


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

Application of STATA for hypothesis testing and introduction to multiple linear regression model Part - 3

(Refer Slide Time: 00:14)



$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ki} + u_i ; i=1, 2, \dots, n$


In the context of 2 var CLRM,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \dots + \beta_k x_{k1} + u_1 \rightarrow 1^{st} \text{ person}$
 $y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \dots + \beta_k x_{k2} + u_2 \rightarrow 2^{nd} \text{ } \dots$
 $y_3 = \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \dots + \beta_k x_{k3} + u_3 \rightarrow 3^{rd} \text{ } \dots$
 \vdots
 $y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \dots + \beta_k x_{kn} + u_n \rightarrow n^{th} \text{ } \dots$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{pmatrix}_{n \times (k+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1) \times 1} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}_{n \times 1}$$

$Y = X\beta + U ; y_i = \beta_0 + u_i$



How do you estimate multiple linear regression model? Let us write our generalized multiple linear regression model as $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_k x_{ki} + u_i$ This is our model. Now, in this model how do you estimate $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$. Now, in the context of 2 variable CLRM, how we get the $\hat{\beta}_1$?

You remember? $\hat{\beta}_1$, we derived as
$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$
, i running from 1 to n. Here, i equals to 1, 2, n. So, i here stands for individual and we have k number of explanatory variables here. So, if we know x and y, we can simply apply this formula to get our beta1 hat.

Now, if we want to estimate so many beta hat, then what we need to do? We need to apply this formula for all this k number of variables which is very very time consuming process. You can

understand if we have 10 or 15 explanatory variables and you need to compute $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, then you have to apply this formula for so many times to calculate this $\hat{\beta}$.

So, if we apply the technique, what we have learnt in our 2 variable classical linear regression model in this setup to estimate the parameters, then we need to apply the formula so many times. But, we do not actually do that in the context of multiple linear regression model. What we do? We represent this equation using metrics notation and then, from that metrics notation we can easily solve all the parameters at one shot. That is why metrics algebra becomes very very important in this context.

Let us see how do you represent the equation in a metrics notation. So, this equation what I have written, $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_k x_{ki} + u_i$ is for ith individual. Now, we can substitute i for 1, then we will get the equation for the first individual, we can substitute i for 2, then we will get same equation for the second individual. Likewise, we will get the equation for the third individual, fourth individual and then up to nth individual.

How does it look like? Let us substitute i for 1, then we will get $Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \beta_3 x_{31} + \beta_k x_{k1} + u_1$. So this is for the first individual, first person. Then I will write the same equation for the second person and it will look like $Y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \beta_3 x_{32} + \beta_k x_{k2} + u_2$ that is for the second person. Likewise, what we will have, we will write the equation for the nth person – $Y_{ni} = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \beta_3 x_{3n} + \beta_k x_{kn} + u_n$. This is how we will write.

Now, this is a system of equations and this is for the nth person. We can write this system of equations in a more compact way. We can take this intercept beta 0 outside. Then what will happen? This would become 1, 1, 1, then we will get $x_{11}, x_{21}, x_{31} \dots x_{k1}$. Then $x_{12}, x_{22}, x_{32} \dots x_{k2}$. So, then we will get $x_{1n}, x_{2n}, x_{3n}, x_{kn}$.

This is your x metrics and that should be multiplied with your $\beta_0, \beta_1, \beta_2 \dots \beta_k$ and $u_1, u_2, \dots u_n$. This is how we can write. Now, we have to think about the dimensions of these metrics. First of

all, this y_1, y_2, \dots, y_n , what is the dimension of this metrics? Here you see? You have n rows but 1 column. So, this is $n \times 1$.

Here, you have how many rows? Here also you have n rows but how many columns? $k+1$. So, that means dimension of this would be $n \times (k+1)$. What is the dimension of this metrics $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. So, that means you have $k+1$ rows but only 1 column. So, $(k+1) \times 1$. And what is the dimension of this metrics? Here also you have n rows but only 1 column. So, $n \times 1$.

Now, if you multiply these two, you see $n \times (k+1)$ and $(k+1) \times 1$, so this would become $n \times 1$. So, this entire thing, $n \times 1$, $(n \times k+1)$ multiplies by $(k+1) \times 1$, when you multiply, the entire thing would become $n \times 1$. When you add a $n \times 1$ with another $n \times 1$ metrics that would also become $n \times 1$.

Then at the end you will see $n \times 1$ equals to an $n \times 1$ metrics. The dimension of the metrics should match otherwise you cannot go for metrics operation. Now, after this what we can write actually in more compact form? $Y = X\beta + u$, where Y is a $n \times 1$ metrics, X is $n \times (k+1)$ metrics.

This beta is basically this metrics which indicates $k+1$ by 1 , that means all $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ they are all denoted by this metrics and this u is again $n \times 1$. Now, if you compare this with the format what we were discussing $y_i = \beta x + u$, you can observe the difference.

Here it is $\beta x + u$ but when you represent this in a system of equation and then you write this in a metrics notation and then ultimately, you express this in a most precise fashion, then you will get $Y = X\beta + u$. This is how you have estimated. Now, what is our objective? Our objective is basically to estimate this and I will show you estimation here.

(Refer Slide Time: 13:11)



Estimation: $\min \sum \hat{u}_i^2 \rightarrow 2 \text{ var CLRM}$

$$U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix}_{n \times 1} \Rightarrow \min U'U \Rightarrow \min (u_1 \ u_2 \ u_3 \ \dots \ u_n) \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix} \Rightarrow \min (u_1^2 + u_2^2 + \dots + u_n^2) \Rightarrow \sum \hat{u}_i^2$$

By minimizing $U'U$, we get

$$\hat{\beta} = (X'X)^{-1} X'Y \quad X^{-1} = \frac{\text{Adj } X}{|X|}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \text{ at one go}$$



Now, in two variable linear regression model, what we were doing? We were minimizing $\sum \hat{u}_i^2$. This is the case for two variable CLRM. Now, in this context when you have multiple linear regression model, we should have a similar expression like this. Here, our u is basically, $u_1, u_2, u_3, \dots, u_n$.

Now, to get a similar expression like \hat{u}_i^2 , what actually we should do? We should then minimize instead of $\sum \hat{u}_i^2$, $u'u$. This is your u metrics, this is $n \times 1$, so you should minimize this. What is u prime u ?

That means this is a column, so this would become your $u_1, u_2, u_3, \dots, u_n$ and that you multiply with $u_1, u_2, u_3, \dots, u_n$. So, this is basically your $u'u$. You are minimizing this. Now, if you multiply this with this, that means $u'u$, that means actually you are minimizing $\sum \hat{u}_i^2$. So, that means this and this now become similar. So, what we minimize here then basically? This is the form we have to keep in mind that we have to minimize $u'u$. And if you minimize $u'u$ we get our desired $\hat{\beta}$.

So, after minimizing this at the end, what we will get? We will get $X'X^{-1} X'Y$, where this X is basically the earlier X what we discussed. This X means this. This is our X and this is our Y . So, this would become $X'X^{-1} X'Y$. And you know what is the, that

means to get beta hat, we need to derive the inverse of this metrics following this, X^{-1} , a simple metrics X^{-1} as you all know, that is adjoint X by determinant X .

So, for getting a solution and to get all the beta hats, this inverse would exist. If for any reason, this inverse metrics does not exist, for example, if this determinants of X becomes 0 for any reason, then you cannot actually estimate $\hat{\beta}$. You have to keep this thing mind, this would be useful later on when we discuss some important things. So, that means from this, if you estimate your $\hat{\beta}$ in this way, then at one go, what we will get? We will get $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$

So, everything will get at one go. We do not have to apply that summation $\sum_{i=1}^k (y_i - \bar{y})^2$ divided by summation $\sum_{i=1}^k (x_i - \bar{x})^2$ formula k times. So, that means in general, our approach would be to represent the equation in metrics notation and then we solve for beta hat following this $X^{-1} X' Y$.

That is the general notation and how we are getting it? We are getting it by minimizing u' . So, you know the u metrics, you take the transpose and minimize this and after this, you will get $X^{-1} X' Y$ where X^{-1} is $\frac{1}{|X|} \text{adjoint } X$, as you all know adjoint X by determinant of X , so that means determinants of this metrics should exist otherwise you cannot estimate your beta hat.

Then, you might be thinking - does that mean that whenever we get multiple linear regression model, we need to first represent this in metrics notation and solve this because as it may look like solving this metrics is also not always easy, because you have so many parameters, then you construct this X metrics, multiply with X , take inverse and again multiply with X' and then multiply with Y . That is also not very easy.

You are getting all these parameters at one go, but you have solve this. Fortunately, we do not have to do anything manually. Rather our statistical software that we are discussing, that is the statistical software Stata, that will help us estimating everything by a single click. Then why we are discussing all this? To understand what is the theory behind this multiple linear regression model. That means what Stata is actually estimating for us.

Unless we understand a theory behind this, what will happen? We will not be able to appreciate what Stata is reporting to us. This is how we can estimate.

(Refer Slide Time: 21:16)

NPTEL

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + u_i$$

$\hat{\beta}_1$: for a unit change in x_{1i} , y_i changes by $\hat{\beta}_1$ amount on an average

$$E(y_i | x_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

$$\frac{\partial E(y_i | x_i)}{\partial x_{1i}} = \hat{\beta}_1$$

for a unit change in x_{1i} , y_i changes by $\hat{\beta}_1$ amount on an average keeping the impact of other factors constant

$\hat{\beta}_1$: net impact of x_{1i} on y_i

Handwritten notes: How do we interpret the impact of other factors on y_i ? or we can't calculate it as we are not sure about both x_{1i} and y_i .

So, once we estimate, lastly what we need to do? We need to interpret $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_k x_{ki} + u_i$. So, when I am writing this model, this is for k variable model, how do you interpret $\hat{\beta}_1$? Earlier, we learnt the interpretation of this model, $Y_i = \alpha + \beta_1 x_{1i} + u_i$.

If you do not have any other variable, then in this context how do you interpret? For a unit change in x_{1i} , y_i changes by $\hat{\beta}_1$ amount on an average. But here, how do you interpret β_1 ? You have so many other factors. That means, to understand the interpretation of β_1 , what I will tell you, you first take the expectation of y_i , then what will happen? Expectation of y_i given x_i , that would become $\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki}$, after expectation. Now, if you differentiate this expectation of y_i given x_i partially, with respect to x_{1i} , what you will get?

This is expectation of y_i given x_i delta x_{1i} , that is nothing but your β_1 hat. Now, you can easily interpret this. That means, if you know the interpretation of partial differentiation from mathematics, you can easily get the interpretation of β_1 hat. This basically says that for a unit

change in x_{1i} , y_i changes by β_1 amount on an average, keeping the impact of other factors constant.

That means, the earlier interpretation is true. But additionally, we have to add that keeping the impact of other factors constant, since this is a multiple linear regression model, if you cannot keep the impact of other factors constant, then actually this $\hat{\beta}_1$, what you will get? You are going to claim that this is actually impact of education, let us say on your wage, but actually, this is not the net impact, this would become a combined impact of other factors constant as well.

When you have multiple variables in your model. So, that is why the explanation would become for a unit change in x_{1i} , y_i changes by $\hat{\beta}_1$ amount on an average keeping the impact of other factors constant, this is the additional thing we need to add in the context of multiple linear regression model.

So, that means $\hat{\beta}_1$, we can say alternatively, in this context is net impact of x_{1i} on y_i . When you have two variable linear regression model, the concept net impact does not arise because you have only one variable. Whatever impact x_{1i} will give you on y_i , that you are getting directly by $\hat{\beta}_1$.

But here, we need to keep the impact of other factors constant otherwise, we will not be able to get the net impact of x_{1i} on y_i . Now, the question that comes to our mind is very challenging. In the context of multiple linear regression model, now we are facing a challenging question. What is that question?

When I am giving you data, let us say I am giving you data on x_{1i} , x_{2i} , x_{3i} and your y_i and I am asking you, you estimate your $\hat{\beta}_1$, you tell me what is the net impact of x_{1i} on y_i , that means how do you keep the impact of other factors constant empirically? That is a very challenging question, is it not?

Keeping the impact, interpretation looks very simple, that keeping the impact of other factors constant, what is the net impact of x_{1i} on y_i , that is given by $\hat{\beta}_1$. But when you are working with

empirical data, the question then is, how do we keep the impact of other factors constant in an empirical set up while dealing with data on x_{1i} , x_{2i} , x_{3i} ... x_{ki} and y_i ?

So, even though the interpretation looks very simple theoretically, empirically it is very challenging because unless you keep the impact of other factors constant, it is really difficult to get the net impact of this. So, what we will learn in our next class? With the data set, then we will learn how to actually keep the impact of other factors constant empirically, that is going to be very very interesting. But that we will discuss in our next class. Thank you.