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

Lecture - 17A
Spatial Dependence in a Regression Model

Welcome back to lecture 17 of Spatial Statistics and Spatial Econometrics. In today's lecture, we will sort of build on the recap of regression analysis as an overview of the regression analysis that we completed in the previous lecture. More particularly, we will start to relax the assumptions of a multiple linear regression model that ensures that the least squares estimator was the best unbiased linear estimator right?

So, we will start to relax those, particularly when spatial dependence exists in data, and then try and see what are the implications or consequences of spatial dependence on the properties of a least squares estimator. And, in case there are adverse impacts or consequences, how do we sort of go about fixing them?

In time series analysis, the concept of a lag or shift as denoted by y_{t-k} *is the groundwater level observation shifted by k periods in the past.*

$$G_t = \beta_1 + \beta_2 G_{t-1} + \beta_3 G_{t-2} + \ldots + \beta_k G_{t-k} + Z\underline{\gamma} + u_t$$

Parallel $-$ *y we can apply the above notice for a regular lattice in space*

$$Shift\ Up \quad y_{i-1,j} \quad \{G(i,j) = f\left(G_{i-1,j'}, G_{i+1,j'}, G_{i,j-1}, G_{i,j+1}\vec{X}\right) + u(i,j)$$

$$Shift\ Down \quad y_{i+1,j} \quad \{G(i,j) = \sum_{l=1}^{q} X_l \beta_l + u(i,j) \qquad u(i,j) = \eta u(i-1,j) + \gamma u(i+$$

$$Shift\ Left \quad y_{i,j-1}$$

$$Shift\ Right \quad y_{i,j+1}$$

However, there is no analog for the irregular spaces (like for groundwater level data).

This is why we should the notion of spatially lagged variables.

So, let us begin by specifying spatial dependence in the regression model. So, I have started, I have sort of proposed to start with this time series analysis analog where we have this concept

of a lag or a shift. So, in the case of time series regression models or time series models, or time series econometrics, we are typically modeling a variable dependent variable y at time t as a function of its own value in the previous period t minus 1 or the periods before it let us say t minus k in general.

So, you know if I am trying to model, for example, groundwater levels data then at time period t, I have my dependent variable as the groundwater level at time t which is let us say G_t. And, I typically modeling it as a function of beta 1 plus beta 2 G_t minus 1. And, often even going beyond just the first-time lag on to let us say a general value of let us say G_t minus k right?

So, then, I should have beta k plus 1 just for consistency. And, then apart from that I will have some other covariate Z, let us say it's a matrix. And, then we have a vector of coefficients gamma which are non-ground water levels data. I mean you can have data on agricultural activity, on policies, on the aquifer structure, on weather or climatic conditions and so on and so forth right?

So, this idea of a time lag in the case of a time series regression model is now extended to this idea of a spatial lag in the case of a spatial lag model, right? So, in case of spatial lag what we can think of doing is that we can have you know lags especially in the northward direction, in the southward direction, in the westward direction, and the eastward direction, that is shift up, shift down, shift left and shift right.

So, corresponding to G_t, now we do not just have G_t, but we have G i comma j where i and j are nothing but x y coordinates right? So, instead of just working with one index, I work with two indices because groundwater data are observed in a two-dimensional real space. This is then modeled as a function of the neighborhood groundwater levels, which is where the spatial spillovers will come from the right.

Similarly, for prices of houses some examples we saw last time, you will have a spillover from the nearby community or the homes located in the nearby community. So, to specify spatial dependence, I will definitely what I do is I include these shift up, shift down, shift left and shift right variables along with some other variables X or Z. Those are the control variables plus the error term u_{ij}.

Now, spatial dependence can be modeled as a function of G_{ij}, but also as a function of the unobserved factor. So, that is to say, I could just have included G_{ij} is equal to summation l equals 1 to q X_l beta l plus u_{ij} where u_{ij} itself can be modeled as a function of eta u_i minus 1_j plus gamma u_i plus 1_j and so on and so forth plus an error term epsilon ij.

Now, spatial dependence in this second type of model is modeled and specified through regression errors, that is to say, some systemic factors will affect groundwater levels; they may or may not have spatial dimensions. So, $X_{l's}$ may or may not be spatially delineated. For example, the intercept is coming from this column of 1s. Now, that column of 1s is not quite spatially variable, right?

I mean it is the same at every location and its value is just 1 right? So, some of the exercises, on the other hand, might be spatially delineated for example, the amenities right; public schools, roads, the crime rates in that locality and so on and so forth will be spatially variable, right? But, the error term which is all of that we could not include in the forms of X when explaining the variation of you know G_{ij} is exhibiting spatial dependence right?

So, the spatial dependence can be specified through lags of the variable itself or the lags of the error structure. Now, this type of specification; however, depends on you know what is called a regular lattice, that is to say, that I must have data that is distributed spatially in these regular lattices. So, if I am at any point i comma j, I should be able to shift up, shift down, shift left, and shift right and exactly find one observation in all these directions.

What we have seen in the past is that real-world data sets do not work like that. They are irregular lattices; you may not be able to find an opportunity of shifting up oftentimes. For example, in the northward boundary of Uttar Pradesh data, you do not have any scope of shifting up. You may not even have scope for some in shifting right or shifting left-right.

So, in those cases such definitions will not be sufficient; that is why we specifically study this notion of spatially lagged variables. So, we will study this notion of spatially lagged variables formally and this is where the weights matrix and all those concepts start to arise.

First, let's study the consequences of spatial dependence on inference in a spatial regression model

Suppose we have housing price data, $P\left(\vec{S}_n\right)$ *or* $P(i,j)$ *or* P_{ij}, *where i and j represent*

$location\ coordinates\ on\ x\ -\ axis\ and\ y\ -\ axis,\ respectively.$

$$P\left(\vec{S}_n\right) = \sum_{l=1}^{q} \beta_l X_l\left(\vec{S}_n\right) + \delta\left(\vec{S}_n\right); \quad \vec{S}_n \epsilon \left\{\vec{S}_1, \vec{S}_2, \ldots, \vec{S}_n\right\} \equiv DcR^2$$

$\left\{x_l(.):\ \ l = 1, 2, \ldots q\right\}$ $is\ a\ collection\ of\ q\ non\ -\ random\ explanatory\ variables$

$that\ may\ or\ may\ not\ depend\ on\ location.$

$\delta\left(\vec{S}_n\right)$ $is\ zero\ -\ mean,\ finite\ variance\ regression\ error\ proves\ that\ exhibits\ spatial$

$dependence$

$$E\left(\delta\left(\vec{S}_n\right)\right) = 0$$

$$Cov\left(\delta\left(\vec{V}\right), \left(\vec{V}\right)\right) = \sigma^2 \rho^{\|\vec{u}-\vec{v}\|} \ ; \ \ \rho\epsilon(0, 1)$$

$\neq 0 \Rightarrow we\ have\ relaxed\ A3\ (refer\ Lecture\ 16)$

$\{when\ \bar{u} = \bar{v} \qquad then\ Cor\left(\delta\left(\vec{u}\right), \delta\left(\vec{u}\right)\right) \qquad = V\left(\delta\left(\vec{u}\right)\right) = \sigma^2 (\sigma\ const)$

$whenever\ \vec{u} \neq \vec{v}\ than$

$$Cor\left(\delta\left(\vec{u}\right), \delta\left(\vec{v}\right)\right) = 0$$

$Errors\ are\ spherical$

$Homoscedasticity\ assumption$

But, before we get there, let us study the consequence of spatial dependence on inference in a spatial regression model. Let us study the consequences of spatial dependence on inference in a spatial model. So, we are going to start with an example. We are going to say suppose we have a housing market, where we have data for such a market.

So, we have housing price data which we denote as P at location vector Sn right? This could be alternatively written as P at i comma j or even P_{ij}. All of these are alternative analogous, you know notations right where i and j, i and j represent location coordinates, represent location coordinates on x and y axes respectively right? So, we are writing a model P S_n equals summation l equals 1 to q beta l x l S_n.

Remember, I am writing $x_{l's}$ as if they are spatially delineated, but many of them might not be. They might not be variable by space. So, there you know you have the index i comma j,

but the value is the same; plus delta S_n such that Sn by itself belongs to a set or domain of interest d which consists of these locations S_1 till S_n.

This is the definition of my domain which is in the two-dimensional real space xl such that l equals 1 to q is a collection of q non-random; remember when I say non-random, I am actually talking about assumption 5 of the regression analysis in the previous lecture; that axis are considered to be non-stochastic.

So, when I am specifying the model, I am actively suggesting that is a non-random explanatory variable, that may or may not depend on location, depend on or you can say vary by; it is, all the same, may not depend on location or may not vary by location. Delta S is my regression error. So, I am going to say this is zero mean, finite variance error process.

You can also call it a regression error, just to be very clear that we are talking about the error term in the regression model. Something we had used the notation u_{ij} consistently, but till now we are just using delta just for the sake of being flexible in notations. I think Cressy mostly uses this notation. So, that is my guess why I have this in my notes. So, now this error process exhibits spatial dependence.

So, all the spatiality I am assuming the model that I am studying here is carried forward through the error process. The prices may be just dependent on each other locally. And, because I did not include lags of prices themselves in this model or Excels are not lags of prices, all that effect is now sucked into u or delta S, which is the error term, which is giving me everything that did not include the model.

So, it is possible that the delta is exhibiting spatial dependence in accordance with the price variation locally or the local dependence, local spatial dependence in prices. But, it is also possible that some of the x's might exhibit spatial dependence. For example, if I look at school quality, it is possible that good schools are clustered together.

So, then if I am including xl which let us say you know represents school quality or a certain type of a public community, it might have its own spatially dependent variables. And, because I did not include those in my regression model on the right-hand side, they are also sitting in delta, right? So, delta by itself is exhibiting spatial dependence of complex types.

It can come from the prices that is the P S_n, you know local clustering there or and slash or local clustering in one or two more of the x, you know explanatory variables that may exhibit spatial dependence. So, mathematically if we say that we have an expectation of delta S_n equals 0 and we have the covariance of the delta at location u and at location v equals sigma squared rho to the power u minus v, that is the distance or the spatial lag that determines the distance between the locations u and v right.

Now, rho is between 0 and 1. So, it is not equal to 0; that means that the covariance structure is non-zero. So, this is non-zero which implies that we have relaxed assumption 3 in the traditional regression analysis, I would say refer to lecture 16.

So, if you refer to lecture 16, you will now realize that we have relaxed assumption 3, where assumption 3 required that had two conditions, that you know when u is equal to v. Then, the covariance of delta u comma delta u which is nothing but equal to the variance of delta u is equal to sigma squared which is a constant and does not depend on the location right?

So, this is a constant. The second thing it required was that whenever u is not equal to 3 v, that is we are standing on two different locations in space; then this covariance of delta u and delta-v is equal to 0. So, this is what A3 required right? A3 entailed that the covariance of errors at two different locations is 0.

And, whenever you are at the same location u and v are the same, that is the diagonal element of the variance-covariance matrix that is the same throughout this you know the data sample, right?

So, this amounted to what we called errors are spherical. We also say that this assumption is also called the assumption of homoscedasticity right? Now, that we have that covariance is not equal to 0, we have covariance not equal to 0 implies that A3 is violated. So, now, we have established that whenever spatial dependence exists through the error terms in a regression model, assumption A3 which is the assumption of spherical errors, homoscedastic errors is violated.

So, as a next step what we are going to do is, we are going to work with a simpler example, a simple linear regression model. And, figure out the consequence of this spatial dependence on the estimation of the least squares, you know coefficients of the regression model.

$$\textit{Let us work with a simple linear regression model}: \quad q = 2$$

$$P\left(\vec{S}_n\right) = \beta_l + \beta_2 R\left(\vec{S}_n\right) \quad + \delta\left(\vec{S}_n\right) \; ; \quad \vec{S}_n \epsilon \left\{\vec{S}_1, \vec{S}_2, \ldots, \vec{S}_N\right\} \epsilon R^2$$

$$Cov\left(\delta\left(\vec{u}\right), \delta\left(\vec{v}\right)\right) = \sigma^2 \rho^{\|\vec{u}-\vec{v}\|} \qquad \rightarrow Relaxed\ A3$$

$$x_1\left(\vec{S}_n\right) = 1 \quad \forall\ \vec{S}_n \epsilon D$$

$$x_2\left(\vec{S}_n\right) = R\left(\vec{S}_n\right)$$

$$\#\ of\ rooms\ in\ house\ located\ at\ \vec{S}_n = (i_n, j_n)$$

$$\textit{In Lecture } 16 \textit{ we introduced the Least Squares algorithum where we minimize the sum of square errors and recover data} - \textit{driven estimates } \hat{\beta}_{1,LS}; \hat{\beta}_{2,LS}$$

$$min\ \beta_1, \beta_2 \qquad \sum_{n=1}^{N} \left(P\left(\vec{S}_n\right) - \beta_1 - \beta_2 R(S_n)\right)^2$$

$$Upon\ solving,$$

$$\hat{\beta}_{2,LS} = \frac{Cov\left(P\left(\vec{S}_n\right), R\left(\vec{S}_n\right)\right)}{V\left(R\left(\vec{S}_n\right)\right)} = \frac{\sum_{n=1}^{N}\left(P\left(\vec{S}_n\right)-\bar{P}\right)\left(R\left(\vec{S}_n\right)-\bar{R}\right)}{\sum_{n=1}^{N}\left(R\left(\vec{S}_n\right)-\bar{R}\right)^2}$$

$$V\left(\hat{\beta}_{2,LS}\right) = \frac{\sigma^2}{\sum_{n=1}^{N}\left(R\left(\vec{S}_n\right)-\bar{R}\right)^2}$$

So, we are going to work with, you know, I am going to say let us work with a simple linear regression model. So, when I say that we are working with a simple linear regression model, all I mean is that I am going to set q equals 2, that is I have a model of the price which looks like P S_n equals beta 1 plus beta 2. So, the coefficient beta 1 is coefficient to just 1.

So, x_1 is just 1 at all locations S_n; plus beta 2, I am going to keep this as R S_n, just to keep continuity from our previous lecture where I modeled the price of a home as an index of spaciousness or the number of rooms, that are available in that home or house. Plus the error term delta S_n such that S_n belongs to S_1 S_2 keeps going till S capital N, which is in the two-dimensional real space; just to keep my notation consistent.

You are encouraged to write these things down with your pen and paper. You know, although when you look at these things, it might look like they are flowing naturally. But, trust me when you actually do them by yourself, you will get stuck and that is where the opportunity to learn comes. So, please do not just look at the screen and let you know see me do it.

You should definitely write these things out at least twice on your own, you know apart from the lecture itself. So, that is when you will start to really get a sense of what are we doing in these you know, how the math is flowing, how we are interpreting it with English and vice versa. With that, I am saying that the covariance of delta u and delta-v is equal to sigma square rho u minus v right?

So, we have relaxed A3, right? We are also saying that x_1, if I go by the notation that I had used earlier was equal to 1 for all $S_{i's}$ or $S_{n's}$ in D. And, x_2 is now specified to be the number of rooms in the house for which we are studying, the houses that we are working on, the real estate price model for right? So, this is nothing but the number of rooms in the house located at S_n which is nothing but an x and a y coordinate.

So, we can say i_n j_n. So, these are just notations, right? I mean do not get stuck with notation or do not start worrying about notation so much, that we do not understand. We need to understand what we are saying here right? Notation is fluid, we can use different notation, you can use you know not i comma j, you like k comma l; go ahead use it.

You like m comma n use it, a comma b whatever right; that is not the point. The point is that we need to understand that some indices are representing location, the location is being represented with a two-dimensional real space. So, it is the x, y coordinate, if it were three-dimensional, we will have the x, y, and z coordinates right? So, the spatiality will become more and more complex, but these tools will remain consistent throughout. So, that is the point.

So, I am going to say in lecture 16, we introduced and went over the least squares algorithm, where we minimize the sum of squared errors and recover data-driven estimates beta 1 hat least squares and beta 2 hat least squares right? So, the hat is a notation that this is a data-driven answer. This is a number that we have backed out from the data and the algorithm that we have applied is minimizing the sum of squared errors.

This algorithm is intuitive, where ultimately we are trying to fit a model, find a very good as good a fit as possible in terms of modeling you know P prices, housing prices. And, we want as we do that we want to minimize the error. Now, one can say why do not sum the error and minimize it. Now, the error could be positive or negative.

If you do not square them, what is going to happen is the positive errors are going to cancel the negative errors. But that is, just because an error is positive and in nature and the other one is negative in nature, does not mean that they are actually supposed to be canceled. They are supposed to be both penalized for being erroneous at the time of you know in terms of prediction. So, to penalize both positive and negative errors, we square them right?

So, once we square them both positive and negative values convert into a positive value, moreover as u moves away from 0, the penalty is increasing. So, you are penalizing more and more as the error is departing from 0 either on the positive or the negative side. So, squaring the errors that are adding the positives and the negatives and then penalizing them by an exponent of 2 as they depart from 0 which is my point of desirability that errors are 0.

As they depart from there, I penalize them by an exponent 2, then I sum them together which is the sum of squared error. Somehow, the variation that I could not explain in the model. I want to minimize, I want to choose beta 1 and beta 2, such that this u_n modeled variation is minimized; given the model specification, the systemic portion of the regression. So, once I have this, what I am trying to do here is, when I say that I am going to minimize the sum of squared errors.

All I have said is that I minimize summation n equals 1 to capital N P S_n minus beta 1 beta 2 R S_n squared and while I do that, I choose beta 1 and beta 2 right? Upon solving, solving means writing the first-order conditions. We have two first-order conditions, two equations, and two unknowns we can solve them. Upon solving, we get our beta 1 hat least squares as follows. It is going to be covariance between the price variable, and the dependent variable and so, this is beta 2.

So, this is the coefficient on the spacious number of rooms. So, the coefficient that we back out from data is going to be how well the prices and number of rooms are co-varying. So, the covariance of P S_n and R S_n is divided by or normalized by the variance of R S_n. This is a typical popular you know when I teach econometrics, I ask my students to learn this by heart.

This is one thing they should learn by heart, other things they can you know they do not need to learn anything. But, this is one of those things that you know even if I wake you up at 3 am and I ask you what beta 2 hat OLS which is the coefficient of x when we are looking at the impact of x on y? It is going to be the covariance of x and y, that is how well are they covarying, normalized by the variance of x itself right? So, this is the formula and you know it turns out to be looking as the following.

n equals 1 to N so, we will write the covariance formula that we know from our basic statistics course divided by n equals 1 to N R S_n minus R bar, the whole square. Now, very interestingly you have to see that you know when we defined when we sort of specifying the model, we said that $x_{I's}$ are assumed to be non-random or non-stochastic.

So, in the formulation of beta 2 hat OLS, you have R which is an analog of the exponent or variable which is going to be assumed to be non-stochastic. So, it is a deterministic variable. There is no, it is not that R will take multiple values and there are some you know when we represent R, we do not really say that number of rooms is 3 plus minus something. We say the number of rooms in a house is 3, exactly 3 or 4 or 2 or 1 or whatever right?

So, this number is exact and it is non-stochastic, right? It is a degenerate variable. The price of the whole house on the other hand is considered a random variable right? As a statistician or econometrician, I begin by saying that I am viewing the world as stochastic and I am going to try and explain it with some systemic portion which is beta 1 plus beta 2 R right?

But, inherently this process is overall random in nature and all that randomness is now you know in the regression model sitting with the error term. But, having said that the fact that beta 2 hat OLS is a function of P which is considered as a random variable. This means that the estimator by itself is a random variable; that is to say that when we look at the estimator, it is not going to be a point estimate only.

But, also there will be a precision metric for it, that is it could by itself is a random variable right? So, this is a random variable; which means, it can take multiple values with different probabilities. And, that implies that I should be able to draft a precision metric for this beta 2 hat OLS. And, indeed a precision metric that is the variance of this beta 2 hat OLS how to vary and vary this, you know how precise is this beta 2 hat OLS as a point estimate.

And, in terms of what is possible for beta 2 hat OLS, what are the extremes that are possible, with let us say 95 percent probability or something like? So, it is given as a sigma square which remembers is again a model parameter, in the covariance, variance-covariance structure of the error term. So, this covariance is called the variance-covariance structure of model error.

So, the variance of beta 2 hat OLS is given as summation sigma squared over summation n equals 1 to capital N R S_n minus R bar the whole squared. These are the results that you are going to get. This is the result that you are going to get when you apply the minimization of this sum of squared errors in the presence of spatial dependence. Now, if we had assumed no spatial dependence, we would get the same results. So, then what does spatial dependence really do right?

So, the next step is to evaluate the implication of spatial dependence in model errors on the least squares estimators. So, we will have to evaluate the form that we see beta 2 hat OLS, that is the covariance of P and R over the variance of R.

And, the variance of beta 2, beta 2 hat OLS which is sigma squared over the variance of R. How does this, you know what is the implication of spatial dependence on these? Because, if we had no spatial dependence, we will still get the same beta 2 hat OLS and the same variance of beta 2 hat OLS. So, what changed right? What is it that we should worry about then? That is what, we are going to look at in the next step and we will come back to it in the next part of this lecture.

Thank you.