R Module 10: Spatial regression in R

Regression residuals are the primary way to assess if a regression violates any of the assumption of the general linear model. Although all the assumptions are important, the violation of the assumption of independence is common with spatially organized data. This has led to the development of a family of alternate regression models that can accommodate spatial dependence. We'll work with two of these models from previous R Module; we'll simply call these the spatial lag model and the spatial error model. Both models account for spatial dependence by including new parameters that must be estimated but each model handles these differently. The spatial lag model, shown below, introduces a new explanatory variable and an associated coefficient that must be estimated.

$$y = \rho Wy + \beta x + \epsilon$$

The new variable is called the lag variable and it is a weighted average of the values of the observations defined as the neighbors of any given observation. This new lag variable is referred to as *Wy*. Like any other variable, *Wy* has a coefficient which must be estimated; this is called rho or ρ. Just like the coefficients for any independent variable, there is a hypothesis test associated with ρ; the null hypothesis is that ρ=0 which would mean that there is no spatial dependence in the dependent variable (or the *Wy* variable). A statistically significant result for this parameter test means that ρ≠0 and that we have adjusted for any spatial dependence. The remaining independent variable(IV) coefficients can then be interpreted as we would use OLS. Like any coefficient, ρ can be either positive (indicating positive spatial autocorrelation among neighbors) or negative (negative spatial autocorrelation among neighbors).

The other model is called the spatial error model. Unlike the lag model, the error model addresses spatial dependence by separating the error term into two components: one that contains the random error components and one that contains the spatially autocorrelated error components. The formula looks like this:

$$y = \beta x + \epsilon, \epsilon = \lambda W\epsilon + \zeta$$

Just like with the lag model, the coefficient (λ or lambda) of the spatially autocorrelated error term must be estimated. The remaining errors (ξ or xi – pronounced like 'sigh' or 'zigh') should be randomly distributed as per the conventional general linear model. The interpretations of the λ coefficient are similar to that of ρ in the lag model as is the hypothesis test informed by the coefficient's test statistic and p-value. Both of these models are computationally challenging and require the extra steps of defining the neighborhood construct and its associated weighting scheme. So why bother? The problem of spatially autocorrelated residuals means that we can't really trust our results. Our parameter estimates are inefficient and perhaps biased which leads us to draw potentially incorrect conclusions about the hypothesis tests associated with the coefficients of the independent variables. This is a long way to say that any inferences we make based on such models can be very wrong. The coefficient of determination also tends to be inflated if we don't account for spatial autocorrelation in our residuals. The lag and error models are a meaningful advance toward making reliable inferences in the presence of spatial dependence.

This week, you'll estimate spatial models on *new* data in R and compare the results. The data is a shapefile of election results available through Havard website. A word of caution you may also need to review the code used in previous R modules to brush up on the creation of spatial weights matrices needed by both the lag and error models.

Start by loading the rgdal and spdep librarys. We will be loading data through R rather than a shapefile (hopefully reducing error messages found throughout previous R modules).

```
## Set working directory
setwd("/Users/kovachmm/Documents/ R")
load("Datasets.RData")
ls()

##save(laos,crime,cities,volcano,election,dat88,mat88,file="Datasets.RData"
```

) I also recommend exploring the data using the following functions:

```
##Explore Data
summary(election)
names(election)
data <- election
```

Create a weights matrix of polygon centroids

```
##Create a matrix of polygon centroids
map_crd <- coordinates(data)

## Contiguity Neighbors
W_cont_el <- poly2nb(data, queen=T)
W_cont_el_mat <- nb2listw(W_cont_el, style="W", zero.policy=TRUE)

## Plot the connections
par(mar=rep(0,4))
plot(W_cont_el_mat,coords=map_crd,pch=19, cex=0.1, col="gray")
```

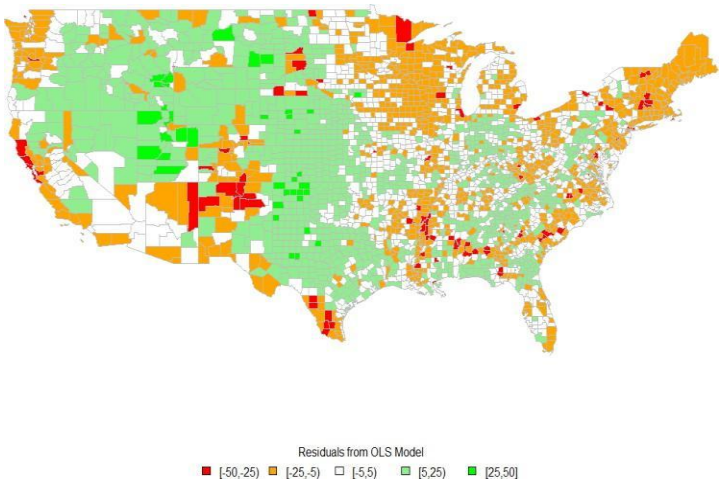
We can run a global Moran's I test or Geary's C test on both variables that we'll use in the regression models. First, test for spatial autocorrelation in the percent Bush voters (dependent variable) and income (independent variable).

```
## Global Autocorrelation Tests: Moran's I
moran.test(data$Bush_pct, listw=W_cont_el_mat, zero.policy=T)

## Global Autocorrelation Tests: Geary's C
geary.test(data$Bush_pct, listw=W_cont_el_mat, zero.policy=T)

## Global Autocorrelation Tests: Moran's I
moran.test(data$pcincome, listw=W_cont_el_mat, zero.policy=T)
```

Global Autocorrelation Tests: Geary's C
geary.test(data$ pcincome, listw=W_cont_el_mat, zero.policy=T)

Both variables are strongly and positively spatially autocorrelation using the queen contiguity neighbors spatial weight matrix. This is a clue that we might expect any residuals from a regression using these variables to also display spatial autocorrelation. If so, we'll need to think about using a spatial regression model to estimate the relationship.

1.) Use lm() to estimate Bush_pct ~ pcincome using ordinary linear least square (OLS). Plot the residuals and test for spatial autocorrelation. Provide an example of your map and an output of of your spatial autocorrelation results. Hint your results should look similar to the plot on the right.



2.) Estimate the same relationship as above using the spatial lag model instead. The lag model function is in the spdep library and is called lagsarlm. Provide the results of your regression.

3.) Perform a Moran's I test on the residuals of the lag model and report your results. Is there evidence of remaining spatial autocorrelation in the residuals?

4.) This time, estimate the same function using the error model (errorsarlm). Provide the results of the error model and explain any differences in the coefficient variable from the lag to the error model. Also perform a Moran's I test on the residuals of the error model and provide those results.