Spatial Statistics and Spatial Econometrics

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

Lecture - 16A
Spatial Regression Analysis

Welcome back everyone to the 16th lecture on Spatial Statistics and Spatial Econometrics. In this lecture, we will make a formal break from spatial statistics and start to focus on spatial econometrics. Within spatial econometrics, we will begin with what is called the Spatial Regression Analysis.

Spatial regression analysis will provide us with an opportunity to conduct multivariate statistical analysis. What does it mean? We will see in a minute, and we will sort of you know kick off this module, we will do a review of the traditional regression analysis for all of you who may not have seen or studied regression analysis earlier.

We will go over the basic material and I will sort of you know provide you references which you can go back and read if in case you want to learn any of these subtopics in traditional regression analysis in depth. We are not only covering this review of regression analysis which is non-spatial in nature for just sort of doing a recap of it, but also this recap will provide us with an opportunity to learn, where exactly spatial regression analysis departs from regression analysis.

Just like in the case of statistics you know we always started with a traditional statistic and then we found points of departure from there, right? So, we will look at crucial assumptions of traditional regression analysis, we will study this pathway from correlation or association to causation and also we will look at the impact of heteroskedastic errors on the least squares estimator.

So, we will talk about these things if you have not heard of these terms, there is nothing to worry about, but this is very crucial to also learning where spatial dependence shows up in regression analysis. Then, we will move on from this non-spatial regression to spatial regression, where we will look at the specification of spatial dependence, what are the alternative ways of specifying spatial dependence and regression models.

Then, we will study the estimation of model parameters in the presence of spatial autocorrelation or spatial dependence and finally, we will think about causal inference in a spatial regression, right? So, we will review these things. Towards, the sort of last sub-module within spatial econometrics will be about this idea of spatial lags to account for spatial dependence.

It is a convenient specification for specifying spatial dependency regression models and it allows us very powerful interpretations of how spatial spill over's can impact the processes, the random processes that we see around us in the real world.

Within that we will talk about something called a spatial weights matrix, we will look at some of these spatial dependent summary statistics specifically Moran's I, the Geary's C which is based on the definition of this spatial weights matrix, and then we will study regression analysis in presence of spatial lags and finally, we will introduce hypothesis testing we will ramp you up and then of course, there is a large literature to be explored at your own time, ok.

So, let us get started with a review of regression analysis, alright. So, when I do that, you know what I want to do is I want to sort of start with a mathematical formulation of a regression model and also a graphical representation of a regression model. So, among the regression you know regression within regression analysis, we have something called the simple regression, simple linear regression model, ok.

The simple linear regression model goes as follows it is y equals beta 0 plus beta 1 x_i plus u_i, where i grows from 1 to n. So, we have a data set of $y_{i's}$. These $y_{i's}$ could be you know groundwater levels something that we have looked at quite extensively throughout this course, but also we could now sort of, think of something a newer example let us say the price of homes, right?

I mean when we started this lecture series we studied or looked at these you know spatial real estate phenomena for the National Capital Territory of Delhi. And at the time, I had said that we want to be able to explain what goes into this phenomenon as it comes up in space as it shows up in space, right? So, from that standpoint going forward in this module we will take the example of you know home prices.

So, let us say we take an example let us specify an example of home prices or house prices. So that means, that y_i is defined as some price value of a home i^{th} home in a given community or in a city or a state or whatever ok; and we have n such homes tool study and these homes as we view them, the prices of these homes as statisticians and econometricians, we are going to view them as random processes, right?

So, we are going to say that y_i or let us say we can say P_i is represented by a probability density function f of P_i right? it could be a normal distribution, it could be anything. If it is a parametric distribution you know you will have a parameter the density parameter you know f of theta, right? But, the idea is that P_i can take multiple values, right?

Each home each i^{th} unit of analysis can take multiple values as far as their price is concerned and it is going to be drawn from this PDF f of P_i, right? So, there is some variation in the $P_{i's}$ and the regression model that it wants to do it, wants to explain this variation using a variable x_i, right? So, y_i by itself is called a predicted dependent you know or a response variable.

We want to explain the variation that is you know the PDF that is the true PDF you want to explain it is properties using a covariate x_i, right? Now, when I said covariate all I meant was x_i by itself is called an independent variable, it is also called an explanatory variable right? it is a variable that explains variation in P_i or y_i, right? It is also called a predictor variable, it is called a control and it is also called a covariate of y_i, right ok, alright?

So this particular portion of the regression that is beta 0 plus beta 1 xi if you realize it is a systemic source of variation in y_i right? it is a systemic source of variation in y_i. Why is it called systemic? Because it is a tangible source of variation. So, if x_i varies a little bit more P_i will vary a little bit more, but I can measure that change, right?

I can observe and measure x_i, I have data for x_i just like I have data for y_i, right? So, what would be a good example of you know x_i, x_i would be you know a property of the home let us say how spacious it is. We can include an index of the spaciousness of a home or a house by looking at the number of rooms in that house, right? So, we can say x_i is the number of rooms in house i, right?

So, that means, we can say this is R_i, right? So, we are writing a model where we are trying to explain variation in P_i using a separate variable R_i, right? So, there is a systemic portion, a systemic component of variation that is you know going to be the one which is our modelled

variation, right? And what remains, which we could not explain through x_i is called the model error which is called the disturbance or error term, ok.

This is a random source of variation, right? Now, a random source of variation is critical to regression analysis, right? Why is that well, it is because the way I sort of articulated, we say ok as an econometrician if someone tells me can you explain the variation in P_i in a given community? why are some homes of less price, why are some homes of a high price you know and so on and so forth?

Why do we see the type of spatial data patterns or simple non-spatial data patterns in terms of why are all homes you know in the world not priced equally? Well, you know it can then be explained by a systemic portion, but as an econometrician, I must view the prices of homes as random variables. Now, where is this randomness going to arise from in my regression model, it is going to be embedded in this term u_i, right?

Whatever is x_i that is the number of rooms in a home that is you know that variable is non-stochastic or non-random in nature? When I say that a home has 4 rooms or 3 rooms, I do not mean that it is 3 plus minus 2 right? I do not mean, it is 4 plus minus you know 0.2, right? I mean 4 exact 4 rooms when I say R_i equals 4. So, that is you know a degenerate form of explaining the variation in P_i, right?

And the source of randomness is this disturbance or the error term which my model could not explain. So, we cannot be perfect in explaining every real-world phenomenon, right? So, in that spirit, you know the regression model has two components one is called a systemic portion which is the modelled portion. The modelled portion has an independent variable and it has what is called the model parameters, right?

In this case, we have the intercept that is beta 0 and we have the slope that is beta 1. Why are they called, what they are called let us look at that term below, ok. So now, let us conceptualize these data. So, we have a data set which comprises columns Pi and Ri. So, I have data, right? So, I have prices for let us say, I have i_s or the ids of homes and I have let us say data for 1000 homes where I have for each home a price level P_i and a corresponding number of rooms or the space over which this is built.

Let us say indexed by R i, ok. Then with these data, I can definitely conceptualize the scatter plot of P_i versus R_i. Let us say that the scatter plot for these 1000 data points looks like the

following. Of course, this is not a perfect plot, but I am trying to try and plot it, ok. So, let us say the number of rooms could be 1 could be 2 could be 3 could be 4 could be 5 could be 6, lets us say maximum there are about 8 rooms homes there are huge homes that we are looking at.

Now, what a regression is doing. The systemic portion of the regression is saying that this P_i can be explained by a real line a real number line which has an intercept of beta 0 and a slope of beta 1 right? this is a regression line, right? But, for every given x_i let us say when x_i equals 3 if I go and look at the graph you know the model will give me beta 0 plus beta 1 times 3 as the systemic component of P_i.

This is the modelled version, the regression model version of P_i. The true P_i, however, is not the same, the true P_i for this one lies either here or it lied you know here. In that case, this distance between the truth and the modelled version is my error representation, ok. Remember also that here beta 0 and beta 1 are data-driven you know understanding of what the intercept and the slope should be.

So, if I have data-driven representations, I use this terminology hat, ok. If I estimate or predict these parameters you know to be predicted or estimated you know, once I estimate them using the data, the intercept and the slope are called beta 0 hat and beta 1 hat, because they are no longer the true model, they are no longer the parameters of a true model.

But they are the data-driven you know estimates of what beta 0 and beta 1 would be if I were to believe that y_i is indeed going to follow the true model that I have specified as an analyst, ok. So, this gives you a gist of what a regression model is it. At the end of the day, it is a linear regression equation, right? when I say linear it means linear in parameters, right?

If I were to include x_i squared here it will still be a linear regression model, because it is linearity between y_s and betas the model parameters and the dependent variable. So, the linearity is not between y and x in a linear regression model ok, it is very important as a review you know from a review point of view ok, alright? So, this gives you a gist of what a regression model is, how it views the world, and how or what are we trying to really explain here.

We are trying to explain the variation in prices that we view in real-world data sets, the fact that all these prices are different for different homes, can be ideally unique for each home.

But, then what explains this? Well, one thing that could explain is how big is the home that we are trying to price in the market, right?

So, that is the way to sort of conceptualize it, there is a regression equation, the regression equation has a systemic portion a non-stochastic portion and a stochastic or a random portion or a component, right? So, u_i is random in nature, at every x_i, u_i can be a different value, right? And the fact that y_i was viewed by an econometrician as a random variable is contained in this error term which is u_i.

So, a part of it is deterministic fine, but a part of it which remains, which I could not explain with xi is you know explained by u_i which is the error term, but the error term is random in nature, right? So, that is the basic premise of a regression model. Now obviously, the linear regression model is a bit restrictive you know it only has one regressor now.

The price of a home is not only explained by the number of rooms. It is also explained by things like amenities around right? Pub, number of parks in the community, right? What is the quality of schools you know in that community? What is the state of public infrastructure roads and you know crime control or kind of security safety and so on and so forth in that community?

What is the access to a local market for the convenience of residents of that community? Many factors drive how a home is priced eventually in a market, right? Now, a simple linear regression model is one which only has one covariate in this case it is x_i or R_i if y_i is P_i, right?

Now, you know to be able to relax this restriction we have what is called the multiple regression model, ok. Now, you know the multiple regression model allows me to include more than one covariate, that is it allows me to include you know a general number of many covariates as many as one analyst thinks that will explain the variation in y_i ok; again u_i has the same representation, right?

So now, you have, your systemic component is more complicated right, but the error term is still you know exactly the same. Although it is real-world formulation will be different, because now when we included only 1 x_i let us say x_{1i} what we did was we put all the other $x_{i's}$ x_2 to x_k into u_i.

When we included x_2 to x_k now u_i has whatever is not included from x_{1i} till x_{ki}, right? So, whatever we exclude from the regression you know simple linear regression, it is all in u_i, right? Now, obviously, you know it is a multiple linear regression model because the linearity is between y and all the betas. So now, if x_{2i} was for example, x_{1i} squared and x_{ki} was let us say x_{1i} you know to the power k, then you know you will still have a linear model, right?

It does not matter how non-linear is the formulation of x, but what matters is you know the fact that the model parameters are still linear in y, ok.

So, that is a note, alright. The other thing that I want to sort of talk about, going forward is that a regression model links the mean value of y_i which is the dependent variable with the mean value of x_i which is the covariate. What does that mean? Well, that really just says that let us say I begin with my simple linear regression model SLRM, I have y_i equals beta 0 plus beta 1_{xi} plus u_i.

Now, if I were to take an expectation on this you know on both sides of this equation, I am going to have the expectation of y_i equals the expectation of beta 0 plus beta 1_{xi} plus u_i. Now, the expectation operator is a linear operator so it will start entering and applying to each term.

Here beta 0 and beta 1 are constants, they do not vary at all even across i as they do not vary. So, I have beta 0 plus beta 1 expectation of x_i plus the expectation of u_i. Now, the expectation of u_i is assumed to be 0 without loss of generality, I will just come to what it means by this assumption of without loss of generality, but the idea is that now what you see here is that the equation.

Once we assume the expectation u_i to be 0, this equation is really linking the expectation of y_i with the expectation of x_i which given the n values are nothing but the sample average of y_s and the sample average of x. So, what happens is that at the end of the day, a regression model is a model on the mean, right? It is linking the mean value of x_i with the mean value of y_i, ok.

Now, it can happen for you know, if I go back to my graph, what is happening is that this x regression line is nothing but the expectation of P_i, the expectation of P_i equals beta 0 plus beta 1 now these are hats these are data-driven slopes and intercept right beta 1 hat expectation of R_i, right? That means, that the average of you know R_i. Let us say the R bar wherever it touches this line will also refer to the y bar value.

And all this is doing is that at every point in x_i on x_i, I am actually going and figuring out what is the average y value, ok. So, it is a model on the mean, if I keep x_i as you know variable then, it will be y bar to be figured out at every x_i. Otherwise, the y bar and x bar will pass through a regression line, right? Now, so this is the representation of a regression.

A regression line is a model of the mean values of an estimator of the dependent variable and the independent variable. Now, while we explained this, we introduced an assumption right away on which we had the definition of the regression line is based. The fact that a regression line is a model on the mean depends on the expectation u_i to be equal to 0. And I have said that we are calling this a without loss of generality assumption, so what is a without loss of generality assumption?

To move forward, I am going to start talking about the crucial assumptions of a linear regression model, ok. So, the first assumption that we are going to talk about is the expectation u_i equals 0 which I am calling a without loss of generality assumption when beta 0 is included in the linear regression model.

That is to say that, when the intercept is included then, the expectation u_i equals 0 without loss of generality assumption what does it mean? So, let us sort of see what it means. So, we have y_i equals beta 0 plus beta 1_{xi} plus u_i. Now, the expectation of y_i equals beta 0 plus beta 1 expectation of x_i plus the expectation of u_i. Now, say that the expectation of u_i is equal to a bar.

Note or notice that as long as beta 0 is included we cannot differentiate between it is self and a bar right? we cannot differentiate between beta 0 and a bar, right? So, if we assume a bar equals 0 then, you know any and all of the impacts of this assumption were not true. Let us say we assumed a bar equal 0, but it was non zero all of that impact will be captured by beta 0, right?

If indeed this was in the non-negative number you know you could simply think of beta 0 as beta 0 plus a and then, you can write expectation of u_i equals 0, right? So, then you know any and all of its impact will be captured by beta 0, even if it were you know even a bar was non-zero. Hence, the expectation of u_i equals 0 is known as a without loss of generality assumption, ok.

Now, you know we can represent this assumption in vector format. So, we can see that now so, A 1 in vector form is written as an expectation of u equals 0. So, u is now a vector which is n by 1 size, because we have $u_{i's}$ going from u_1 to u_n. So, this is a column vector just like an Excel sheet column going from u_1 u_2 all the way till un. Similarly, the 0s are also n by 1.

Now, I can rewrite this as an expectation of u_1, u_2, and u_3 all the way till un equals 0 0 0. Now, the expectation is a linear operator just like, just enters the vector as if it were a multiple lambda right? it is just like if it were a constant lambda. It would simply enter and what will happen, will apply to each element of this vector.

So, this means I am talking, what I am saying here is an expectation of u_1 and expectation of u_2 to keep going till expectation of un all of these will be 0s, right? So, all I am saying is that expectation u_i equals 0 for all I going from 1 to n ok, now that is it. So, from here to here what I am using is that expectation is a linear operator we have used this property earlier in this course as well, but I hope this will make things very very clear.

So, this is a first crucial assumption of a linear regression model. Thankfully when the intercept is included this assumption is a without-loss of generality assumption.

So, let us move on to the 2^{nd} and perhaps the most crucial assumption of the regression model, right? I am going to call it the most crucial assumption and we will see why in a minute. I am going to write down this assumption as the expectation u_i has given x_i is 0. What this means is that not only overall the error should you know the sum to be 0, but for every given x_i the errors must sum to 0, right?

If I go back to my scatter plot between P_i and R_i right? if I go back to my scatter plot let us say I have again 1000 homes, I have my data you know and I am simply going to now draw a regression line. I know this regression line is on the expectation of P_i given R_i which is nothing but beta 0 hat beta 1 hat R_i, right?

Now, obviously, for every value of R_i, the model is suggesting that there is a level of you know P_i, but this is nothing but the expectation of P_i given R_i you know let us say R_i equals R tilde; now this will be R tilde, right?

But for every R_i, you have to understand that there is a predicted value and there is truth; now the truth can be different. So for example, the truth is here and the predictive value is given

by the green dot. The distance between the 2 is the error of the model it is that model error, right? So, every data point and it is the distance from the regression line actually represents an error.

Now, this error is sometimes positive, it is sometimes negative. The assumption that expectation u_i is 0 is basically saying that when I sum these errors throughout my sample, my scatter plot will sum to 0, that is A 1, which is assumption 1. Assumption 2 is a bit more restrictive. Assumption 2 is saying that for every value of R_i that I can find my data for there is we can we will have the expectation u_i equals 0.

That is to say, there is going to be an error, where the predicted value is an underestimate of the truth, but there is also an equivalent value of the truth where the predictive value is an overestimate of the truth. In a way that this underestimation over estimation they actually cancel each other, ok. That is what this second assumption is trying to you know trying to tell you.

Now, you know what it means? well before I go on to what it means mathematically let me just say that this is a crucial assumption for causal inference. So, the impact you know just because I have written y_i equals beta 0 plus beta 1 R_i plus u_i just because as an analyst I have decided to put y P_i on the left-hand side and R_i on the right-hand side does not mean that you know the directional relationship is from R_i to P_i, right?

Equivalently I could have just flipped the model and put R_i on the left-hand side and P_i on the right-hand side. Does that mean that you know the price impacts the number of rooms, did the market price come before the number of rooms well no it did not, right? Now, how do you decide the directionality, you know how you decide the impact that you are seeing is indeed from R_i to P_i, right?

To be able to say that the assumption in front of your screen is needed, right? If the expectation u_i given R_i is not equal to 0 beta 1 is not providing a causal impact of R_i on P_i it is merely a correlation or association between P_i and R_i, ok. So, if I were to just look at a correlation metric and not even specify a regression I would have been fine, right? The pain that I am taking to come to the regression analysis or regression modelling is to be able to work with this assumption which is expectation u_i given R_i equals 0, ok.

Now, let us see how this assumption you know mobilizes causation. Now, so, let us write down our model of interest P_i equals beta 0 plus beta 1_{Ri} plus u_i. Now, beta 1, beta 1 is nothing but del P_i by del R_i. As an interpretation beta 1 as an English interpretation beta 1 provides a measure of an increase in P_i upon a marginal change.

Remember marginal change means, the delta change in R_i in R_i right? That is R_i goes by a unit 1 well we are talking about the number of rooms. So, you know you cannot have half a room or a 1 4^{th} of a room or 1 8^{th} or 3 8^{ths} of a room. If you have another room you have another room, right? And this relationship will hold at all levels of R_i, right?

It does not differentiate if I am going from a single-bedroom apartment to a double bedroom or 3 bedroom or a four-bedroom and so on and so forth. This specification P_i equals beta 0 plus beta 1 R_i provides me with a beta 1 change in P_i upon a 1 unit change in R_i. Now see that the above interpretation crucially relies on the assumption that expectation u_i given R_i is equal to 0, right?

To be able to see that let us say you have P_i equals beta 0 plus beta 1 R_i plus u_i and P_i tilde when R_i goes from R_i to R_i plus 1 and I have a new error variable u_i tilde, right? Then, P_i tilde minus P_i is attributable, this change is attributable only to R_i delta R_i equals 1 if u_i tilde minus u_i is equal to 0. That is to say that the error term is the change in error when we move on from a smaller house to a bigger house nothing changes in u_i.

Nothing changes in this unobserved uncaptured error term of the regression model. If indeed you know, if you had a situation where u_i minus u tilde is not equal to 0, then we cannot say that this change in P_i tilde minus P_i is not attributable solely to delta R_i.

Rather it will be confusing where is it coming from, is it coming from a change in rooms or is it coming from that unobserved factor sitting in u_i. Now, the question is what would be that unobserved factor? Let us continue that in the next part of this lecture.

Thank you.