Spatial Statistics and Spatial Econometrics
Prof. Gaurav Arora
Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

Lecture - 20B
Lagrange Multiplier Test for Spatial Regression Model

So, welcome back to this last in the series of lectures on Spatial Statistics and Spatial Econometrics.

We are going to today look at the Lagrange multiplier test for spatial dependence in regression models. It is a short lecture on hypothesis testing in spatial regression models. This is just an introduction, the idea of this lecture will be to ramp you up to learn about hypothesis testing and regression models or rather spatial regression models, but in no way this lecture is going to be exhaustive in doing that.

So, having gone through this lecture will then sort of enable you to study these concepts elsewhere in detail. So, of course, we are going to look at the two models that we have seen in detail that is the spatial lag model and the spatial error model. And if you remember the spatial lag model looks like this; so, we had y equals rho W y plus X beta plus u.

So, in the spatial error model I will have y equals x beta plus u; such that, u is equal to lambda W u plus e. Where e is multivariate normal with 0 mean and a homoscedastic variance-covariance structure; so, we have seen these things in detail by now. Now, we can just revise this very quickly; so, let us look at the spatial lag model. So, we have y, y is an N by 1 vector which is just data delineated by space or locations.

W is a weights matrix a very precise or let us say a concise container for spatial relationships between the units of analysis. So, W is an N by N matrix y is again N by 1; so, while I am looking at an N by 1 matrix, we have defined this as y_L which is the spatial lag notion of modeling spatial processes just parallel to the time lag right, x is an N by K matrix of K covariates and beta is a K by 1 vector of coefficients, u again is an N by 1 random term.

Error term comprises everything that we could not model in this quest for measuring or explaining the variation in y. The special error model is slightly different, instead of including the spatial effects in the mean outcome model it includes such effects in the error process.

And in the last couple of lectures ago we saw that such models make sense when you have a spatial mismatch between the underlying population process and the data-generating process.

So, if you are studying land use change, which is a more high-resolution process happening at farm levels, but what when we get data if those data are at the district level, then we are going to have an error term miss specification in the sense that we are fundamentally miss specifying or misrepresenting the population process. And that enters the random error term right which we can then specify as a spatial error model.

What is the aim of hypothesis testing? The hypothesis testing wants to figure out whether it makes any sense to include these spatial effects in the models. Or can we just live with the nonspatial models that we are very used to even before we studied this particular course right?

And that will depend on whether or not rho is equal to 0, if rho equals 0, then I am done I do not need to worry about all the weights matrix and all the computation. And all the techniques that I need to sort of imbibe for conducting spatial regressions, but if rho is not equal to 0 then I cannot run away from this situation.

Similarly, here if lambda is equal to 0, I can go back to my traditional analysis something that I know of regressions before even I studied spatial regressions or spatial statistics. And if it is non-zero then everything that we have studied in this course kicks into action and that will lend us our null hypothesis and our alternate hypothesis for this particular exercise.

And the test that we are going to use is going to be a Lagrange multiplier test. So, let us dive right into it and figure out what is the aim of doing all this, and what are the statistics that we are going to use right.

So, I am going to start with the spatial error model. So, I have y equals x beta plus u such that u equals lambda W u plus e where e is multivariate standard normal. Now, standard normal multivariate normal with mean 0 and variance sigma squared and the covariance is between the e errors are.

Of course, now my null hypothesis is that lambda equals 0 and my alternate hypothesis is that lambda is not equal to 0. If I reject my null hypothesis, then I can just say I can work with traditional tools, I do not need all these spatial regressions. So, the test statistic Lagrange multiplier test statistic LM_e is given as u prime W u divided by S squared is just sample variance, right?

So, S squared the whole squared divided by T, this is distributed chi-squared with one degree of freedom. Now, S squared is nothing but u prime u by N, and T is a trace of W plus W prime dot product with W. So, now, let us look at what are we really writing here, it is a lot of notation. Now what is up with the Lagrange multiplier statistics?

So, this numerator this u prime is 1 by N, W is N by N and this is 1 N by 1; so, I am looking at a scalar 1 by 1. The scalar 1 by 1 is divided by S squared, and S squared is nothing but u prime u; so, again 1 by N, N by 1. I have a scalar sitting here, this u prime u can also be seen as a summation i u_i squared. So, it is the sum of squared errors divided by N which is nothing but the sample variance.

But this sample variance S squared is as if there is no spatial weights. So, the numerator here is looking at the sample variance modification when we have spatial weights versus when we do not have spatial weights. This square of this numerator is then normalized by this factor T, you can think of it as a normalizing coefficient and this is a function of the weights matrix.

So, now I have a weights matrix N by N, it transpose N by N, and you multiply it by N by N I have a square matrix of N by N, the trace of a square matrix is just the sum of its diagonal elements. So, whatever matrix that I get here I will sum all the diagonal elements and I will then normalize. So, I have the trace by itself as a scalar because it is the sum of all diagonal elements.

So, I have a scalar sitting in the denominator I have all scalars, and because ultimately here I am comparing the variation or the variance. In the presence of weights and absence of weights, the appropriate test statistic is chi-squared. If you have studied statistical inference before you will know, that even if you are comparing the variances of two different samples use a chi-squared statistic.

So, the Lagrange multiplier statistic is a very natural statistic to study or test the hypothesis that whether lambda is equal to 0 or not equal to 0. So, of course, what I need is a test decision what is a test decision well if LM_e is greater than chi-squared 1.95? So, we say we reject the null hypothesis at 95 percent confidence and if LM_e is less than or equal to the chi-squared 1.95, then you say you fail to reject the null hypothesis at 95 percent CI.

So, if you fail to reject the null hypothesis then you may not want to specify a spatial error model, but if you reject it then the spatial error model is indeed the right specification. If you ignore it then you are going to have inefficient estimates, and you are going to have larger errors in your beta hats than you would have ideally wanted.

So, if you reject the null right; that means, lambda is indeed going to be less than not equal to 0 and is likely to be not equal to 0 with a very high probability. And so, probably you should apply a feasible generalized least squared model, but if you fail to reject the null hypothesis then you can live with OLS, you are fine; so, that is what this null test hypothesis testing exercise is telling me.

So, I am going to do a similar thing for spatial lag models, let me do that for the spatial lag model. We will quickly write it down y equals rho W y plus x beta plus u and now my null hypothesis is either rho is equal to 0 or rho is not equal to 0 much more natural. Just like earlier, I am going to define a test statistic, this time I am going to call it LM_l which is the lag this is going to be u prime W y.

Because now I am explaining the variation in the mean outcome and not in the error outcome. This is divided by again S squared whole square divided by a normalizing constant n J. This is again going to be chi-squared 1, which is chi-squared distributed with 1 degree of freedom and S squared just like earlier is just u prime u over N. And n J is equal to T plus W X beta prime m W X beta divided by S squared where T is equal to the trace of the same matrix that we are used to by now right?

And M is a projection matrix, I minus X X prime X inverse X prime, and then close the bracket, this is a projection matrix, and the idea is the same. Now, the expression is a bit more complicated, but the idea is the same that we have seen earlier ok.

And of course, what we are interested in is the test decision that is if LM_l is greater than chi-squared 1 at 0.95 this can be sourced from a chi-squared table. Then we will reject the null at 95 percent confidence. And if LM_l is less than or equal to a chi-squared 0 at 0.95 which again will be sourced from the table, we will fail to reject the null at a high confidence of 95 percent.

Now, if you reject the null then you cannot avoid a spatial lag model that is indeed the right specification. And if you reject the null and you still go on to ignoring, if you go on ignoring this rho W y factor, then all the beta hertz are going to be biased, and the higher the rho the greater the degree of bias something that we saw through a simulated plot by Luc Anselin earlier.

And if you fail to reject the null then you can go back to your OLS just like we talked about in the previous case. So, ultimately we are also often interested in choosing between these most appropriate specifications. So, this is the starting point. If I am interested in choosing between whether I should run a spatial error model or a spatial lag model.

Then the starting point is to first figure out whether individually they make sense, if say both individually make sense then probably, we should go back and look at a likelihood ratio test. Something that I am not going to talk about here, but you know basically what you are saying is I am trying to maximize likelihood in estimating the spatial lag, in estimating the spatial error I want to see which fits my data better.

So, we have to apply again as I said this module on hypothesis testing is a very basic introduction, it does not do justice to hypothesis testing in spatial regression models at all. But it is a ramping-up exercise, I am trying to empower you to now go back and read the book, and figure out how to conduct hypothesis testing in various settings with spatial regression models.

So, with that I am going to end the lecture modules on spatial statistics and spatial econometrics, going forward we are going to move to tutorials that are hands-on exercises with AGIS and R. So, I want to congratulate you for making it this far and the rest of this course is all your software driven. And it is basically working with actual data and applying all these theories that we have been learning for last almost 10 weeks now.

So, thank you very much for your attention and I will see you on the other end with the hands-on exercises.

Thank you.