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

Lecture - 18C
Endogenous Effects in a spatial regression model

So, welcome back to the third part of lecture 18, we are going to now specify the endogenous social effects in a spatial regression model. And till now, we have learned how to specify spatial dependence and regression model through spatial weights. So, we are going to specify the regression model using the spatial weights and then adapt the Manski reflection problem onto the spatial you know regression model. And then finally, as part of this lecture, we are going to provide the fixes for this purpose.

So, let us think about the groundwater data that we have worked well till now. So, let us say there is the spatial domain of interest as exhibited in front of your screen and I am interested in explaining the depletion of groundwater depth at each location i in the domain of interest at any given point of time.

So, I have a model G_i equal to, I am going to use spatial weights to specify a spillover effect. So, the point that I am trying to make is that all the values in its neighborhood are going to have a spillover effect on the groundwater levels at location i. Of course, the level the degree of spillovers can vary by the location of the neighbors. So, all the dots in this sample will have contributed to the level of groundwater at location i.

So, G_i is written to be equal to rho which is a coefficient of summation j equals 1 to n w_{ij} G_j ok, j being you know the lag. Now, we are going to use you know row standardized weights that is to say that w_{ij} tilde is w_{ij} over summation j w_{ij}. This is just to ensure that summation j w tilde ij will always be equal to 1, right?

So, this is going to be the case plus some other factors you know let us say beta 1. So, I am including the intercept beta 1 x_{1i} beta 2 x_{2i} and so on till beta k x_{ki} plus u_i, this is the regression model that we are used to ok.

In the matrix form, we can write this as G rho W G plus X beta, where X will be an n by k and k beta will be a k by 1 vector plus u which is n by 1 vector, g is n by 1 vector and so is

this one. So, beta and rho are both model parameters; now, rho specifies the effect on location i due to its neighbor j.

Now, the reflection problem will suggest that first of all, what we are as an analyst, an econometrician, a statistician, a researcher, and an analyst. You will be what you are going to exhibit, what you are going to see is that data is exhibiting a pattern, that is to say, that there are going to be clusters where groundwater depth will be very high. So, the groundwater situation will be very bad, let us say, in one cluster it is an urban area you have very low values of groundwater levels.

And towards the east in this direction these groundwater levels are not so bad, you know that they are fine. So, what is happening; obviously, is that if I focus on any value in the north you know west cluster, I am going to have, I am seeing some kind of a spillover going on to this location i. What I do not know is whether or not this location itself is also exhibiting an equal or stronger spillover on all its neighbors.

So, when we see the average neighborhood effect, average neighborhood level groundwaters. So, this is average, it is a weighted average; so, W G is a weighted average; so, its average neighborhood groundwater depth. And G is individual well-level groundwater depth. What we cannot ascertain by this, but by just specifying the model in the way we have is whether or not we can expect to have some effect of individual wells on all the wells in its neighborhood.

This type of intuitive understanding will also apply to housing prices. If I am looking at a home or a house that has a high price and its clustered with other high-priced homes or houses. Then we do not know whether we are seeing the reflection of this house or how the price of this particular house on you know in terms of what is happening in the neighborhood. Or this house itself is simply mirroring what is happening in its neighborhood, who is affecting whom well.

What happens generally in the real world is everybody is affecting everybody else. When we are talking about groundwater data, the groundwater as we have looked at extensively, they are connected through aquifers. And so, long as these aquifers beneath the ground are connected, the wells that we see that we dig from the ground. If they go on into the same aquifer, a cross-section, then the levels that you observe are going to be connected with each other.

We cannot just ascertain simply by specifying G equals rho W G as a single directional or unidirectional effect, there is circularity there which we cannot circumvent, right? And if you think about it this W G which is nothing, but some kind of a weighted average is an analog of the expected value of y given x in Manski's notation.

So, what we are looking at here is an endogenous social effects problem playing out in spatial regression models, and it's quite parallel; so, then one can use Manski's framework. So, if you read Manski's paper, Manski provides constraints for parameters and some conditions for finding instruments to ensure that one can infer upon an effect of the neighborhood scale of outcome onto the individual outcome at every location in a spatial data set.

Now, the next step is reconciling this spatial endogeneity. The first fix that I am presenting to you is coming from Luc Anselin, Luc Anselin is the Professor of spatial econometrics who is perhaps the father of spatial econometrics and provided us with this idea of spatial weights and space weights matrices. Anselin says that weights matrix base spatial lag is not group mean remember, we looked at this idea of a moving window average.

In this moving average what we learned was the difference between a time series moving average window and a spatial moving average window is that wii is 0 that is to say the mean, the center point of this moving average window is not counted. So, what Anselin is trying to say here mathematically is that the effect WG which is the spatial lag, the spatial lag defined using the weights matrix is not equal to expectation y given x.

Is or let us say is not equal to expectation G given sum x, because you are not accounting, a location i is not her own neighbor. So, every location is not its own neighbor, they are not accounted for while calculating this group average, the individual is not counted when we are calculating the group average.

So, the problem of endogenous social effects arising due to this term does not exist is what Anselin's argument is that the spatial lag regression is still identified even though this troubling revelation is due to the endogenous special effects or social effects.

Now, this argument has been critiqued by Gibbons and Overman in a highly cited paper called mostly pointless spatial econometrics. Now, the argument that Gibbons and Overman are presenting is as follows, they are suggesting that causal inference, this idea of ceteris paribus or causal inference is achieved by controlling for economic channels of effect.

In the sense that unless you understand what is the channel of effect, you cannot really ascertain a causal inference. You cannot really get to the point, what is this channel coming from the left-hand side to the right-hand side? And merely using these geographic weights, which are only an index of who is near me in terms of my geography in terms of my location is not going to be sufficient to account for the economic channels of endogeneity, right?

So, if you are not able to account for the economic channels of endogeneity, then a weights matrix-based argument that just because mathematically, you know WG is not equal to an expectation, a moving average expectation especially does not suffice the fact that it is going to now identify causal inference just because we have not quite done that, right?

This is a very good paper, you must read this paper, but the point is that the Anselin argument may not be sufficient. And there are in practice what we find is that people do not just stop at spatial lag models and say they are identified automatically, people do more than that.

And one of the things that is done more is what is called the instrumental variable approach. Here I provided you with an example paper that is using the instrumental variable approach to account for endogenous social effects in a spatial regression model.

I am going to go over this idea of an instrumental variable approach very briefly, because of the because of lack of time. But you can go back and study instrumental variables and instrumental variable approaches for identifying endogenous effects in general not just endogenous spatial effects in the voltages book.

So, I am going to go over the instrumental variable approach very very briefly as follows. So, I am going to provide steps in case steps for the IV estimation approach, step 1, is to find regressors, independent variables, or covariates, that are likely correlated with the group effects. But not correlated with the individual decision or outcome y_i.

So, we want to find variables, we want to seek out some explanatory variables which will affect the group effect. So, group effects are nothing but expectation y given x something that is happening at the group level in a spatial neighborhood or in general in a peer network or wherever.

You want regressors that are correlated with the group effects, but not correlated with the individual decision right. So, let us say we call these variables li's; so, I am using just the term li to sort of say this is one such variable.

Such that the correlation of li with the expectation of y given x, x being G, you can also call it G because it is a group effect. Then this will be non-zero and the correlation of l_i with individual effect is 0 that is to say that l_i will impact y_i only through the group effect. So, we are after this regression y equals alpha plus beta expectation y given in g plus eta z plus u, and I am interested in beta as an analyst.

So, I want to find a variable l which is correlated with expectation by a given g, but uncorrelated with y. That is to say that the only way l will impact y is through g, then I will go to my next step, step two, and I will say run an auxiliary regression also known as the first stage regression.

So, you will run this estimate this regression model even before you are estimating your parent model or the model of interest. The model or a primary model whichever name suits, is favorable in your opinion, right? So, the first stage model will be that I will go ahead and I will regress these average group effects which is nothing but expectation y given g on gamma 0 plus gamma 1 l plus epsilon.

So, I am going to now regress these neighborhood-level variables onto this instrument l, these are the instrument. And then obtain your gamma 0 hat and gamma 1 hat. and hence you are going to get your y bar hat which is nothing but gamma 0 hat, plus gamma 1 hat l_i.

And then in step three, we run this parent regression y_i equals alpha plus beta and we replace y bar g with y bar hat g which is the predicted value in the first stage regression, plus we have eta z_i plus u_i. Step three is called the second stage regression. So, you are now conducting the same analysis in two stages, this is also called the two-stage least squares method right? This is also called AKA (2 SLS method), remember in this lecture I am not doing a good job in terms of explaining what a two-stage least squares method is.

But you have all the tools to learn what it is, and I encourage you strongly to go back to the Gold Ridges book and study what a two-stage least squares method is. It is very general; it is very powerful it is used a lot in econometrics research.

Now, by using this we will be able to identify beta which is the endogenous social effect; so, we are able to reconcile the issue, the reflection problem that Manski is raising because we are not using y bar.

So, I am just defining expectation y given g, I am defining this as y bar g, just to represent group level average of the outcome. Now, the point that I am making here is that instead of using the raw group-level data we use it as a function of an auxiliary variable l_i which does not directly impact y_i. So, we are trying to out the problem of endogenous social effects, we are trying to sort of get away with it.

Because we have not used y bar g in its original form we have also bypassed or we have sort of just avoided the issue of endogeneity, right? So, I am not going to call this an endogenous effect anymore, beta is of course, you know it's going to be the social effect that I am after. So, it is indeed an endogenous social effect; so, the beta hat will represent that.

So, this is the second fix that I am proposing in this lecture, the third one comes from again a very good paper by William Brock and Steven Durlauf. Wherein they are suggesting that instead of using y_i in the left-hand side in their linear form if we use a non-linear form of it that is to say that if we were to model h of y_i as the average level effect.

Then he will not have the endogenous social or reflection problem as provided by Manski, it's mainly driven by linearity. So, I can have alpha plus u, you can read this paper, it is mainly for binary outcome variables, but it is sort of you know also a way to reconcile the endogenous problem.

So, with this, we are coming to a close of lecture 18, but by lecture 18 we have now relaxed assumption A2 and assumption A3. And we have provided ways to estimate the models when these two classical assumptions of the regression model fail for the spatial data sets.

So, thank you very much for your attention, in the next coming lectures we are going to spend more time with the spatial lag model, the spatial cross-regressive model, the SLX model, and the spatial error model. And we will study a bit of hypothesis testing with them, and after that, we will move on to the hands-on exercise with ArcGIS and on RStudio.

Thank you very much for your attention and see you in the next lecture.