Texas A&M Census RDC - Spatial Analysis Workshop

Corey S. Sparks, Ph.D. - The University of Texas at San Antonio

May 14, 2018

Introduction

About me

Conceptual Stuff

What's special about spatial?

- Spatial data have more information than ordinary data
- Think of them as a triplet - Y, X, and Z, where Y is the variable of interest, X is some other information that influences Y and Z is the geographic location where Y occurred
- If our data aren't spatial, we don't have Z
- Spatial information is a key attribute of behavioral data
- This adds a potentially interesting attribute to any data we collect

- Spatial data monkey with models
- Most analytical models have assumptions, spatial structure can violate these models
- We typically want to jump into modeling, but without acknowledging or handling directly, spatial data can make our models meaningless
- Some of the problems are..

The Ecological fallacy

- The tendency for aggregate data on a concept to show correlations when individual data on a concept do not.
- In general the effect of aggregation bias, whereby those studying macro-level data try to make conclusions or statements about individual-level behavior
- This also is felt when you analyze data at a specific level, say counties, your results are only generalizeable at that level, not at the level of congressional districts, MSA's or states.
- The often-arbitrary nature of aggregate units also needs to be considered in such analysis.

MAUP

- This is akin to the ecological fallacy and the notion of aggregation bias.

-The MAUP occurs when inferences about data change when the spatial scale of observation is modified.

- i.e. at a county level there may be a significant association between income and health, but at the state or national level this may become insignificant, likewise at the individual level we may see the relationship disappear.

-This problem also exists when we suspect that a characteristic of an aggregate unit is influencing an individual behavior, but because the level at which aggregate data are available, we are unable to properly measure the variable at the aggregate level.

-E.g. we suspect that neighborhood crime rates will the recidivism hazard for a parolee, but we can only get crime rates at the census tract or county level, so we cannot really measure the effect we want.

Spatial Structure

- Structure is the idea that your data have an organization to them that has a specific spatial dimension
- Think of a square grid
- Each cell in the grid can be though of as being neighbors of other cells base on their proximity, distance, direction, etc.
- This structure generally influences data by making them non-independent of one another
- At best, you can have a correlation with your neighbor
- At worst, your characteristics are a linear or nonlinear function of your neighbors

Spatial Heterogeneity

- Spatial heterogeneity is the idea that characteristics of a population or a sample vary by location
- This can manifest itself by generating clusters of like observations
- Statistically, this is bad because many models assume constant variance, but if like observations are spatially co-incident, then variance is not constant
- This is really cool

Stationarity

- Stationarity simply means that the process is not changing with respect to either time (i.e. time series analysis) or space.
- This implies that the process that has generated our data is acting the same way in all areas under study.
- The implications of Stationarity are that we can use a global statistic to measure our process and not feel too bad about it.
- It also implies that our observations are iid (independent and identically distributed) with respect to one another
- e.g. the parameters estimated by the regression of X on Y are the same throughout our area of study, and do not have a tendency to change.
- Also, it means the model estimated is equally well specified at all locations. *This is our general assumption in regression models*

Non Stationarity

- If a process is *non-stationary* then the process changes with respect to time or space.
- This implies that the process that has generated our data is not acting the same way in all areas, or the expected value (mean, or variance) of our data are subject to spatial fluctuations.
- If our data are subject to such fluctuations, the this implies that our global statistics are also subject to major local fluctuations.
- Meaning areas in our data can tend to cluster together and have similar values.

Autocorrelation

- This can occur in either space or time
- Really boils down to the non-independence between neighboring values
- The values of our independent variable (or our dependent variables) may be similar because:
- Our values occur
- closely in time (temporal autocorrelation)
- closely in space (spatial autocorrelation)

Basic Assessment of Spatial Dependency

- Before we can model the dependency in spatial data, we must first cover the ideas of creating and modeling neighborhoods in our data.
- By neighborhoods, I mean the clustering or connectedness of observations
- The exploratory methods we will cover today depend on us knowing how our data are arranged in space, who is next to who.
- This is important (as we will see later) because most correlation in spatial data tends to die out as we get further away from a specific location

Tobler’s Law

- Waldo Tobler (1970) suggested the first law of geography
- *Everything is related to everything else“, but near things are more related than distant things.*
- We can see this better in graphical form: We expect the correlation between the attributes of two points to diminish as the distance between them grows.

Basic Spatial clustering

- Clustering means that observations that are close geographically are close in other attributes. Autocorrelation is typically a local process. Meaning it typically dies out as distance between observations increase.
- So our statistics that correct for, or in fact measure spatial association have to account for where we are with respect to the observation under present consideration.
- This is typically done by specifying/identifying the spatial connectivity between spatial observations.
- To measure clustering, we must first see who is next to who

Spatial Connectivity

- Spatial connectivity, or a spatial neighborhood, is defined based on the interactions/associations between features in our data.
- This connectivity is often in terms of the spatial weight of an observation, in other words how much of the value of a surrounding observation do we consider when we are looking at spatial correlation.
- Typically the weight of a neighboring observation dies out the further it is away from our feature of interest.
- There are two typical ways in which we measure spatial relationships
- Distance and contiguity

Distance based association

- In a distance based connectivity method, features (generally points) are considered to be contiguous if they are within a given radius of another point. The radius is really left up to the researcher to decide.
- For example we did this in the point analysis lab, where we selected roads within a mile of hospitals.
- We can equally do it to search for other hospitals within a given radius of every other hospital.
- The would then be labeled as neighbors according to our radius rule.

- Likewise, we can calculate the distance matrix between a set of points
- This is usually measured using the standard Euclidean distance
- $d^2 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
- Where x and y are coordinates of the point or polygon in question (selected features), this is the as the crow flies distance. *There are lots of distances*

Spatial Neighbors

- There are many different criteria for deciding if two observations are neighbors
- Generally two observations must be within a critical distance, d, to be considered neighbors.
- This is the Minimum distance criteria, and is very popular.
- This will generate a matrix of binary variables describing the neighborhood.
- We can also describe the neighborhoods in a continuous weighting scheme based on the distance between them

K nearest neighbors

- A useful way to use distances is to construct a k-nearest neighbors set.
- This will find the "k" closest observations for each observation, where k is some integer.
- For instance if we find the k=3 nearest neighbors, then each observation will have 3 neighbors, which are the closest observations to it, *regardless of the distance between them* which is important.
- Using the k nearest neighbor rule, two observations could potentially be very far apart and still be considered neighbors.

Measuring Spatial Autocorrelation

- If we observe data Z(s) (an attribute) at location i, and again at location j, then the spatial autocorrelation between $Z(s)_i$ and $Z(s)_j$ is degree of similarity between them, measured as the standardized covariance between their locations and values.
- In the absence of spatial autocorrelation the locations of $Z(s)_i$ and $Z(s)_j$ has nothing to do with the values of $Z(s)_i$ and $Z(s)_j$
- OTOH, if autocorrelation is present, close proximity of $Z(s)_i$ and $Z(s)_j$ leads to similiarity in their attributes.

Types of autocorrelation

Positive Autocorrelation - This means that a feature is positively associated with the values of the surrounding area (as defined by the spatial weight matrix), high values occur with high values, and low with low

Negative autocorrelation - This means that a feature is negatively associated with the values of the surrounding area (as defined by the spatial weight matrix), high with low, low with high

Measures of autocorrelation

- The (probably) most popular global autocorrelation statistic is Moran's I (1950):
- $I = \frac{n}{(n-1)\sigma^2 w_{..}} \sum_n^i \sum_n^j w_{ij}(Z(s_i) - \bar{Z})(Z(s_j) - \bar{Z})$
- with $Z(s)_i$ being the value of the attribute at location i, $Z(s)_j$ being the value of the attribute at location j, σ^2 is sample variance, w_{ij} is the weight for location ij (0 if they are not neighbors, 1 otherwise).
- Very similar in interpretation ot a Pearson Correlation

Geary's C

- RC Geary in 1954 derived the C statistic
- $C = \frac{n-1}{2\sum_{ij} w_{ij}} \frac{\sum_{ij} w_{ij}(x_i - x_j)^2}{\sum_{ij}(x_i - \bar{x})^2}$
- Similar in interpretation to the Moran statistic, C, measures whether values are similar in neighboring areas.
- C $==$ 1 $==$ No autocorrelation, C$<$ 1 $==$ positive autocorrelation, C $>$ 1 negative autocorrelation

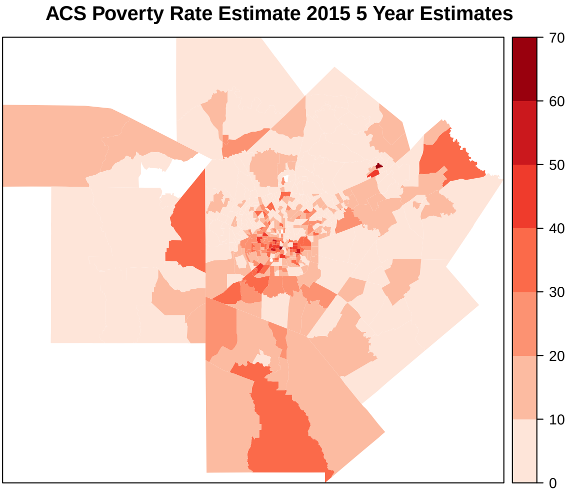
Getis-Ord G

- `"{Too Ugly to Show}"` See the paper
- Similar to Geary's C in interpretation
- High values next to high values, and so on

San Antonio Poverty Rate Map

Here is the overall poverty rate map for San Antonio

- the Global Moran's I is 0.364
- the Geary C value is 0.618



Spatial Lag of a Variable

- If we have a value $Z(s_i)$ at location i and a spatial weight matrix w_{ij} describing the spatial neighborhood around location i, we can find the lagged value of the variable by:
- $WZ_i = Z(s_i) * w_{ij}$
- This calculates what is effectively, the neighborhood average value in locations around location i, often stated $Z(s_{-i})$

- Again, if we had the adjacency matrix from above, a *Rook-based* adjacency weight matrix.

$$w_{ij} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Typically this matrix is standardized, by dividing each element of w_{ij} by the number of neighbors, this is called *row-standardized*:

$$w_{ij} = \begin{bmatrix} 0 & .5 & .5 & 0 \\ .5 & 0 & 0 & .5 \\ .5 & 0 & 0 & .5 \\ 0 & .5 & .5 & 0 \end{bmatrix}$$

and a variable z, equal to:

$$z = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}$$

When we form the product: $z'W$, we get:

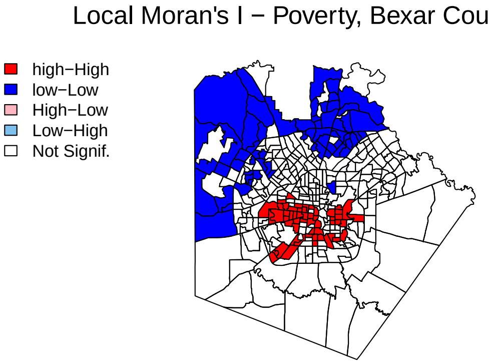
$$z_{lag} = \begin{bmatrix} 2.5 & 2.5 & 2.5 & 2.5 \end{bmatrix}$$

Which, now we see where we get the y of the moran scatterplot. It is just the lagged version of the original variable.

Local Autocorrelation Statistics

- So far, we have only seen a *Global* statistic for autocorrelation, and this tells us if there is any *overall* clustering in our data.
- We may be more interested in *where* the autocorrelation occurs, or where *clusters* are located.
- A local version of the autocorrelation statistics are avaialble as well.
- This basically calculates the statistic from above, but only for the *local neighborhood*.
- It compares the observation's value to the local neighborhood average, instead of the global average. Anselin (1995) referred to this as a "**LISA**" statistic, for Local Indicator of Spatial Autocorrelation.

Here is a LISA map for clusters of poverty in San Antonio:



which shows areas of low poverty clustering in blue, and high poverty clustering in red.

- These are so-called *spatial clusters*, becuase they are areas with higher (or lower, for the blues) than average poverty rates, surrounded by areas with with higher than average poverty rates.
- The red clusters are so called “high-high clusters”, likewise the blue areas are called “low-low clusters”.
- We also see light pink and light blue polygons. The light pink polygons represent areas that have high poverty rates, but are in a low poverty spatial neighborhood, and are called high-low outliers.

What these methods tell you

- all of these statistics are *descriptive statistics ONLY*,
- It simply indicates if there is spatial association/autocorrelation in a variable
- Local autocorrelation statistics tell you if there is significant localized clustering of the variable

Introduction to Spatial Regression Models

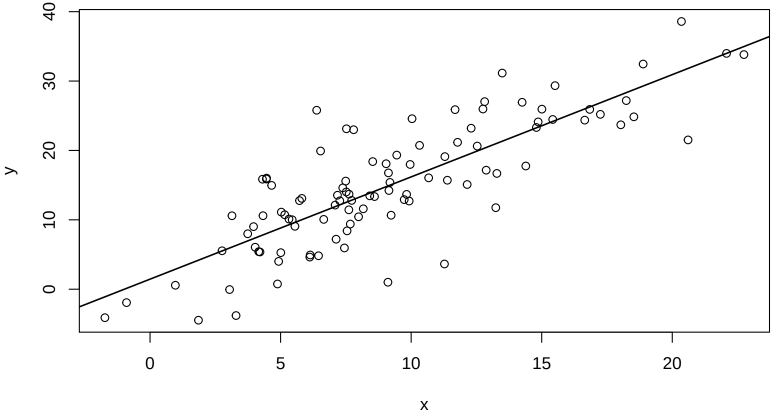
How to break a linear model

- Up until now, we have been concerned with describing the structure of spatial data through correlational, and the methods of exploratory spatial data analysis.
- Through ESDA, we examined data for patterns and using the Moran I and Local Moran I statistics, we examined clustering of variables.
- Now we consider regression models for continuous outcomes. We begin with a review of the Ordinary Least Squares model for a continuous outcome.

OLS Model

- The basic OLS model is an attempt to estimate the effect of an independent variable(s) on the value of a dependent variable. This is written as:
- $y_i = \beta_0 + \beta_1 * x_i + e_i$
- where y is the dependent variable that we want to model,
- x is the independent variable we think has an association with y,
- β_0 is the model intercept, or grand mean of y, when x = 0, and
- β_1 is the slope parameter that defines the strength of the linear relationship between x and y.
- e is the error in the model for y that is unaccounted for by the values of x and the grand mean β_0.

- The average, or expected value of y is : $E[y|x] = \beta_0 + \beta_1 * x_i$, which is the linear mean function for y, conditional on x, and this gives us the customary linear regression plot:



```
##                     Estimate Std. Error   t value      Pr(>|t|)
```

- We assume that the errors, $e_i \sim N(0, \sigma^2)$ are independent, Normally distributed and homoskdastic, with variances σ^2.
- This is the simple model with one predictor. We can easily add more predictors to the equation and rewrite it:
  $y = \beta_0 + \sum^k \beta_k * x_{ik} + e_i$

- So, now the mean of y is modeled with multiple x variables. We can write this relationship more compactly using matrix notation:
- $Y = X'\beta + e$
- Where Y is now a $n * 1$ vector of observations of our dependent variable, X is a $n * k$ matrix of independent variables, with the first column being all 1's and e is the $n * 1$ vector of errors for each observation.

- In matrices this looks like:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$x = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,k} \end{bmatrix}$$

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The residuals are uncorrelated, with covariance matrix $\Sigma =$

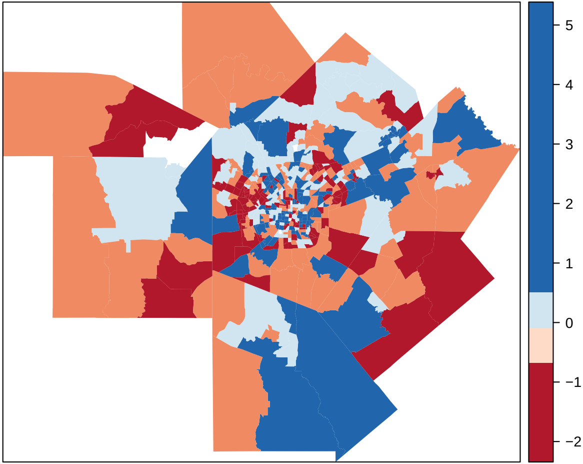
$$\Sigma = \sigma^2 I = \sigma^2 * \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- To estimate the β coefficients, we use the customary OLS estimator
- $\beta = (X'X)^{-1}(X'Y)$
- this is the estimator that minimizes the residual sum of squares:
- $(Y - X'\beta)'(Y - X'\beta)$
- or
- $(Y - \hat{Y})'(Y - \hat{Y})$

Model-data agreement

- Do we (meaning our data) meet the statistical assumptions of our analytical models?
- *Always ask this of any analysis you do, if your model is wrong, your inference will also be wrong.*
- Since spatial data often display correlations amongst closely located observations (autocorrelation), we should probably test for autocorrelation in the model residuals, as that would violate the assumptions of the OLS model.
- One method for doing this is to calculate the value of Moran's I for the OLS residuals.

- Here's a simple OLS model of the form:
- Poverty rate = %with a BA + % Hispanic + % Black
- In R: `lm( pvrty ~bapls+ phspn+pnhbl)`
- I extract the residuals and map them :



Which, in this case, there appears to be significant clustering in the residuals, since the observed value of Moran's I is .221, with a z-test of 8.54, p < .0001

Extending the OLS model to accommodate spatial structure

- If we now assume we measure our Y and X's at specific spatial locations (s), so we have Y(s) and X(s).
- In most analysis, the spatial location (i.e. the county or census tract) only serves to link X and Y so we can collect our data on them, and in the subsequent analysis this spatial information is ignored that explicitly considers the spatial relationships between the variables or the locations.
- In fact, even though we measure Y(s) and X(s) what we end up analyzing X and Y, and apply the ordinary regression methods on these data to understand the effects of X on Y.
- Moreover, we could move them around in space (as long as we keep the observations together y_i with x_i) and still get the same results.

- Such analyses have been called *a-spatial*. This is the kind of regression model you are used to fitting, where we ignore any information on the locations of the observations themselves.
- However, we can extend the simple regression case to include the information on (s) and incorporate it into our models explicitly, so they are no longer *a-spatial*.
- There are several methods by which to incorporate the (s) locations into our models, there are several alternatives to use on this problem:
- The structured linear mixed (multi-level) model, or GLMM (generalized linear mixed model)
- Spatial filtering of observations
- Spatially autoregressive models
- Geographically weighted regression

How to model spatial data correctly

-We will first deal with the case of the spatially autoregressive model, or **SAR model**, as its structure is just a modification of the OLS model from above.

Spatially autoregressive models

We saw in the normal OLS model that some of the basic assumptions of the model are that the: 1) model residuals are distributed as iid standard normal random variates 2) and that they have common (and constant, meaning homoskedastic) unit variance.

- Spatial data, however present a series of problems to the standard OLS regression model. These problems are typically seen as various representations of spatial structure or *dependence* within the data. The spatial structure of the data can introduce spatial dependence into both the outcome, the predictors and the model residuals.
- This can be observed as neighboring observations, both with high (or low) values (positive autocorrelation) for either the dependent variable, the model predictors or the model residuals. We can also observe situations where areas with high values can be surrounded by areas with low values (negative autocorrelation).

- Since the standard OLS model assumes the residuals (and the outcomes themselves) are uncorrelated:
- the autocorrelation inherent to most spatial data introduces factors that violate the iid distributional assumptions for the residuals, and could violate the assumption of common variance for the OLS residuals.
- To account for the expected spatial association in the data, we would like a model that accounts for the spatial structure of the data.
- One such way of doing this is by allowing there to be correlation between residuals in our model, or to be correlation in the dependent variable itself.

- I have introduced with the concept of autoregression amongst neighboring observations.
- This concept is that a particular observation is a linear combination of its neighboring values.
- This autoregression introduces dependence into the data.
- Instead of specifying the autoregression structure directly, we introduce spatial autocorrelation through a global autocorrelation coefficient and a spatial proximity measure.

- There are 2 basic forms of the spatial autoregressive model: the spatial lag and the spatial error models.
- Both of these models build on the basic OLS regression model:
- $Y = X'\beta + e$

The spatial lag model

- The spatial lag model introduces autocorrelation into the regression model by lagging the dependent variables themselves, much like in a time-series approach .
- The model is specified as:
- $Y = \rho WY + X'\beta + e$
- where ρ is the *autoregressive* coefficient, which tells us how strong the resemblance is, on average, between Y_i and it's neighbors. The matrix $\mathbf{W}$ is the spatial weight matrix, describing the spatial network structure of the observations, like we described in the ESDA lecture.

The spatial error model

- The spatial error model says that the autocorrelation is not in the outcome itself, but instead, any autocorrelation is attributable to there being missing *spatial covariates* in the data.
- If these spatially patterned covariates *could* be measures, the the autocorrelation would be 0. This model is written:
- $Y = X'\beta + e$
- $e = \lambda We + v$

- This model, in effect, controls for the nuisance of correlated errors in the data that are attributable to an inherently spatial process, or to spatial autocorrelation in the measurement errors of the measured and possibly unmeasured variables in the model.

Another form of a spatial lag model is the **Spatial Durbin Model** (SDM). This model is an extension of the ordinary lag or error model that includes spatially lagged independent variables.

If you remember, one issue that commonly occures with the lag model, is that we often have residual autocorrelation in the model. This autocorrelation could be attributable to a missing spatial covariate.

We *can* get a kind of spatial covariate by lagging the predictor variables in the model using $\mathbf{W}$.

This model can be written:

$Y = \rho WY + X'\beta + WX\theta + e$

Where, the θ parameter vector are now the regression coefficients for the lagged predictor variables. We can also include the lagged predictors in an error model, which gives us the **Durbin Error Model** (DEM):

$Y = X'\beta + WX\theta + e$

$e = \lambda We + v$

Generally, the spatial Durbin model is preferred to the ordinary error model, because we can include the *unspecified spatial covariates* from the error model into the Durbin model via the lagged predictor variables.

Futher extensions of these models include dependence on both the outcome and the error process. Two models are described in LeSage and Pace. The **Spatial Autocorrelation Model**, or SAC model and the **Spatially autoregressive moving average** model (SARMA model). The SAC model is:

$$Y = \rho W_1 Y + X'\beta + e$$

$$e = \theta W_2 e + v$$

$$Y = (I_n - \rho W_1)^{-1} X'\beta + (I_n - \rho W_1)^{-1}(I_n - \theta W_2)^{-1} e$$

Where, you can potentially have two different spatial weight matrices, W_1 and W_2. Here, the lagged error term is taken over all orders of neighbors, leading to a more *global* error process, The SARMA model has form:

$Y = \rho W_1 Y + X'\beta + u$

$u = (I_n - \theta W_2)e$

$e \sim N(0, \sigma^2 I_n)$

$Y = (I_n - \rho W_1)^{-1} X'\beta + (I_n - \rho W_1)^{-1}(I_n - \theta W_2)e$

which gives a “locally” weighted moving average to the residuals, which will avereage the residuals only in the local neighborhood, instead of over all neighbor orders.

Examination of Model Specification

- To some degree, both of the SAR specifications allow us to model spatial dependence in the data. The primary difference between them is where we model said dependence.
- The lag model says that the dependence affects the dependent variable only, we can liken this to a diffusion scenario, where your neighbors have a diffusive effect on you.
- The error model says that dependence affects the residuals only. We can liken this to the missing spatially dependent covariate situation, where, if only we could measure another really important spatially associated predictor, we could account for the spatial dependence. But alas, we cannot, and we instead model dependence in our errors.

- These are inherently two completely different ways to think about specifying a model, and we should really make our decision based upon how we think our process of interest operates.
- That being said, this way of thinking isn't necessarily popular among practitioners. Most practitioners want the *best fitting model*, 'nuff said. So methods have been developed that test for alternate model specifications, to see which kind of model best summarizes the observed variation in the dependent variable and the spatial dependence.
- These are a set of so-called Lagrange Multiplier (econometrician's jargon for a score test) test. These tests compare the model fits from the OLS, spatial error, and spatial lag models using the method of the score test.

- For those who don't remember, the score test is a test based on the relative change in the first derivative of the likelihood function around the maximum likelihood.
- The particular thing here that is affecting the value of this derivative is the autoregressive parameter, ρ or λ.
- In the OLS model ρ or $\lambda = 0$ (so both the lag and error models simplify to OLS), but as this parameter changes, so does the likelihood for the model, hence why the derivative of the likelihood function is used.
- This is all related to how the estimation routines estimate the value of ρ or λ.

Using the Lagrange Multiplier Test (LMT)

- In general, you fit the OLS model to your dependent variable, then submit the OLS model fit to the LMT testing procedure.
- Then you look to see which model (spatial error, or spatial lag) has the highest value for the test.
- Enter the uncertainty. . .
- So how much bigger, you might say?

- Well, drastically bigger, if the LMT for the error model is 2500 and the LMT for the lag model is 2480, this is NOT A BIG DIFFERENCE, only about 1%.
- If you see a LMT for the error model of 2500 and a LMT for the lag model of 250, THIS IS A BIG DIFFERENCE.
- So what if you don't see a BIG DIFFERENCE, HOW DO YOU DECIDE WHICH MODEL TO USE???
- Well, you could think more, but who has time for that.
- The econometricians have thought up a better LMT test, the so-called robust LMT, robust to what I'm not sure, but it is said that it can settle such problems of a not so big difference between the lag and error model specifications.
- So what do you do? In general, think about your problem before you run your analysis, should this fail you, proceed with using the LMT, if this is inconclusive, look at the robust LMT, and choose the model which has the larger value for this test.

```r
lm.LMtests(model = fit, listw=wts, test = "all")
```

Non-Normal outcomes - So, if you have a normally distributed outcome, then the SAR model is a good choice.

- If your data are multi-level or if your outcome is not normal, then the SAR model is not so good
- This afternoon, we will examine the use of Bayesian models for analyzing both multi-level outcomes or non-normal outcomes.