Geographical Information Systems Institute

Center for Geographic Analysis, Harvard University

Spatial Regression with GeoDa

This lab includes discussion of two types of models of spatial dependence

1) Spatial lag
2) Spatial error

It then shows how to estimate simple spatial regression models in **GeoDa**

1) Classical OLS regression with diagnostics
2) Spatial lag model vs. Spatial error model

Data and Shapefiles:

In this exercise we will use American Community Survey (ACS) 3 year data, ending in 2016. There is also a dataset for the same geography (New York City) from the 2000 decennial census, if you are interested. (Note: GeoDa comes with several data files as samples (e.g., Crime in Columbus [tracts], SIDS in North Carolina [counties])).

The data is in "GeoDa2018" in the GIS Data folder.

The shapefile NYC_Data.shp is the map of New York City with ACS 2016 data. These are socioeconomic attributes for census tracts in the five boroughs. It includes the following variables:

NYC_Data.shp

PCT_WHT	Percent of Population that is white
STATEFP	State FIPS* code
COUNTYFP	County FIPS* code
CT	Census Tract #
TOT-POP	Estimate of total population
TOT_WHT_POP	Estimate of total White population
TOT_BLK_POP	Estimate of total Black population
TOT_ASN_POP	Estimate of total Asian population
HSHLDS_TOT	Estimate of total households
FAM-TOT	Estimate of total families
Highschool	Estimate count of population with a high school diploma
Postseccond	Estimate count of population with some post-secondary education
Bachelor_o	Estimate count of population with a Bachelor's degree or higher
MED_House	Median House Value
MED_Income	Median Household income (derived by compiling three other median income variables

*Federal Information Processing Standard

1. Standard linear regression - *Ordinary Least Squares (OLS)*

The general purpose of linear regression analysis is to find a (linear) relationship between a dependent variable and a set of explanatory variables:

$$y = X\beta + \varepsilon$$

The method of ordinary least squares (OLS) estimation is referred to as a Best Linear Unbiased Estimator (BLUE). The OLS estimates β by minimizing the sum of squared prediction errors, hence, least squares. In order to obtain the BLUE property and make statistical inferences about the population regression coefficients from the estimated b, certain assumptions about the random error of the regression equation need to be made. These include:

a) the random errors have a mean of zero (there is no systematic misspecification or bias in the population regression equation);

b) the random errors have a constant variance (homoskedasticity) and are uncorrelated;

c) the random errors have a normal distribution.

2. Spatial dependence

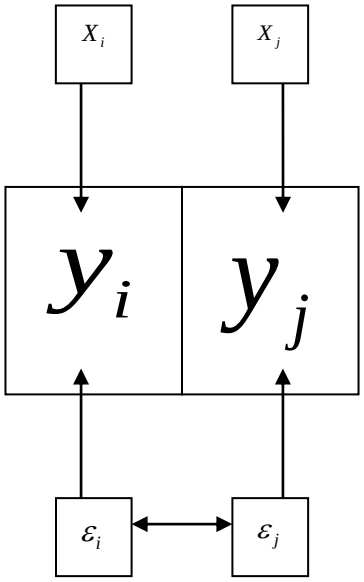
These assumptions may not be always satisfied in practice. When a value observed in one location depends on the values observed at neighboring locations, there is a spatial dependence. And spatial data may show spatial dependence in the variables and error terms.

Why should spatial dependence occur? There are two reasons commonly given. First, data collection of observations associated with spatial units may reflect measurement error. This happens when the boundaries for which information is collected do not accurately reflect the nature of the underlying process generating the sample data.

A second reason for spatial dependence is that the spatial dimension of a social or economic characteristic may be an important aspect of the phenomenon. For example, based on the premise that location and distance are important forces at work, regional science theory relies on notions of spatial interaction and diffusion effects, hierarchies of place and spatial spillovers.

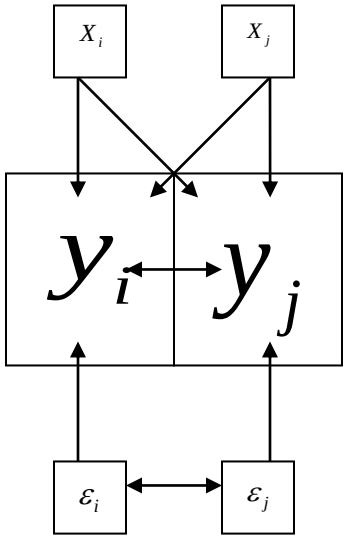
There are two primary types of spatial dependence:

a. Spatial error -the error terms across different spatial units are correlated



With spatial error in OLS regression, the assumption of uncorrelated error terms is violated. As a result, the estimates are inefficient. Spatial error is indicative of omitted (spatially correlated) covariates that if left unattended would affect inference.

b. Spatial lag- the dependent variable y in place i is affected by the independent variables in both place i and j.



With spatial lag in OLS regression, the assumption of uncorrelated error terms is violated; in addition, the assumption of independent observations is also violated. As a result, the estimates are biased and inefficient. Spatial lag is suggestive of a possible diffusion process – events in one place predict an increased likelihood of similar events in neighboring places.

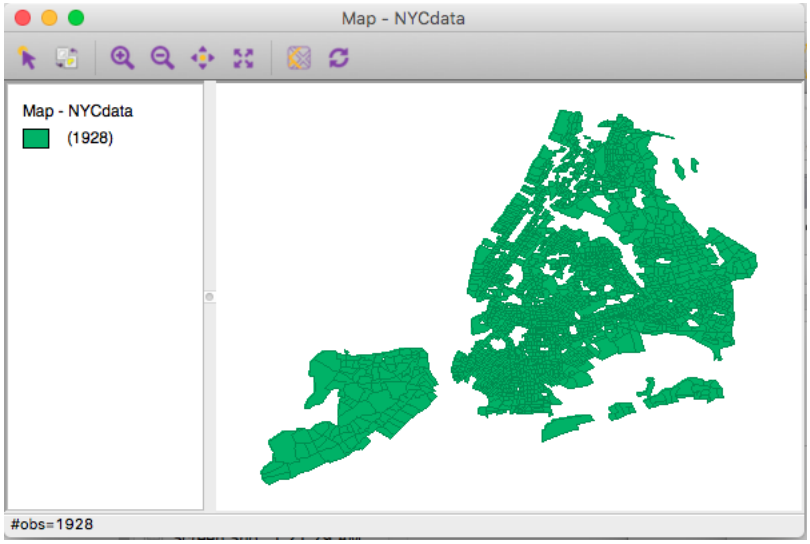
GeoDa provides a range of diagnostics to detect spatial dependence. It also provides unbiased regression estimates using a Maximum Likelihood approach (ML Spatial Lag or Spatial Error models).

3. Spatial regression in GeoDa

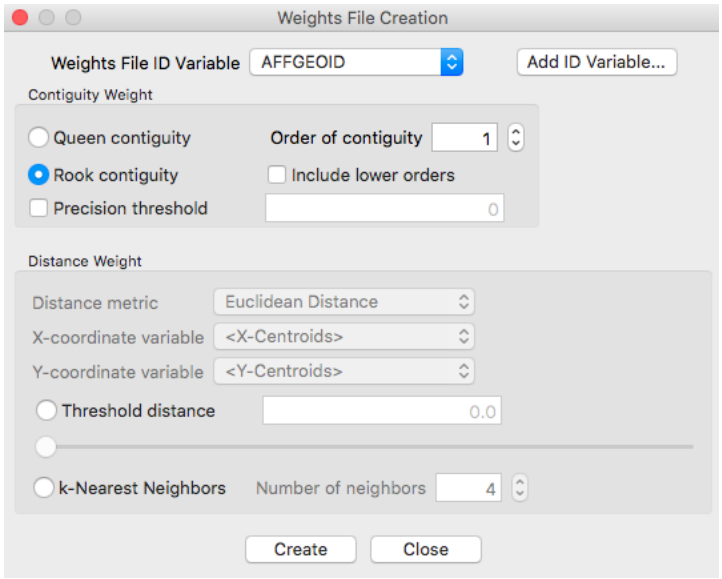
Start a project and create weights matrix

a. Start **GeoDa** by,

b. Go to **File > New Project,** open our most recent NYC data file (the one in which you created the new variables).



c. Create a weights matrix, or use the existing weights file. Go to **Tools > Weights > Create** to open the

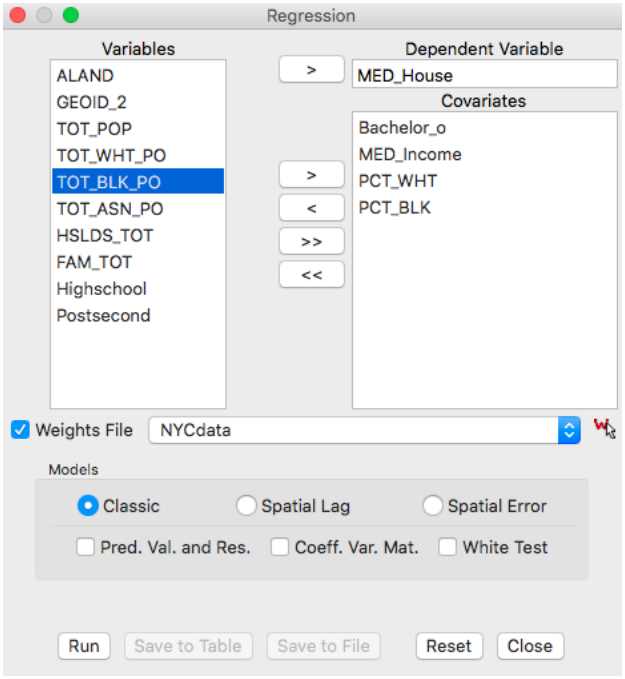


3.1. Classical OLS regression with diagnostics.

Now, we are ready to perform an OLS regression and evaluate the spatial dependence in this regression.

a. On the menu bar, choose **Regression,** then **Regression**. A dialogue box will appear:

b. A variable selection box will appear:



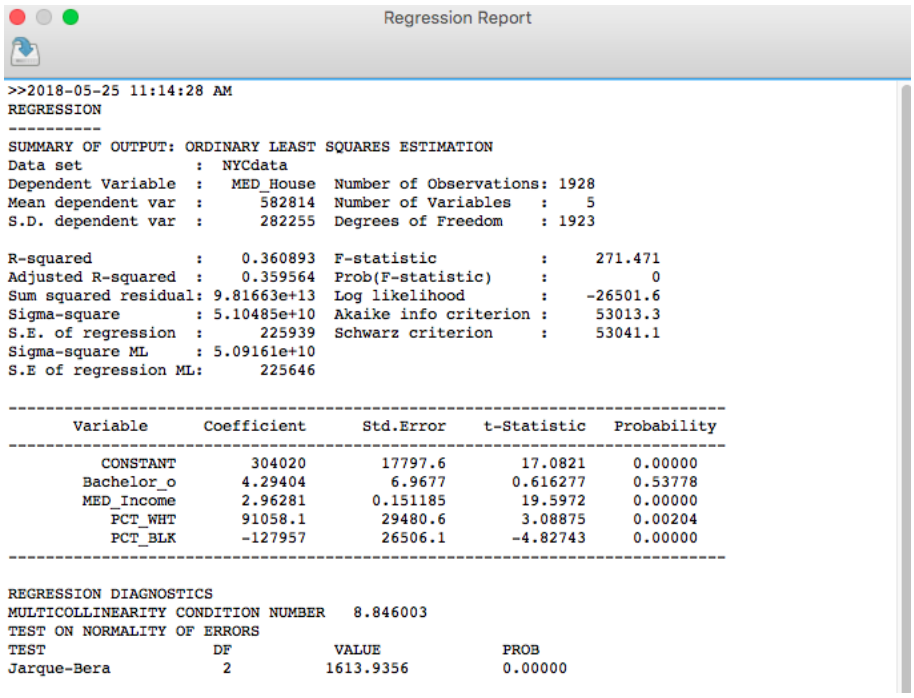
In this example, we will predict House Price (value) with several indicators, % with a bachelor's degree or higher, Median Income, % White and % Black.

Browse to locate the weights matrix file you just created and check the **Weight Files** icon to the right of the file name window.

c. Check **Classic**- this will run classical OLS regression with spatial dependence diagnostics, click **Run.** When done, click **Run.**

d. A new window of regression output will appear, and it has several sections.

1) The first section is the summary output of OLS regression:

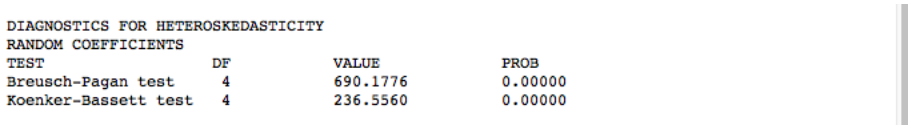


It first shows general information of the run, including the mean and standard deviation of the dependent variable, the model coefficient of determination, F-test probability, and Log likelihood. Then, the coefficients, standard errors, and significance are shown. We can see that among the indicators Incom and % white are positively related to housing value, while % Black is negatively related, and education has an insignificant effect.

2) The next section deals with regression diagnostics:

GeoDa tests multicollinearity of the model- one should be alarmed when **MULTICOLLINEARITY CONDITION NUMBER** is greater than 20. The Jarque-Bera test is used to examine the normality of the distribution of the errors. This tests the combined effects of both skewness and Kurtosis. The low probability of the test score indicates non-normal distribution of the error term. Since the following tests of variance and spatial dependence are conditioned upon normal distribution, in real research, one should be very cautious to interpret the test results. Here, we simply give the illustration of how to interpret the diagnostics when non-normality is not encountered.

3) **DIAGNOSTICS FOR HETEROSKEDASTICITY-** a test of the variance of the error term as the BLUE requires constant error variance.



The low probabilities of the two tests point to existence of heteroskedasticity. This is not necessarily a surprise because the error variance could well be affected by the spatial dependence in the data.

4) **DIAGNOSTICS FOR SPATIAL DEPENDENCE**

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : NYCdata
   (row-standardized weights)
TEST                           MI/DF        VALUE           PROB
Moran's I (error)              0.5188       34.0626         0.00000
Lagrange Multiplier (lag)         1       1167.2037         0.00000
Robust LM (lag)                   1         71.4116         0.00000
Lagrange Multiplier (error)       1       1143.6047         0.00000
Robust LM (error)                 1         47.8127         0.00000
Lagrange Multiplier (SARMA)       2       1215.0164         0.00000
============================== END OF REPORT ==============================
```

There are six tests performed to assess the spatial dependence of the model. First, the Moran's I is highly significant, indicating strong spatial autocorrelation of the residuals. In addition, the output reports the estimates of tests chosen among five statistics for testing for spatial dependence in linear models. The statistics are the simple LM test for a missing spatially lagged dependent variable (Lagrange Multiplier (lag)), the simple LM test for error dependence (Lagrange Multiplier (error)), and robust variants of these are also included (Robust LM (lag) `and` Robust LM (error)- which tests for error dependence and the possible presence of a missing lagged dependent variable, Robust LM (lag) is the other way round), and a portmanteau test (SARMA, in fact Lagrange Multiplier (error) + Robust LM (lag)).

We can see both simple tests of the lag and error are significant, indicating presence of spatial dependence. The robust tests help us understand what type of spatial dependence may be at work. The robust measure for error is still significant, but the robust lag test becomes insignificant, which means that when a lag dependent variable is present the error dependence disappears.

3.2. Spatial lag model

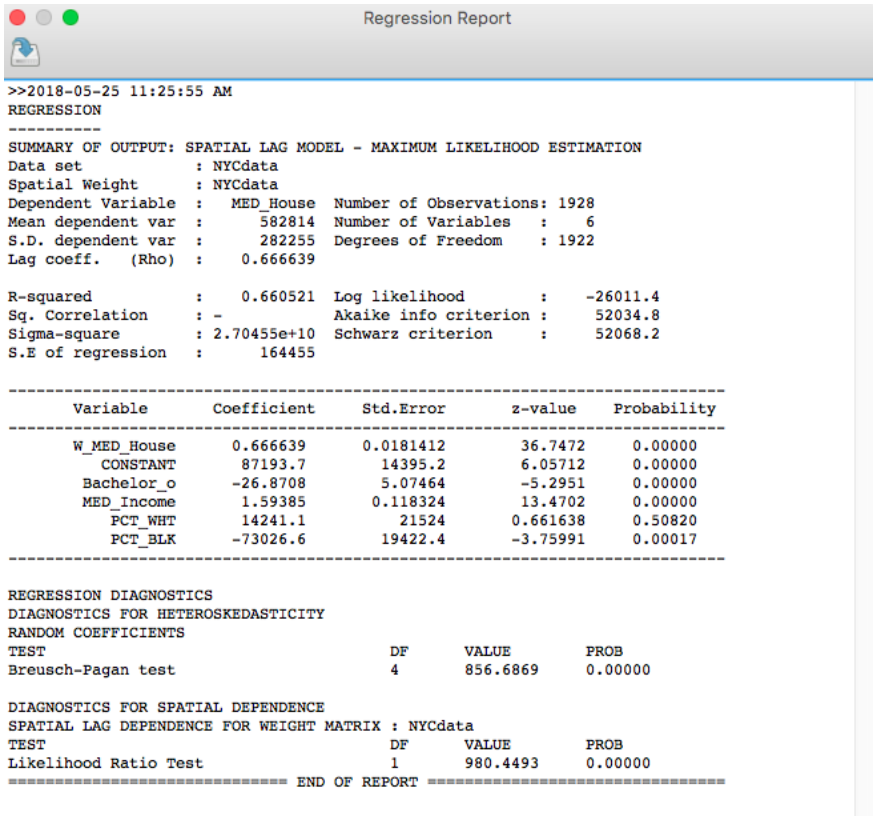
After identifying the presence of spatial dependence, we will use **GeoDa** to re-estimate the model with maximum likelihood approach while controlling for the spatial dependence.

On the menu bar, again choose **Regress**. Check **Moran's I z-value** in the output box, and click **OK**.

In the variable selection box, set dependent and independent variables as before. Check **Weight Files**, and browse to locate the weights matrix file.

Check **Spatial Lag** in the **Models** selection, and click **Run.**

A new window of regression output will appear presenting results from each of the three models of the new Lag model. We should keep the non-spatial OLS results window open for comparison.



Regression Report

```
>>2018-05-25 11:25:55 AM
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : NYCdata
Spatial Weight      : NYCdata
Dependent Variable  :   MED_House  Number of Observations: 1928
Mean dependent var  :      582814  Number of Variables   :    6
S.D. dependent var  :      282255  Degrees of Freedom    : 1922
Lag coeff.   (Rho)  :    0.666639

R-squared           :    0.660521  Log likelihood        :    -26011.4
Sq. Correlation     : -            Akaike info criterion :     52034.8
Sigma-square        : 2.70455e+10  Schwarz criterion     :     52068.2
S.E of regression   :      164455

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
       W_MED_House      0.666639      0.0181412        36.7472      0.00000
          CONSTANT       87193.7        14395.2        6.05712      0.00000
         Bachelor_o      -26.8708        5.07464       -5.2951      0.00000
         MED_Income       1.59385       0.118324        13.4702      0.00000
            PCT_WHT       14241.1          21524       0.661638      0.50820
            PCT_BLK      -73026.6        19422.4       -3.75991      0.00017
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                 DF      VALUE        PROB
Breusch-Pagan test                   4       856.6869     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : NYCdata
TEST                                 DF      VALUE        PROB
Likelihood Ratio Test                1       980.4493     0.00000
============================== END OF REPORT ================================
```

Notice, besides the information that appeared in previous OLS regression output spatial weights file is specified below the data file: rook.GAL. For the lag model we also have a new variable for the spatial lag term of homeownership W_MED_House. (Its coefficient parameter (Rho) reflects the spatial dependence inherent in our sample data, measuring the average influence on observations by their neighboring observations.) It has a positive effect and it is highly significant. As a result, the general model fit improved, as indicated in higher values of R-squared, Log likelihood, AIC (Akaike info Criterion). You should note the absence of adjusted R-squared (which takes into account the degrees of freedom in the model); generally when comparing models of the same set of observations (the same census tracts, in this case) it is advised to use AIC as a measure of model fit, smaller AIC indicates a better fit. The effects of other independent variables remain virtually the same.

The following sections show tests of Heteroskedasticity and Spatial dependence

We can see that the low probability in the Breusch-Pagan test suggests that there is still Heteroskedasticity in the model after introducing the spatial lag term. And in the Likelihood Ratio Test of Spatial Lag Dependence, the result is still significant. Therefore, we conclude that although the introduction of spatial lag term improved the model fit, it didn't make the spatial effects go away.

What should we conclude?

Comparing the spatial lag to the OLS, we can see the new models yield improvement over the non-spatial model. Therefore, we should conclude that controlling spatial dependence will improve our model performance. Understanding the role that spatial dependence of the outcome variable has on predictors is complicated with worthy of attention and thought. For instance when controlling for nearby home values, education becomes a significant predictor (in the negative direction: higher rates of education are associated with lower home values). % white goes from a significant to non-signifcant predictor. Hmmmm…

In this session, we learned to use **GeoDa** to examine and control spatial dependence in the regression models. By now, you should know the general strategy applied in **GeoDa** to deal with the spatial dependence issue. You should be able to understand the various diagnostics and regression tests, and able to interpret the results and make decisions accordingly.