Lecture - 55

Practical
Spatial Statistics and Spatial Econometrics
With R

Prof. Saif Ali
Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

Session - 09
Lecture - 55
Spatial linear regression in R - II

Hi, welcome to another practical session in which we are learning Spatial Statistics and Spatial Econometrics with R. My name is Saif and today we will continue our discussion of spatial regressions in R.

So, we will try to develop the programming skill needed to run spatial regressions in R, but this skill can only be acquired based on a solid understanding of the theoretical principles.

And what we need to understand to fully appreciate the programming that we will do in this session is the idea of linear regression using ordinary least squares, the assumptions that we make before we run OLS, what is spatial autocorrelation, and how it violates those classical assumptions as well as an understanding and familiarity with the basic spatial models the SLX, SAR, and SER models.

In the last session, what we did was we prepared our spatial data for regression analysis, if you recall it was groundwater levels data and we were trying to explain groundwater levels in space using some observations on rainfall, temperature, and cropping; so, trying to combine climatic and anthropogenic factors to explain the groundwater level over a region of space in Uttar Pradesh.

So, we prepared this data and we built our neighborhood or weights matrix that is required to indicate the spatial relationships in our region. We need these two steps before we start running our regressions. I will briefly review these steps for those of you who might have missed the last session, if you did miss the last session I encourage you to pause the video now and go back and watch that session, which was session 8, which was part 1 of spatial regressions.

So, that you have a firm and clear grasp of how to prepare the data for regression analysis and how to build your weights matrix, your neighborhood, or your adjacency matrix. Once we have done that we will today start to actually run the spatial regression models using R and then we will look at the results and try and compare what are the differences between the results when we run different models.

So, quick review, this is something that we should already know, but just to refresh your memory. There are three simple spatial regression models, the first one is the SAR or spatial auto-regressive model in which we include on the right-hand side a spatially lagged version of the dependent variable.

So, this W which is the weights matrix indicates the neighborhood relationships and this rho is the coefficient of this spatially lag variable. The W times y is the spatially lagged version of y so if y is the groundwater level at the location I then W y is the groundwater level in the neighbors of I. So, if we are in a sub-district, let us say Agra subdistrict then W I will be all of the neighbors or W y would be all of the neighbors of that subdistrict.

And how we would pick out the neighbors because the weights matrix will contain ones for all of the neighbors and zeros for everything else. So, we are trying to explain the groundwater level in the Agra sub-subdistrict division using the groundwater levels of neighboring subdistricts. And the coefficient that we want to estimate is rho hat and as usual, we have the other covariates and the error term, when do we use this model well?

We use this model when we expect that some quantity is a function, and is directly influenced by the same quantity in the neighborhood. So, if we believe the groundwater level at some position at some location in space is directly influenced by the groundwater levels around it then we should use this model, and the reason to believe that is because of hydraulic gradient.

If groundwater levels are lower all around me then water will tend to flow away from me into those areas or the other way around if the groundwater levels are higher or the hydraulic head is higher in around me the water will tend to flow in.

So, because of this subsurface flow, ground waters groundwater levels are spatially autocorrelated. So, we try to capture this effect using this spatial lag variable. This is also called the spatial lag model, simply spatial lag and you might find other terms, but SAR or spatial lag model is the most common. Well, you can also have the spatial error model in which case your primary model remains the same, but your error terms are spatially auto-correlated.

Lecture - 55

Now what is the difference between this model and the previous one, well this model is saying that there is spatial autocorrelation in the errors of each observation. Meaning that the spatial autocorrelation is coming from unmeasured factors, it is not coming directly from variables that we include, but from unmeasured factors which is a little different from, when the spatial autocorrelation is coming from direct relationships between the dependent variable at various locations, this is this we do not know the source of this.

So, we do believe that the groundwater levels are spatially autocorrelated, we do not know why and we do not include any variables in our regression model to capture that. So, it's going to kind of sit in the error term and this is if that is our belief about the process, then this model is most appropriate.

And then the third one is also a spatial lag model, but the spatial lag is not in the dependent variable, it is in the covariates in x. So, it is sometimes called spatial SLX, spatial lag in X, where we include along with the covariates. So, remember X beta these are the covariates at the same location.

So, if Y is groundwater level at location I then X is rainfall temperature and cropping at the same location, but theta which is W X the second term W X is rainfall temperature and cropping in the neighborhoods of I. So, we include both.

So, we believe that groundwater levels at a location I, are explained by rainfall temperature and cropping at the location I and by rainfall temperature and cropping in the neighboring locations, which is a kind of reasonable assumption because we do expect that in the neighborhood if it rains a lot or if it is very hot or there is was a lot of cropping going on we do expect these geographical climatic phenomena or the anthropogenic processes to be spatially related to make an effect felt at this location. So, we will run this model.

Now usually in realistic situations, you have two, one or more, or two or all three of these effects happening simultaneously and for those you need more complex models, but to start with, it will be good to get a grasp on these three simple models how to run them in R and try to interpret the results before you start combining things.

So, if that is clear, I am going to move on to the coding session, if it is not clear please pause now go back, and listen to this video again maybe another video in the course where Dr.

Lecture - 55

Gaurav Arora talks about these models, please review your materials at this point and when you come back I will be right here.

So, I am going to go ahead with the coding session with the understanding that all of us are kind of clear on these matters, alright?

So, before I go to today's code, I am just going to briefly review our code from last time.

So, remember as step 1, we kind of prepared our spatial data. We first read some data from a CSV file that has data on post-monsoon groundwater levels, rainfall, temperature, and cropping. And then we converted it to spatial data and then we loaded our state boundaries or subdistrict boundaries for Uttar Pradesh, we aggregated our data to these boundaries. So, we have a value for every subdistrict as opposed to a value for every well.

So, we aggregated the well-level observations to a higher sort of unit and then we did some other work with the data and we sort of plotted our data. So, this is what the post-monsoon plot looks like.

Now notice that in one of the subdistricts, there is no data here. So, it is this data is missing. So, this will cause problems for us later on.

So, what I am going to do is, I am going to take this district out, I am going to take it out of the data set and that is what I am doing here, I am going to remove the unit with missing data. So, if you remember, we will use this condition dot is n a is a way to check if an observation is missing, and when I say not is dot na; that means I want only those where the observations are not missing.

So, I only want available observations for post-monsoon. So, I only want to keep the rows where post-monsoon observations are available, if it is missing I do not want that. So, I think that gets rid of the one district. So, I create a new data set up level 3 dot spatial dot 1. So, dot 1 is the 1 where nothing is missing. So, that is the data set that we will use going forward.

And then we build our weights matrix and visualized the weights matrix as follows.

So, please see that now for this district where data was missing, this is not part of the weights matrix, it is not because we have taken it out of the regression completely. So, if this is clear, now that we have our spatial data, we have our weights matrix, our weights matrix is contained in this variable called w dot list w, and we can start to run our regression models. All of this should be completely clear, if any of this is missing then you know what is coming is not going to make a lot of sense.

So, please review as many times as you need before going forward. So, this is our regression equation, we want to regress the post-monsoon level that is our dependent variable on the rainfall of this year's monsoon rainfall, this year's summer temperature, and last year's cropping.

Those are the three variables, that are our three covariates, that is the basic regression equation that we would use in ordinary least squares if there was no spatial component. So, that is what we will do first. We will first run ordinary least squares to see what kind of results we get and this is done with the command lm and then let us look at the results.

So, what we see is that rainfall is highly significant and the sign is negative, which is good news. So, you might, if I may just tax you with a small exercise, what does it mean that the sign is negative you might pause the video and think about it for a few seconds before you come back. We are seeing a significant coefficient for rainfall this year's rainfall and the sign is negative, and what that means, is that the rainfall is sort of inversely related to groundwater levels.

So, that means areas where rainfall is high meaning the magnitude of rainfall in meters is high, at those locations the value of groundwater level is smaller. So, the groundwater level being smaller means that the water is closer, and the depth is smaller.

So, water is closer to the ground which is what we expect in very humid wet areas, we expect water levels to be shallower lower, and in areas where rainfall is small like areas that border on Rajasthan, or hotter areas and drier areas we expect groundwater levels to be much deeper. So, the value of depth will be higher.

So, that is what the negative sign is telling us. The other two variables do not seem to be very significant although the sign of percentage cropping seems to be correct that if there is more cropping then we expect the groundwater level to be deeper. And the sign of temperature is negative which means hotter areas have shallow groundwater levels which is counterintuitive, first of all, we must realize that this model may not be correctly specified, we may have omitted variable bias, and we may have other problems in addition to spatial autocorrelation.

Lecture - 55

So, while I am interpreting these results for you, they may not necessarily be interpretations that we want to use for analysis or for gaining any kind of understanding, this is just a demonstrative discussion of how we might interpret regression results. So, now, what we want to do is, well before we start running spatial regression models, we have to understand whether are they even required, do we even need spatial regression models, and maybe these ordinary least squares are good enough.

So, what we do for that is, what we can do, we can do a lot of things, but the one thing we can do is, one of many things that we can do is, we can try to look at the residuals.

So, from this ordinary least squares regression we can pull out the residuals. So, we will have a residual for every sub-district right, because we have 213 observations, which you can see from here, here it says 209 degrees of freedom because we have 4 variables. So, the number of observations we can see from here in our data is 213, right? So, for each one of these, we will have a residual and we want to see if these residuals are spatially autocorrelated.

So, if there is some residual spatial autocorrelation then; that means, that the ordinary least squares is not good enough, we need a spatial model; that means, there is some residual spatial autocorrelation that is not being captured by the model. So, we can get the residuals using this command here and store them in a variable called residuals and then we can plot those residuals.

So what does that look like? I mean now that we have gotten used to looking at spatial data, does that look like random like a spatially random distribution or is there some spatial autocorrelation? Well, it does seem that there is some clustering, that there are areas where the residuals are high like here and here maybe here, and then there are other areas where residuals are 0 close to 0, and others where they are negative. So, areas of positive residuals, negative residuals, and 0 residuals seem to be clustered together, which signals spatial autocorrelation.

So, what we can do is we can compute the Moran's I for the residuals to get a sense, a quantitative measure and we see that the Moran's I is 0.4 which is not necessarily very low, there does seems to be positive spatial autocorrelation p-value is pretty low. So, it does seem that we have some residual spatial autocorrelation. So, we might benefit from running spatial regression models.

Lecture - 55

So, let us run our first spatial regression model. So, let us go with SAR the autoregressive model first. So, the command to do that is lagsarlm. So, lag means spatial lag, sar means spatial autoregressive, and lm means linear model, and this is available in the spatial reg package.

So, you can look at the help file for this by typing this and this is available in the spatial reg package.

So, you can look at the specification here.

So, it gives you a little explanation and then shows you how to use it. Here what I am doing is I am passing in my regression equation which is the same as before, the data of course, and one additional parameter, which is the weights matrix because remember now on the right-hand side we have a spatial lag variable so; obviously, we need to pass in the weights matrix otherwise R cannot compute that spatially lag variable. So, let us run this model and look at the results.

So, now we see sort of something interesting that as before temperature and percentage cropping are no longer significant whereas, rainfall is. So, remember this is rainfall at this location, at the current location is still significant, the p-value is 0.01, which is less than 0.05. But the coefficient is much smaller in magnitude.

So, it went down from minus 20, which was in the value in OLS to minus 5. So, that seems to indicate a bias and then the rho which is the coefficient of the spatial lag term is 0.74 and, significantly, the p-value is less, close to 0, and that seems to be a fairly high spatial autocorrelation that is close to a value of 0.8.

So, we see that the spatially lagged term has a fairly high explanatory power, that it takes away the exploratory power of rainfall. So, we see that not only groundwater levels in this region are influenced by rainfall in this region, but also strongly influenced by rainfall, by groundwater levels in the neighboring regions, and then gives us some more information here. So, for example, the one thing that it gives us is the AIC criteria, I do not know how to say the name.

So, I will not say it. So, the AIC criteria are 1153.8 for this model and 1248 for the linear model. So, the AIC criteria are slightly higher for ordinary OLS and slightly lower for this.

Lecture - 55

So, while we want to know, did we do any better, did we catch more of the, is there residual spatial auto-correlation now? Now that we accounted for some spatial relationships so, we can repeat the same exercise.

We can get the residuals for this model and then plot them and then see whether there are still any.

So, we still see some spatial clustering we see high residuals in these areas, but there is a little less clustering in areas of low or negative residuals, it is a different picture from the previous one, but we still do not know. So, we can run a Moran's I test and see.

So, now we see that Moran's I test is not even that significant, and the statistic has a very low value it is close to 0, 0.04. So, it seems that residual spatial autocorrelation has reduced significantly from the ordinary. So, that potentially means that we did a little better at capturing spatial relationships.

So, we can now repeat the same process for our two other models. So, this is the SLE model, the command is errorsarlm, same as the previous name except the word lag has been replaced by the word error. And it has the same parameters, we also pass in a tolerance because it is an iterative algorithm. So, let us run that and store it in an object called fit dot sle, this is our spatial error model and then run a summary of that. So, what do we get?

So, rainfall is still just about significant, the value is different from both the ordinary least squares and it is kind of between the spatially lag model and the ordinary least square. So, the ordinary least square was minus 20, spatial lag was minus 5, now it is minus 10 then close to 0.8 and it is significant.

So that means, that that the error terms also had some spatial autocorrelation. So, let us see if the Moran's I for that kind of is. So, let us plot the residuals and let us see Moran's I.

So, Moran's I is again not significant and it is close to 0. So, that model also helps us, it somehow helps us to capture spatial autocorrelation more than the ordinary least squares model; now we can run SLX, and if you do a summary of this.

So, what you will see is that you get the usual coefficients which are the rainfall temperature and percentage cropped area for this location, and then lagged versions of all of them. And now we see some interesting results so, we see temperature becomes significant in the sense that it has a positive sign. So, that positive sign is something that we can interpret in the sense that hotter areas will tend to have higher groundwater depth.

So, that makes sense, and somehow when we did not have this result when we did not include the spatially lagged temperature variable, but if we include the spatially lagged temperature variable then we have a significant coefficient, but interestingly, the lag temperature variable has a negative sign so; you will have to think about what; that means, is that while if it is hotter here, the groundwater level here will be deeper, but if I am surrounded by cooler areas or hotter areas.

So, the temperature here has a direct relationship, but the temperature in the neighborhood has an inverse relationship. So, that is counterintuitive and so, I would relook at what model I would use for this, similarly, we see that lag rainfall has a very high negative sign and is significant so; which means, the groundwater level here is inversely strongly related to rainfall in the neighborhood. And then we can do the same thing, we can plot the residuals for this and run a Moran’s I.

So, for this, we still see residual spatial autocorrelation. So, this model does not necessarily get rid of the spatial autocorrelation like the spatially lagged and the spatial error models did, one could argue that what; that means, is that we should probably go with the spatially lagged or spatial error models and this is not necessarily the correct model, but in reality I think the best approach would be some kind of mixed model because rainfall in the neighborhood is important, temperature in the neighborhood is important as well as groundwater levels in the neighborhood.

But before you start mixing and matching models, I would encourage you to get a firm handle on how to prepare your data, your weights matrix, and how to run these models and start interpreting the results before you go forward.

So, that is all I had for today’s coding session just to summarize, if you can help me summarize, what did we do, we ran OLS and three simple spatial regression models. We compared the results we ran and how did we compare the results by doing visual plots of the residuals and running Moran’s I test for residual spatial auto-correlation. I am going to leave you here.

Thank you so much for your attention.