2$ ,  $X_3$  and  $X_4$  and so on.

So if we have such kind of situation, then if my  $Y$  actually if I feed a linear regression as  $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3$  and let us say I have a  $\beta_4 X_4$  and up to a higher end of let me say  $n$  variable  $\beta_n X_n$  plus some error term. If this is my actual linear equation, now then if I have such kind of pattern, such kind of nonlinearity existing in the model, so then this model is definitely not be able to capture a nonlinear pattern because essentially this is a linear model.

But if I want to include this nonlinearity, what I can do is I can actually include let us say  $\beta_1 + \beta_2 X_2^2$  and I am not writing so many different terms,  $\beta_3 X_3^3$  and then  $\beta_n X_n^n$  and then I actually do a square of the whole thing, let us say  $\beta_2 X_2^2$  square. So I basically take square of the explanatory variables and this is  $\beta_2'$  and I have a different coefficient here,  $\beta_3' X_3^3$  square and so on.

And then I will have let us say  $\beta_n$ . Then I will have  $\beta_2^3 X_2^3 + \beta_3^3 X_3^3$  and so forth. So basically what I have done, I have simply taken a square term of all these different explanatory variables. Now, it may not just end here, I can have further terms which are essentially interaction terms. So, I will have  $X_2^2 X_3$ ,  $X_3 X_4$  and so forth. So I can add more terms, instead of writing the equation one more time I can basically add here let us say  $\beta_2'' X_2 X_3$ , then  $\beta_3'' X_3 X_4$  and so on.

So, this way I can have many, many different terms in terms of the higher order terms as well as the interaction terms. So, essentially if I want to have that that would be really a daunting sort of expression and more importantly with so many different independent variables, my degrees of freedom will get reduced. So, I do not want that. Instead of that what Ramsey

proposed is that instead of having this squared terms or the cube terms what if we can simply; let me just note it here.

So, instead of having these square terms or the cube terms or the intersection terms here, what if we can simply have something like  $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n$  you have and then you add to that,  $\hat{Y}$  so estimated  $Y$  square, basically instead of squaring so many different terms you simply square the estimated, you first estimate the  $Y$  and then you take a square of that and then you have any sort of gamma.

Let us say we take a gamma or gamma 1 or gamma 2, whatever you want to take. And then you can also have if you want to have cube terms, you can also have gamma 3 and let us say  $\hat{Y}$  cube. So the beauty of taking the square or cube terms is that; let me use a different page.

(Refer Slide Time: 07:35)

$$\hat{Y}^2 = [\beta_1 + \beta_2 X_2 + \dots + \beta_n X_n]^2$$

$$= \beta_1^2 + \beta_2^2 X_2^2 + \dots + 2\beta_1 \beta_2 X_1 X_2 + \dots$$

$\hat{Y}^3, \hat{Y}^4$

$$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_n X_n + \gamma_2 \hat{Y}^2 + \gamma_3 \hat{Y}^3 + \gamma_4 \hat{Y}^4 + u$$

The beauty of using this square cube term is that essentially the  $\hat{Y}$  hat square would actually take into account, so let us say my original equation was  $\beta_1 + \beta_2 X_2 + \dots + \beta_n X_n$ . So if I take a square of  $\hat{Y}$  hat, then I will get square of all these different terms as well as the intersection of these terms. So like, I have  $\beta_1^2 + \beta_2^2 X_2^2 + \dots$  and the intersection terms, so  $2\beta_1 \beta_2 X_1 X_2$ , essentially ideally I am basically capturing that.

So because of these, it is really convenient to actually estimate the  $\hat{Y}$  hat first, and once I estimate the  $\hat{Y}$  hat, I can use a  $\hat{Y}$  hat as a regressor. So it becomes an independent variable,

estimated  $\hat{Y}$ . And once I can take a square of it, I can take a cube of it, I can take a fourth power of it. So all this different power of  $\hat{Y}$  could be used as an independent regressor. Now, in fact when you do Ramsey Reset test, usually the original equation we actually take power up to 4.

So essentially my  $\hat{Y}$ ; let me write down the final equation that we will be using for Ramsey Reset test is  $\hat{Y} = \beta_1 + \beta_2 \hat{Y}^2 + \beta_3 \hat{Y}^3 + \beta_4 \hat{Y}^4$ . So, that is what we are going to use my regression equation where I am going to do the reset test. So, ideally why we actually take the squares and what is the whole point here that we have done so far?

(Refer Slide Time: 10:00)

if there are problems of model misspecification

- i)  $\hat{Y}$
- ii)  $\hat{Y}^2, \hat{Y}^3, \hat{Y}^4$
- iii) Regression with eqn  $\hat{Y} = \beta_1 + \beta_2 \hat{Y}^2 + \beta_3 \hat{Y}^3 + \beta_4 \hat{Y}^4$

$$F = \frac{(RSS_R - RSS_{UR}) / (m-1)}{RSS_{UR} / (n-k)}$$

We have explained that the importance of having the higher order independent variable and the importance of having interaction terms but one thing we have to remember that Ramsey Reset test can at best tell you if there are problems of model misspecification, but it will not be able to tell you which exact variable you need to include, what power of that variable you need to include so that the Ramsey Reset test will not be able to tell you.

It can only tell you that well there is a problem, you need to perhaps include higher order terms of the independent variable. So, essentially it is not very specific in terms of telling you which variables to include, alright. So, with this we will actually see what are the different steps that we need to do here. So, one we understood that first we get the  $\hat{Y}$ , and then we take the  $\hat{Y}$  square,  $\hat{Y}$  cube,  $\hat{Y}$  to the power 4 and then you run the final regression with equation.

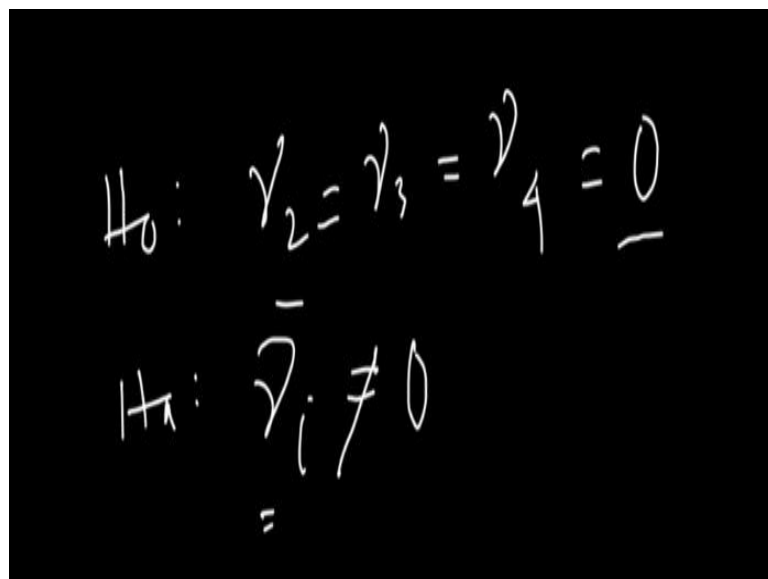
So let us say this is my equation star, regression equation star. Regression with equation star. Now, what exactly happens here when we do that? So, essentially in Ramsey Reset test what it does is it calculates the F statistic and in the first model where I do not have any of these higher order  $\hat{Y}$  terms, so this is my restricted model. So, we remember how to use restricted and unrestricted model to get the F statistic.

So, essentially for the restricted model, we have residual sum square restricted and then you have the second model which is unrestricted model where we have included all these different terms. So, you take the residual sum of squared RSS unrestricted and then you divide that by the degrees of freedom, which is essentially if you have  $m$  number of independent variables in your model including the intercept, so you do  $m-1$ .

Whereas in the denominator you just take the residual sum square for the unrestricted model and you divide it by if there are a total  $n$  number of observations, then  $n$  minus, now in this case I will have a different number of explanatory variable because I am including these different terms of square, cube and to the power 4 and so that is why you have to calculate the F statistic when we try to understand which model is better.

And my null hypothesis in this case is going to be; so all these gammas, so that the gammas that I have written here; gamma 2, gamma 3, gamma 4, so they are going to be 0.

**(Refer Slide Time: 13:30)**



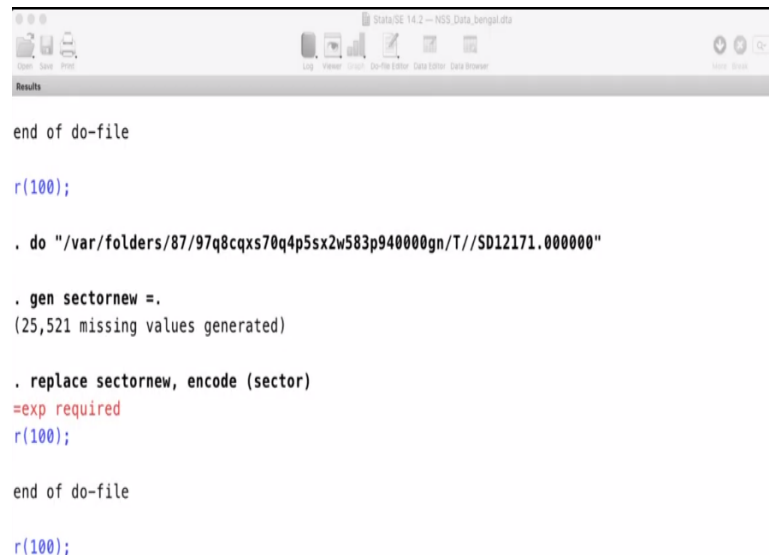
The image shows handwritten mathematical expressions on a black background. The first line represents the null hypothesis:  $H_0: \gamma_2 = \gamma_3 = \gamma_4 = 0$ . The second line represents the alternative hypothesis:  $H_1: \gamma_i \neq 0$ .

So, my null hypothesis is that in  $\gamma_2 = \gamma_3 = \gamma_4$  whichever  $\gamma$  I take that is equal to 0. So, if I take only let us say  $\hat{Y}$  square and not the higher order term, so in that case I will only have  $\gamma_2 = 0$ . So, if my  $\gamma$ s are 0, so then the higher order  $\hat{Y}$  terms will not have any significant contribution towards explaining the model and which is good for me because that means my model that I have actually written at the beginning is correctly specified.

But if we have these higher order terms if they are significant, so that means we actually need to incorporate the higher order terms, higher order variables essentially or the square of intersection terms. So all these different aspects we need to consider. And if my alternative hypothesis is going to be  $\gamma_i$  not is equal to 0. So if any of the  $\gamma$ s are not equal to 0, so then I will say my null hypothesis is rejected.

So that is basically the idea and the theory behind estimating the Ramsey Reset test. Now, what we are going to do is that we are going to actually see using the data that we have already used previously, the Ramsey Reset test.

**(Refer Slide Time: 14:52)**



```
Results
end of do-file

r(100);

. do "/var/folders/87/97q8cqs70q4p5sx2w583p940000gn/T//SD12171.000000"

. gen sectornew =.
(25,521 missing values generated)

. replace sectornew, encode (sector)
=exp required
r(100);

end of do-file

r(100);
```

So let me actually show you the data. So exactly the same data we have used previously that is the national sample survey data for the state of West Bengal.

**(Refer Slide Time: 14:57)**

	fsu	sample	sector	state	region	district	restinfo
1	41910	1	2	19	2	07	0701032193111
2	41910	1	2	19	2	07	0701032193111
3	41910	1	2	19	2	07	0701032193111
4	41910	1	2	19	2	07	0701032193111
5	41910	1	2	19	2	07	0701032193111
6	41910	1	2	19	2	07	0701032193111
7	41910	1	2	19	2	07	0701032193111
8	41910	1	2	19	2	07	0701032193111
9	41910	1	2	19	2	07	0701032193111
10	41910	1	2	19	2	07	0701032193112
11	41910	1	2	19	2	07	0701032193112
12	41910	1	2	19	2	07	0701032193112
13	41910	1	2	19	2	07	0701032193112
14	41910	1	2	19	2	07	0701032193112
15	41910	1	2	19	2	07	0701032193112
16	41910	1	2	19	2	07	0701032193112
17	41910	1	2	19	2	07	0701032193112

So what we are going to do we will actually run the code here and we will see what is the result of the Ramsey Reset test and here let me actually go back to the code. So let me actually run this equation.

(Refer Slide Time: 15:03)

```

1
2 gen exp =,
3 replace exp = age - 5 - genedu
4
5 gen expsq = .
6 replace expsq = exp*exp
7
8 gen lnwagetotal =,
9 replace lnwagetotal = log(wagetotal)
10
11 regress wagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51
12 regress lnwagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51
13
14
15 regress lnwagetotal genedu if prinactSTATUS > 21 & prinactSTATUS <= 51
16 regress lnwagetotal genedu exp if prinactSTATUS > 21 & prinactSTATUS <= 51
17 regress lnwagetotal exp if prinactSTATUS > 21 & prinactSTATUS <= 51
18
19 regress lnwagetotal genedu exp if prinactSTATUS > 21 & prinactSTATUS <= 51
20 regress lnwagetotal genedu exp expsq if prinactSTATUS > 21 & prinactSTATUS <= 51
21 regress lnwagetotal genedu expsq if prinactSTATUS > 21 & prinactSTATUS <= 51
22
23 regress lnwagetotal expsq if prinactSTATUS > 21 & prinactSTATUS <= 51
24
25 regress lnwagetotal genedu exp sex if prinactSTATUS > 21 & prinactSTATUS <= 51
26
27 ovtest
28
29 regress lnwagetotal genedu exp expsq sex sector if prinactSTATUS > 21 & prinactSTATUS <=

```

Let us say my equation is this. So log of wage total, general education, experience, let us say I also had sex and then it is for a specific group of people. So prinactstatus is between 21 and 51. So I have already defined all those variables, I am not doing it again. So if I run this regression, the result I get is here.

(Refer Slide Time: 15:50)

Results						
Model	2016.38158	3	672.127194	F(3, 4255)	=	1229.26
Residual	2326.51656	4,255	.5467724	Prob > F	=	0.0000
				R-squared	=	0.4643
				Adj R-squared	=	0.4639
Total	4342.89815	4,258	1.0199385	Root MSE	=	.73944
lnwagetotal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
genedu	.1873304	.0031932	58.66	0.000	.18107	.1935908
exp	.0221558	.0009104	24.34	0.000	.020371	.0239407
sex	-.2795995	.0301368	-9.28	0.000	-.3386833	-.2205156
_cons	5.551327	.0519324	106.90	0.000	5.449513	5.653142
. end of do-file						

So, I see that the R squared values 0.46 and all these variables are actually explaining the model pretty nicely general education, experience and sex. Now we know that for Mincerian equation, we already have an experience square, which we need to actually take into account which I have not done here in this particular case. What I will do is let me actually do omitted variable bias test.

So, I will simply write a command that is ovtest which is rather simple command, this omitted variable bias test. If I write ovtest, it will give me the results. I will run it and let us see what result I got.

**(Refer Slide Time: 16:32)**

Results						
_cons	5.551327	.0519324	106.90	0.000	5.449513	5.653142
. end of do-file						
. do "/var/folders/87/97q8cqx570q4p5sx2w583p940000gn/T//SD12171.000000"						
. ovtest						
Ramsey RESET test using powers of the fitted values of lnwagetotal						
Ho: model has no omitted variables						
F(3, 4252) = 164.67						
Prob > F = 0.0000						
. end of do-file						

Interesting, so we see that ovtest actually gives you a result Ramsey Reset test, the same thing, ovtest is the command that is actually used for Ramsey Reset test and it says that using



powers of the fitted values of the log of wage total and  $H_0$  is model has no omitted variable. So it means that  $\gamma_2$ ,  $\gamma_3$ ,  $\gamma_4$  they are 0 and essentially in case of omitted variable bias test as I said, we use up to fourth power of the  $\hat{Y}$  variable.

So essentially in this case, my F statistic has a value of 164.67 and the P value is pretty less. So it means that my null hypothesis is rejected and I say that there is a possibility of presence of higher order of the  $\hat{Y}$ . So that essentially says that my model has some omitted variable bias problem. Now, going forward we can also think that in the previous equation, what I had is that I actually did not include the experience square term.

So that is one term I know for a fact that I need to include experience squared. So let us say if we include experience squared what happens? So, I will just run the equation including the experience square.

**(Refer Slide Time: 17:52)**

```

3  replace exp = age - 5 - genedu
4
5  gen expsq = .
6  replace expsq = exp*exp
7
8  gen lnwagetotal = .
9  replace lnwagetotal = log(wagetotal)
10
11 regress wagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51
12 regress lnwagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51
13
14
15 regress lnwagetotal genedu if prinactSTATUS > 21 & prinactSTATUS <= 51
16 regress lnwagetotal genedu exp if prinactSTATUS > 21 & prinactSTATUS <= 51
17 regress lnwagetotal exp if prinactSTATUS > 21 & prinactSTATUS <= 51
18
19 regress lnwagetotal genedu exp if prinactSTATUS > 21 & prinactSTATUS <= 51
20 regress lnwagetotal genedu exp expsq if prinactSTATUS > 21 & prinactSTATUS <= 51
21 regress lnwagetotal genedu expsq if prinactSTATUS > 21 & prinactSTATUS <= 51
22
23 regress lnwagetotal expsq if prinactSTATUS > 21 & prinactSTATUS <= 51
24
25 regress lnwagetotal genedu exp sex if prinactSTATUS > 21 & prinactSTATUS <= 51
26
27 ovtest
28
29 regress lnwagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51
30

```

Let me write. There I have written some other equations, so let us just not get into that. So let me just write down this and I will include experience squared. And following that I will again run the ovtest command so as to see if my omitted variable problem is solved or not. So I have run the regression and then I will run the ovtest.

**(Refer Slide Time: 18:22)**

```

Results
_cons      5.213611   .0603429   86.40   0.000   5.095307   5.331914

.
end of do-file

. do "/var/folders/87/97q8cqs70q4p5sx2w583p940000gn/T//SD12171.000000"

. ovtest

Ramsey RESET test using powers of the fitted values of lnwagetotal
Ho: model has no omitted variables
      F(3, 4251) =    139.05
      Prob > F =    0.0000

.
end of do-file

```

Now, let us look at the results. Well, what I see here is this. I see that the F statistic still; so let me actually see the that regression has already actually happened.

**(Refer Slide Time: 18:34)**

Source	SS	df	MS	Number of obs = 4,259
Model	2076.35749	4	519.089374	F(4, 4254) = 974.26
Residual	2266.54065	4,254	.532802222	Prob > F = 0.0000
Total	4342.89815	4,258	1.0199385	R-squared = 0.4781
				Adj R-squared = 0.4776
				Root MSE = .72993

lnwagetotal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
genedu	.1838988	.0031687	58.04	0.000	.1776864 .1901112
exp	.0555027	.003269	16.98	0.000	.0490937 .0619116
expsq	-.000595	.0000561	-10.61	0.000	-.000705 -.0004851
sex	-.2830172	.0297511	-9.51	0.000	-.3413448 -.2246896
_cons	5.213611	.0603429	86.40	0.000	5.095307 5.331914

And I see that we have all these independent variables education, experience, experience square, sex and all the variables are actually significant. My model has the R squared values improved. Now, when I look at the ovtest result, what I see is that my F statistic has a rather large value of 139.05 for given degrees of freedom and the P value is low and is 0 essentially. So it means that my model there still have omitted variable bias problem even if I have included experience squared term. So essentially, we have to work further to improve the model.