Lab 9. Spatial Regression

SOC 261, Spring 2005
Spatial Thinking in Social Science

Today's task:

1. Basic linear regression
1) BLUE
2) OLS regression

2. Two types of models of spatial dependence
1). Spatial lag
2). Spatial error

3. Spatial regression in **`GeoDa`**
1). Classical OLS regression with diagnostics
2). Spatial lag model vs. Spatial error model
3). Another example

Data and Shapefiles:

The shapefile **newyork.shp** is the map of Manhattan in New York City with Census 2000 data from summary file 3. These are socioeconomic attributes for 297 Census tracts. It includes the following variables:

Vraiable name	Label
POLYID	Polygon ID
STATE	State FIPS
COUNTY	County FIPS
TRACT	Census Tract ID
sctrct00	FIPSID
hvalue	Median housing value
t0_pop	Total population
pctnhw	Percent non-Hispanic white persons
pctnhb	Percent non-Hispanic black persons
pcthsp	Percent Hispanic persons
pctasn	Percent Asian persons
t0p_own	Percent homeowners
t0p_coll	Percent college educated
t0p_prf	Percent of people employed in professional/managerial occupations
t0p_uemp	Percent of people unemployed
t0p_for	Percent foreign born persons
t0p_rec	Percent recent immigrants
t0_minc	Median household income
t0p_poor	Percent total population below poverty

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.

$y = X\beta + \varepsilon$ ---- for population

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. A few of them pertain to our purpose:

a) the random error has mean zero (there is no systematic misspecification or bias in the population regression equation);

b) the random error terms are uncorrelated and have a constant variance (homoskedastic);

c) the random error term follows a normal distribution.

2. Spatial dependence

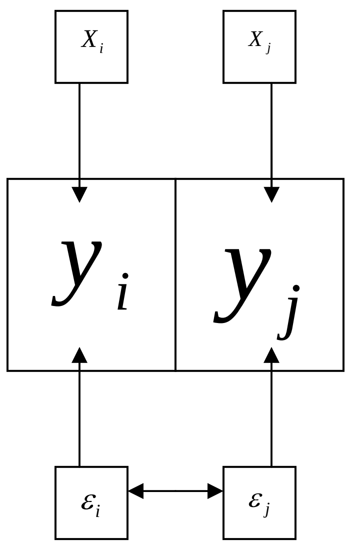
These assumptions may not be always satisfied in practice. When a value observed in one location depends on the values observed at neighboring locations, there is a spatial dependence. And spatial data may show spatial dependence in the variables and error terms.

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 administrative boundaries for collecting information do not accurately reflect the nature of the underlying process generating the sample data.

A second and perhaps more important reason for spatial dependence is that the spatial dimension of social and economic may truly be an important aspect of a modeling. 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.

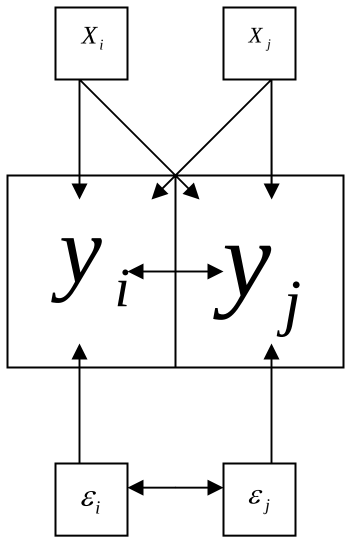
There are two primary types of spatial dependence:

1). 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.

2). 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 with Maximum likelihood approach (ML Spatial Lag or Spatial Error models).

3. Spatial regression in **GeoDa**

1). Starting a project and create weights matrix

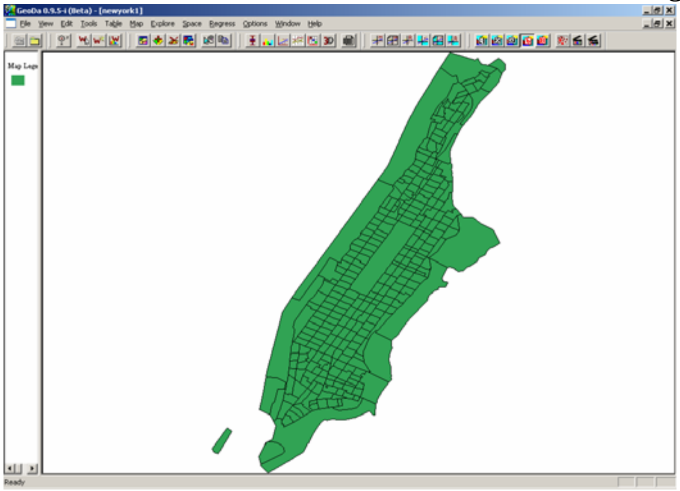
a. Start **GeoDa** by Click the **Start > Programs > Instructional Tools > GeoDa,**

b. Go to **File > Open Project .**

c. **Browse** through the folder to find and select the shape file, **newyork.shp**.

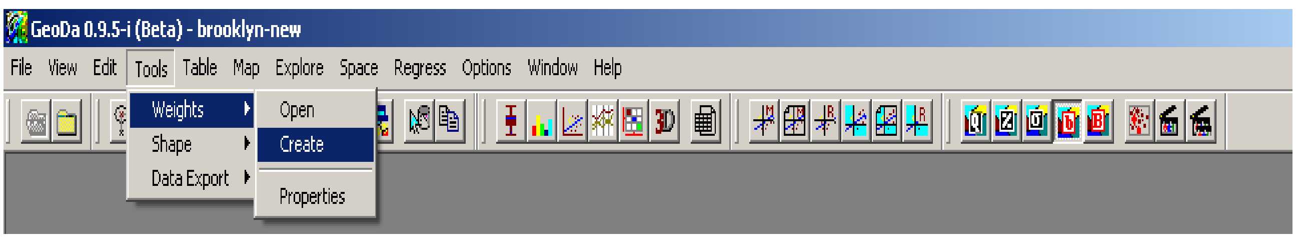
d. Select **POLYID** as the "Key Variable" (by scrolling and clicking on variable name). The Key Variable must have a unique value for each observation (i.e., in this case Census tract). The unique value is used to implement the link between maps and statistical graphs.

e. Click **OK**. Your screen should now look something like …



f. Create a weights matrix. Go to **Tools > Weights > Create** to open the

Creating Weights dialogue box.



In the **Creating weights dialogue** box:

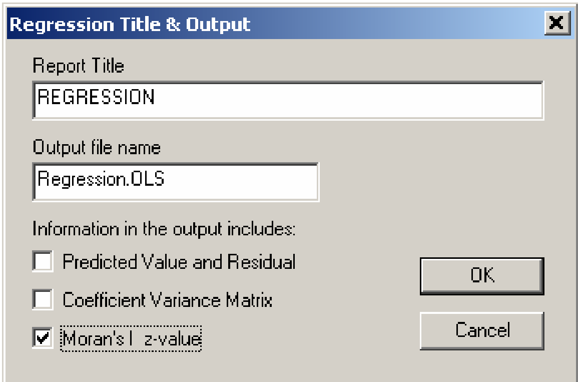
Select **newyork.shp** as the input, **type "rook"** in the **Save output as (**the default extension is **.gal),** Select **POLYID** as the **ID variable for the weights file.**

Select **Rook Contiguity**, click **Create**, then **Done**.

2). Classical OLS regression with diagnostics

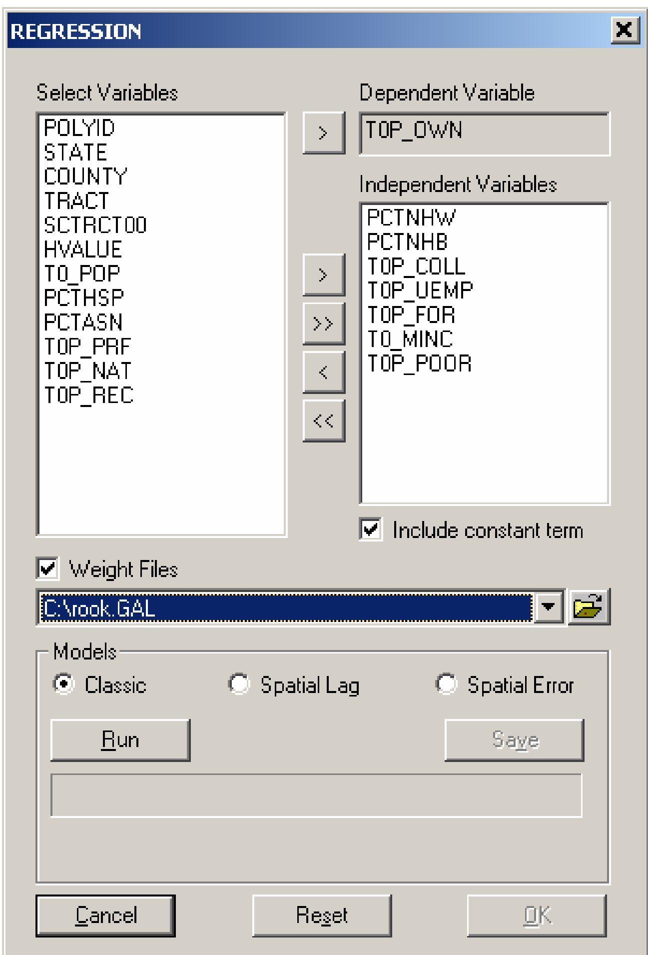
Now, we are ready to perform spatial regression and evaluate the spatial dependence in the regression.

a. On the menu bar, choose **`Regress`**. A dialogue box will appear:



This box inquires the information to be included in the output results. Check **`Moran's I z-value`** as shown, and click **`OK`**.

b. A variable selection box will appear:



In this example, we will predict the homeownership with several indicators, including % non-Hispanic white, % non-Hispanic black, % college+ educated, unemployment rate, % foreign born, median household income, and % person below poverty.

Check **`Weight Files`**, and browse to locate the weights matrix file you just created.

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

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

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

```
GeoDa 0.9.5-i (Beta) - [Regression.OLS]
File  View  Edit  Tools  Table  Map  Explore  Space  Regress  Window  Help

REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set              :  newyork1
Dependent Variable    :      TOP_OWN  Number of Observations:  297
Mean dependent var    :       18.487  Number of Variables   :    8
S.D. dependent var    :      18.7788  Degrees of Freedom    :  289

R-squared             :     0.495409  F-statistic           :     40.5344
Adjusted R-squared    :     0.483187  Prob(F-statistic)     :1.62532e-039
Sum squared residual:        52848.3  Log likelihood        :    -1190.87
Sigma-square          :      182.866  Akaike info criterion :     2397.74
S.E. of regression    :      13.5228  Schwarz criterion     :     2427.29
Sigma-square ML       :       177.94
S.E of regression ML:        13.3394
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT     -0.1177227        4.497595     -0.0261746     0.9790765
      PCTNHW      0.3275164      0.08210561       3.988964     0.0000842
      PCTNHB      0.1429145      0.04944296       2.890493     0.0041380
    TOP_COLL     -0.2569445       0.0814599      -3.154245     0.0017787
    TOP_UEMP      0.1173606       0.1104465       1.062601     0.2888514
     TOP_FOR     0.06952368      0.06435492       1.080316     0.2809013
     TO_MINC    0.000293414    4.489094e-005      6.536152     0.0000000
    TOP_POOR     -0.2369784       0.1065372      -2.224373     0.0268959
-----------------------------------------------------------------------
```

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 seven indicators, % non-Hispanic white, % non-Hispanic black, % college+ educated, and median household income are positively related to homeownership; while poverty rate is negatively related to homeownership; and unemployment rate and % foreign born have insignificant effects.

b). The next section deals with regression diagnostics:

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   18.18592
TEST ON NORMALITY OF ERRORS
TEST                  DF          VALUE             PROB
Jarque-Bera            2           1185.541         0.0000000
```

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 test is a test of 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.

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

```
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF          VALUE             PROB
Breusch-Pagan test     7            102.959         0.0000000
Koenker-Bassett test   7           18.45908         0.0100618
SPECIFICATION ROBUST TEST
TEST                  DF          VALUE             PROB
White                 35           185.7326         0.0000000
```

The low probabilities of the three 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.

d). **DIAGNOSTICS FOR SPATIAL DEPENDENCE**

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : rook.GAL  (row-standardized weights)
TEST                          MI/DF       VALUE          PROB
Moran's I (error)            0.196095     5.7432856      0.0000000
Lagrange Multiplier (lag)       1        21.5140684      0.0000035
Robust LM (lag)                 1         0.1141221      0.7354991
Lagrange Multiplier (error)     1        27.8417603      0.0000001
Robust LM (error)               1         6.4418140      0.0111465
Lagrange Multiplier (SARMA)     2        27.9558824      0.0000009
========================= END OF REPORT ==============================
```

There are six tests performed to assess the spatial dependence of the model. First, Moran's I score of 0.196 is highly significant, indicating strong spatial autocorrelation of the residuals. In addition, the function 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)), variants of these robust to the presence of the other (Robust LM (lag) and Robust LM (error)- which tests for error dependence in 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 lagged dependent variable is present the error dependence disappears.

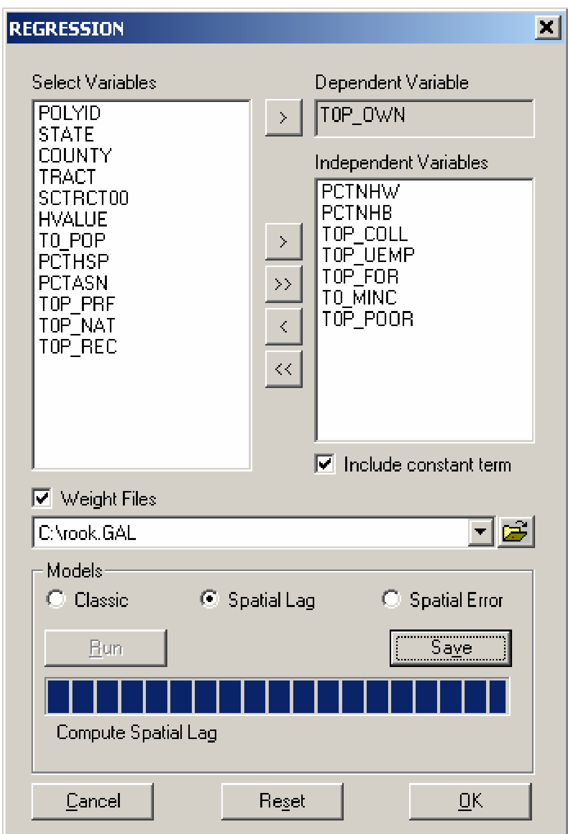
3). Spatial lag model vs. spatial error model

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.

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

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

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



d. Check **`Spatial Error`** and click **`Run`**. Then click **`OK`**.

e. A new window of regression output will appear, and it presents results for both spatial lag and spatial error models.

a). First, we see the results from spatial lag model. Again, it consists of multiple sections,

I). and begins with general information and regression coefficients with significance tests.

```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set              : newyork1
Spatial Weight        : rook.GAL
Dependent Variable    :     T0P_OWN  Number of Observations:  297
Mean dependent var    :      18.487  Number of Variables   :    9
S.D. dependent var    :     18.7788  Degrees of Freedom    :  288
Lag coeff.   (Rho)    :    0.244682

R-squared             :    0.526532  Log likelihood        :    -1183.22
Sq. Correlation       : -            Akaike info criterion :     2384.44
Sigma-square          :     166.965  Schwarz criterion     :     2417.69
S.E of regression     :     12.9215

-----------------------------------------------------------------------
    Variable    Coefficient     Std.Error    z-value      Probability
-----------------------------------------------------------------------
   W_T0P_OWN     0.2446815     0.07099979      3.446228    0.0005686
    CONSTANT     -2.435339       4.334687    -0.5618259    0.5742346
      PCTNHW     0.2890341     0.08031075      3.598947    0.0003196
      PCTNHB     0.1438118     0.04732972       3.03851    0.0023776
    T0P_COLL     -0.238261     0.07828735     -3.043416    0.0023392
    T0P_UEMP     0.1286424      0.1055521      1.218757    0.2229365
     T0P_FOR    0.08310842     0.06151475      1.351032    0.1766852
     T0_MINC  0.0002427569  4.391055e-005      5.528442    0.0000000
    T0P_POOR    -0.2224669      0.1020317      -2.18037    0.0292299
-----------------------------------------------------------------------
```

Notice, besides the info appeared in previous OLS regression output, we have a designated spatial weight file: rook.GAL. And the spatial lag term of homeownership W_T0P_OWN appeared as additional indicator. (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 positive effect and it is highly significant. As a result, the general model fit improved, as indicated in higher values of R-squared and Log likelihood. The effects of other independent variables remain virtually the same.

II). The following sections show tests of Heteroskedasticity and Spatial dependence.

```
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      7        104.1549      0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : rook.GAL
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1        15.29485      0.0000920
======================== END OF REPORT ============================
```

We can see that the low probability in 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.

b). Now let's try the Spatial Error model.

```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set             : newyork1
Spatial Weight       : rook.GAL
Dependent Variable   :     TOP_OWN  Number of Observations:  297
Mean dependent var   :   18.487003  Number of Variables   :    8
S.D. dependent var   :   18.778784  Degree of Freedom     :  289
Lag coeff. (Lambda)  :    0.299826

R-squared            :    0.533847  R-squared (BUSE)      : -
Sq. Correlation      : -            Log likelihood        :-1181.858166
Sigma-square         :  164.385385  Akaike info criterion :     2379.72
S.E of regression    :     12.8213  Schwarz criterion     : 2409.266190

-----------------------------------------------------------------------
    Variable    Coefficient      Std.Error    z-value      Probability
-----------------------------------------------------------------------
    CONSTANT   -0.5642749       4.471026      -0.126207     0.8995680
      PCTNHW    0.3535008     0.07917054       4.465055     0.0000080
      PCTNHB    0.1686377     0.05519698       3.055198     0.0022493
    TOP_COLL   -0.2548309     0.08248667      -3.089358     0.0020060
    TOP_UEMP    0.1141496      0.1054046       1.082966     0.2788234
     TOP_FOR   0.08326652     0.06332542       1.314899     0.1885440
     TO_MINC   0.0002531006     4.596173e-005       5.506767     0.0000000
    TOP_POOR   -0.2535857      0.1014921      -2.498575     0.0124694
      LAMBDA    0.2998257     0.08042689       3.727928     0.0001931
-----------------------------------------------------------------------
```

I). In comparison with the Spatial Lag model output, we also have a designated spatial weight file: rook.GAL. And a coefficient on the spatially correlated errors (LAMBDA) is added as additional indicator. It has positive effect and it is highly significant. As a result, the general model fit improved, as indicated in higher values of R-squared and Log likelihood. Like the lag model, the effects of other independent variables remain virtually the same.

II). Diagnostics of Heteroskedasticity and spatial independence

```
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE           PROB
Breusch-Pagan test                      7        114.8941       0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : rook.GAL
TEST                                    DF      VALUE           PROB
Likelihood Ratio Test                   1        18.02336       0.0000218
========================= END OF REPORT ==============================
```

Similar to the lag model, the Heteroskedasticity test remains significant. Also, the Likelihood Ratio Test of Spatial Error Dependence has significant result. Therefore, we conclude that although allowing the error terms to be spatially correlated improved the model fit, it didn’t make the spatial effects go away.

c). What should we conclude?

Comparing the spatial lag and spatial error models, we can see both alternative models yield improvement to the original OLS model. Therefore, we should conclude that controlling spatial dependence will improve our model performance. Now the question is which of the two models is better? To some extent, this is an open question. The general advice is first to look for theoretical basis to inform your choice. If there are strong substantive grounds for one model instead of the other, you should adopt it. When it is not so clear theoretically, you can compare the model performance parameters: the R-squared and Log likelihood. In our case, the spatial error model has greater R-squared and Log likelihood values, therefore, that could be our reason to adopt this solution.

4) Another example

Now, do another exercise with % person below poverty as dependent variable, and % non-Hispanic white, % non-Hispanic black, % college+ educated, unemployment rate, % foreign born, and median household income as independent variables. Run OLS regression, interpret the regression and spatial dependence diagnostics; if spatial dependence exists, try to re-estimate the models with Spatial lag model and Spatial Error model, then compare the output and make your choice.

4. Summary.

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.

This effectively concludes our lab work on **GeoDa**.