Lecture - 54

Practical
Spatial Statistics and Spatial Econometrics
With R
Prof. Saif Ali
Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

Session - 08
Lecture - 54
Spatial linear regression in R - I

Hi, welcome to another practical session, where we are learning Spatial Statistics and Spatial Econometrics With R and today's topic is spatial regression, particularly linear spatial regression. My name is Saif Ali.

And as we have been emphasizing in these practical sessions, our goal is to develop programming skills to do our spatial statistics tasks and spatial econometric tasks. However, the skill acquisition is based on a solid clear, and precise understanding of the theoretical principles of the subject, which you will gain from the theoretical lectures that you have been watching so far.

And for today's session, these are the points that would be good to have a solid understanding of, because if we do not understand this what will happen is that we can run an R command. It is just like running a command, you can type it in, you can download the data run the command, and you will get some results.

And you can even maybe put those results in a paper or something but to interpret those results, to gain insight from those results. And more importantly, to deepen our understanding of our research area / our study area, we will need to go beyond simply the R command to do something and see internally what is going on.

So, to understand the skill-based material, the R code that we will discuss today it would be good if we understand what is linear regression and what is ordinary least squares. This is something that you can gain fairly easily from a textbook or other online lectures and is kind of required background knowledge for this session.

And then, of course, the classical assumptions that we make before, we run ordinary least squares regression. And how spatial autocorrelation when we work with spatial data, and how that violates those assumptions, it would be nice to know why we need spatial regressions in

the first place. Why do we need a different class of models to work with data, which has spatial relationships, spatial autocorrelation, and spatial effects?

And of course, we need to be familiar with the spatial models that have been covered in this course, the SLX cross-regressive model, the spatial auto-regressive model SAR and the spatial error model. I have used some abbreviations here you may find abbreviations that are slightly different if you read a different book or maybe read some papers.

But these are the basic three simple spatial regression models that are commonly known and we would like to have understood those fairly well and fairly deeply to appreciate the code that we will work with today, that is it, what should we already have done in terms of our programming, well we should be fairly comfortable with loading new libraries.

Working with R, we should be fairly comfortable with the interface. We should be able to learn new functions and new libraries by reading the help manuals. I hope that all of you have gained this level of proficiency in R by working through the material in this course so far. If you have not, it would be nice if you could review some of the previous exercises and try and replicate the results that I have shown you.

And of course, you should be comfortable by now working with spatial data. I should know what spatial data is, how is it different from regular tabular data, and what are the different operations that you can perform with spatial data, plotting it, computing variograms etcetera.

So, if all of that is in place, if some of it is not in place, I would highly recommend a brief review, you can pause this video right now. And I will be right here when you come back, you can go back and review some of the materials and then start from this point onwards that would really help you to acquire the skills in a more solid and firm, get a more firm grasp on the skills that we will discuss today.

And what we will discuss today is we will again prepare some new spatial data for regression analysis and this data is specially prepared to run regression models. So, it is slightly different from the data that we used earlier, but it is data that pertains to groundwater levels which we are familiar with by now.

So, that will be step one, step two will be to build a weights matrix and that term I hope should be familiar to you because when we run spatial models, we need a matrix that tells

you about the different neighborhood relationships in your data. And it encodes that information in a matrix form which is called the weights matrix.

So, we will learn how to build this matrix using R commands. So, I am going to move forward, I am hoping most of you are with me if some of you have gone back to review the materials earlier, I will see you on the other side.

So, we will do a review of spatial regression models. But we will not do it right now. Let me get into the code a little bit, let me show you how to prepare the data and I will come back to this slide and we will review these models nearer the end of the session at the correct place.

So, we already know the three basic models which would be working with, we will not actually run these modules in this session. We will run them in the next session, but we will prepare, we will do all of the preparation that is necessary to run these models and the preparation we will do in this session.

So, let us transition to our r window.

In the r window, we have a new file corresponding to session 8, this file is called R Session 8 spatial regression part one, and of course, we will share this code with you.

So, let us begin by loading our libraries, now you are familiar with most of these, but there are three new libraries that we have not used so far. And they are called spdep, spatial reg stands for spatial regression, which provides most of the code that we need to run our spatial regression models, and an rgos, which has some additional code to help us prepare our spatial data.

So, we will go ahead and load these libraries. So, remember if you want to run some code in R, you can just select it and press run and it will load all of those libraries. So, it seems that went well and we have loaded our libraries. So, we are going to be working with a new data set which is called spatial hyphen regression dot csv.

So, let us get into this data, we now know how to load data from a csv and we load it into an object called spreg, and then let us examine it, let us see what is inside. So, it seems there is district-wise or block-wise data. In fact, station-wise so; you have a district block and hydrograph station, type of station here. So, you have data, this is data for the year 2015 well,

id, spatial coordinates, which is of course, necessary otherwise we would not have spatial data.

And then we have four variables called post-monsoon, RFL, TY, which stand for rainfall this year. So, TY is this year so; that means, what is this year this post monsoon observation for groundwater level was made in post monsoon of 2015. So, that is close to the end of August or the beginning of September in 2015.

And rainfall this year means the rainfall received around that particular well in 2015 in the months from June to September. So, the monsoon rainfall occurred right before the observation was made. And similarly, there is temperature this year TEMP TY and this was the mean temperature in April and May of 2015.

And then we have a percentage cropped area LY. So, LY stands for Last Year. So, we have the percentage of cropped area around a given well where an observation was made, what percentage of that area was used for cropping that is what we have here and this is last year. So, this would be for the year 2014.

So, remember that the cropping year begins around May and so, from May 2014 to May 2015 is the cropping year. So, this variable will encompass all the cropping that occurred between May 2014 and May 2015. And because of course, we do not want cropping this year because if you take cropping in 2015 then you are accounting for cropping that occurs after the observation for groundwater level was made. And that does not make any sense if you want to regress groundwater levels and we want to see what factors influenced or were associated with that groundwater level.

Then the cropping of the future is not necessarily something that if the groundwater level is the dependent variable then the cropping of the future as the independent variable does not make sense. So, we have taken cropping of the past and rainfall of the immediate past, and temperature of the immediate past.

I hope that is clear. So, let me just show you this in a better view here by clicking this we can see our data and these four variables. So, what we want to do is we want to understand how this year's rainfall was, how much rainfall was received in some neighborhoods of a well, how hot the summer was, and how much cropping went on in a period of 1 year.

Lecture - 54

How those three factors influence the groundwater level observation made at a particular well and of course, we will aggregate this data to a to a subdistrict unit. So, we will not do it well-wise, but that is the data that we have a dependent variable called post-monsoon and we have three explanatory variables and we want to run a regression to understand the relationship between the explanatory variables and the dependent variable.

So, that is the scope of what we want to do so, now, that we have loaded our data of course, we have to convert it into spatial data, which is done by using the coordinates command. So, let us do that. So, now, we have spreg which was just a regular data frame has turned into a spatial point data frame.

So, let us load our spatial boundaries. So, let us read in a shape file for the administrative boundaries in Uttar Pradesh.

Now, remember we have worked with administrative boundaries previously, but those were level two boundaries. So, level two boundaries are district boundaries, but now we want to work at a slightly finer level. So, we will use level three spatial boundaries which are subdivisional or subdistrict level. So, they are something like within a district there are administrative subdivisions.

And we will be working at that level. So, we will aggregate all of our data to the sub-district level. So, now, that we have read the shape file, let us subset it because we do not want. So, if we plot the shape file, let me just plot it and show you. So, we have called it upstate, but if we plot it, we see that it is actually, let us see what we get, it's actually all of India.

But that is not what we want. So, let us subset it only to include UP, and the way we do that is we use the name one variable, the name one variable contains the names of the states. So, whatever state you want you can subset it using this variable, I hope this syntax is familiar to you, how to subset a data frame. So, we only want the polygons where the state is called Uttar Pradesh. So, let us do that, and then let us plot it again and that is much better.

So, now we only have the boundaries for Uttar Pradesh. So, these are the subdistrict administrative units, we will aggregate all our data, the groundwater level, rainfall, temperature, and cropping. We will add up the wells within each of these units and take the average. So, we will be working with means that are aggregated to the sub-district level.

Lecture - 54

So, we will not work with the data as it is, we will perform an extra step to spatially aggregate the data, and the way you spatially aggregate data is using a command called aggregate, this is a new command which we have not seen so far. So, it would help, if you could pause the video right now. And then read the help file for this command which is done by using the question mark.

So, aggregation of spatial objects. So, what it does is, it takes the data which is point data. So, remember we have data add wells which are individual points. So, for a well, we have the groundwater level at that well.

How deep was the water level at that particular well, we have the rainfall that occurred around that well in some neighborhoods. I have prepared this data, I am not going to get into how I prepared it, but we have the rainfall that occurred in let us say a 1 or 1 to 4-kilometer radius of that well.

We have the temperature from 2 meters from the surface of the earth so, that the sort of the air temperature around that well, is the mean temperature in April and May. So, to get a sense of how hot the summer was because you know if it is hotter then wells dry up and water levels fall.

And similarly, if there is more rain then water levels rise and we want to understand these relationships through a spatial regression and then, of course, cropping. So, how much cropping occurred, what percentage of the area in the neighborhood of the well was used for cropping? So, if it is a high percentage then that well is probably in like a rural area where there is a lot of cropping going on if it is a low percentage.

Then it is maybe in an area that is urban less cropping or industrial or maybe in a forest somewhere close to a road, but there is less cropping going on. So, that gives you a sense of how much water was extracted by farmers around that well, because if you abstract a lot of water around a well then the water level at that well will go down. So, we have these three variables and we have these data at individual points.

So, it is a point data set, but we do not want a point data set, we want to aggregate these individual well observations in a particular subdistrict. So, if I have a subdistrict let us say Agra subdivision and I have 10 wells then I want to take the average of all those wells and just compute one value for the whole subdistrict.

So, that is what I am doing here using this command aggregate and the first parameter is the regression data that we just loaded, remember we turned it into a spatial data set. So, if you pass it a special data frame, aggregate will perform a spatial aggregation. And then we pass in the polygon boundaries that will be used for aggregation. And these are the boundaries.

So, we want to aggregate this data which is point data for individual wells, which are at these spatial coordinates onto these boundaries. So, let me just show you, let me try and show you a map of both of those together. So, let me show you a map of both of those together, this is just some live coding here. So, let us say we do this. So, now, we want to also show the wells on top of this, and then let us see.

So, that is the data set; however, the points are way too big. So, we maybe make them a little smaller, even smaller, maybe even smaller.

So, let me just zoom that in for you. So, we have UP and we have data at these point locations.

So, what we are going to do with this aggregate function is that we are going to go into every sub-district, take all of the wells, add up all of the data variables by variable, and then take the average. So, we will compute an average post-monsoon level for this subdistrict then for this subdistrict, also the average rainfall, the average temperature, and the average percentage of the cropped area.

And we will do this for every sub-district of course, we are not going to do it, r will do it for us, but just to show you that is actually what we're doing when we run this aggregate command. Please pause the video, go back and watch it again if that was not clear, and repeat it as many times as you feel comfortable. I am going to assume that you understand what aggregate does.

So, here, I am giving a function which is mean because I want the average, I could also ask for the sum, or the maximum, but I want the mean and na dot rm equals TRUE tells R to ignore missing values. So, some data will be missing and we want to ignore it.

So, let us run this. So, I get some error saying. So, spatial data has a projection system which I cannot teach you right now.

Lecture - 54

In the interest of time because there you could have an entire subcourse on projection systems, but the idea is if you want to aggregate points over a polygon of spatial data then two of them must have the same projection system. So, let me do that, let me set the projection of the points to the projection of the polygons and then that gets rid of that.

And now aggregate, so, that seemed to go well, and also for the percentage cropped area, I do not have a percentage, I have a proportion. So, I have things like 0.2, 0.4. So, I am just going to multiply that by 100 just so that it is truly a percentage.

Now let us plot, let us see how that went. So, let us plot our post-monsoon value. So, now, we do not have values over individual wells, we have values over entire subdistricts. So, we have a value over this subdistrict, this one, this one, and this has been obtained by averaging over individual wells. So, you can see the darker orange colors, these areas have maybe deeper groundwater levels. So, maybe this is a depletion sort of hot spot, this area borders, it is around Agra, but it borders on Rajasthan.

So, in this area, we expect that there is not a lot of rainfall because it borders on Rajasthan which is an arid desert area. So, maybe that is why groundwater levels are not so good here then up here we have a lot of sugar cane farming. And then these darker yellow regions have slightly deeper groundwater levels and the lighter yellow groundwater levels are a little shallower.

The water is closer to the earth, the ground surface. So, this gives us a sense of the spatial distribution of groundwater levels.

So, let us look at rainfall and we see that rainfall clearly shows a trend from the east, from the west to east right. So, the western part which borders on Rajasthan is kind of dry. So, there is not a lot of rain as you go closer to the eastern parts near the mountains and nearer Bihar and those areas of West Bengal you get more and more rain.

And clearly, like right along the foothills of the mountains, you get a lot of rain and then as you come down into the plains and go into the desert. So, it is a very clear sort of trend that we see. So, we do expect that groundwater levels in the western half maybe you know groundwater level to be kind of deeper. Because there is less rainfall that is something that we expect.

And let us look at temperature. So, you can look at each of these and try to develop some theories. So, the temperature has a south-to-north gradient and we know that the Northern parts of India are much colder, Southern parts are warmer and we see that across UP being that trend manifesting across UP as well. Right here, you know when you are getting into Uttarakhand, it's quite cold, and then down here you know when you are getting into MP and down here into Bihar etcetera it is much hotter.

And similarly, you can look at cropping. So, cropping clearly is more in the western, in western UP there is more cropping going on area-wise, and then here in the south maybe a little more, and then some parts of northeast UP.

These are rice districts with a lot of rice cropping. So, now, that is all good, we have examined our data, and we have aggregated it, but now we want to build a weights matrix. So, how are we going to do that, we want to build the weights matrix and also get a visual sense of what it looks like. So, the way to do that is using the spdep package which we have already installed.

So, the first command is we have our spatial data which is called UP level 3 spatial data. So, we use a command called poly2nb which stands for the polygon to the neighborhood. So, it takes a set of polygons and gives you the neighborhood relationships.

So, what is a neighborhood? If for example, this district here, sub-district borders this district and this one and this one and this one and this one. So, any district that it shares some edge with or some vertex with that is a neighbor, that is what we defined as a neighbor. So, for each subdistrict, we want a list of all the neighbors and we want to organize it in a matrix form.

And R does this for us. So, we run the command poly2nb first and then we get the centroids of each of the sub-districts.

And then we can plot the neighborhood relationships, and we can see. So, we can go into each sub-district. So, we can see this sub-district is connected to all of those around it and this one is connected to all of those around it and this is called a queen neighborhood pattern because even if you share a corner with another sub-district you are counted as a neighbor. There are other kinds of topologies that you can consider.

But, this is the most commonly used one, we will stick with this one for now. So, this is fairly easy to understand. So, this sub-district right here, only neighbors this one. So, there is only one edge, this one neighbor only has one, and this one has two neighbors one to the north and one to the south.

And then most of these sub-districts are kind of connected to all of those around it which is why we see this Web-like pattern. And then we have to do one last step, we want to convert the data format from one type to another, do not worry too much about this, this is just a data conversion nothing is happening, we are just converting the same information to another data format because later we need it as a list and not as this. So, well do that and we will stop here for this session that is all were going to do.

So, just to summarize what did we do, we prepared spatial data to run spatial regressions. And then we used this spatial data and the preparation of the spatial data was many operations loading the data converting it into spatial data then aggregating it, plotting it, and looking at the various variables then we built a weights matrix to encode the neighborhood relationship and then we visualized that relationship just to see that everything turned out well.

Now, we are set, we have all the information, all the variables, and the data in place to start running our spatial regressions, which we will do in the next session. I encourage you to review this material thoroughly. So, that both you and I are ready for the next session, I will see you there.

Thank you so much for your attention.