GIS & SPATIAL ANALYSIS - SOC SCI, 2023

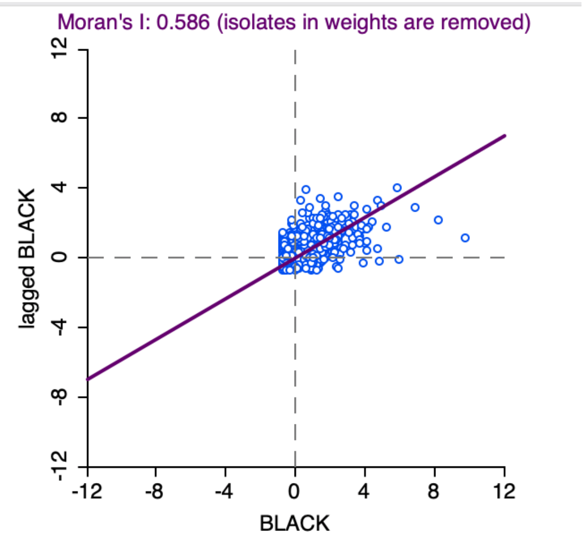
Author: Oguz Alp Eren

Homework 6: Spatial Autocorrelation Diagnostic and Spatial Lag/Error Models

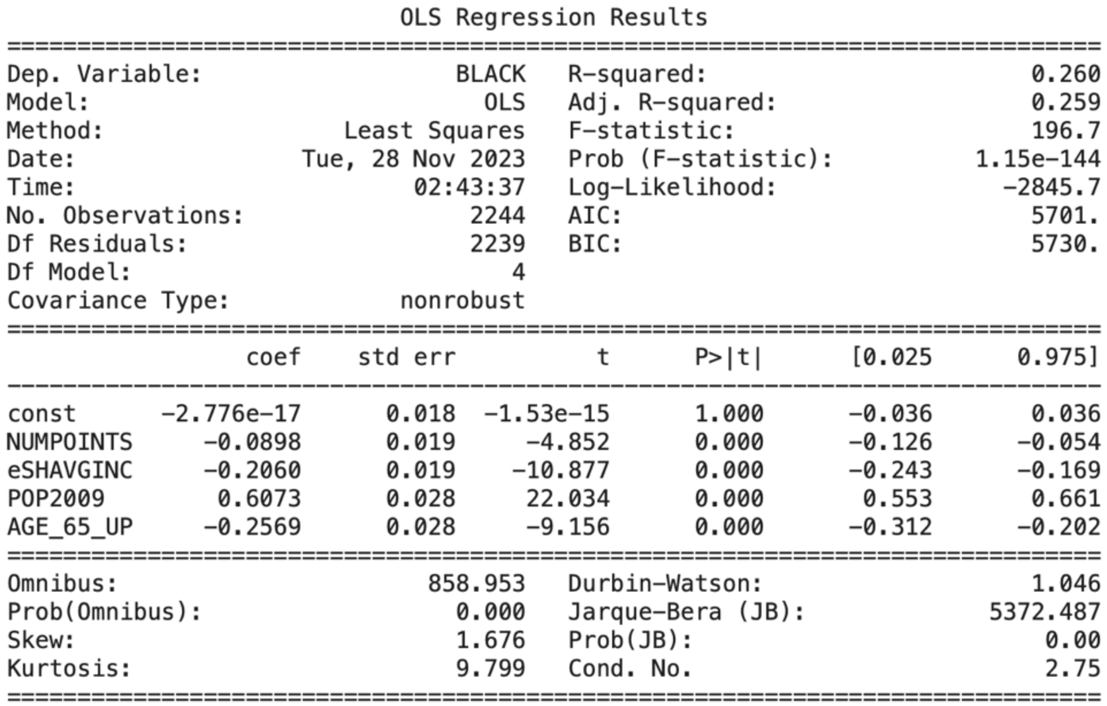
Model 1, Ordinary Least Squares:

I am testing whether there is a significant correlation between the black population and number of trees in tracts in NYC. I have included household income, population, and age 65 and up as control variables. These control variables serve the purpose of accounting for potential influences on the relationship between the number of trees and the Black population. Household income helps control for economic disparities that might affect resource allocation, while population size considers the demand for amenities like green spaces. The proportion of the population aged 65 and older acknowledges the potential influence of age-related preferences for outdoor spaces.

Queen Contiguity defines spatial relationships between geographic units based on shared edges or vertices. In this case, it considers two geographic units as neighbors if they share a common boundary or vertex. This weight structure is appropriate for considering both contiguity and adjacency in defining spatial relationships.



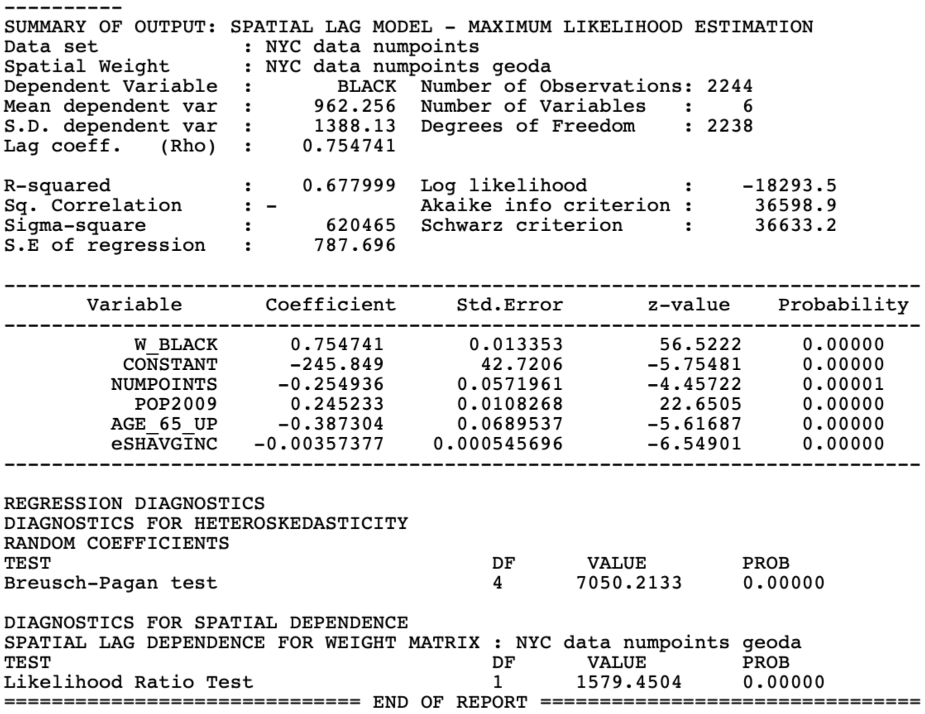
The Moran's I statistic with a value of 0.58 indicates that the distribution of the Black population in NYC exhibits positive spatial autocorrelation, signifying that areas with similar Black population values tend to cluster together in space. This spatial pattern suggests that neighborhoods or regions with a high concentration of Black residents are geographically adjacent to one another, and the observed clustering is statistically significant. This insight reveals a non-random spatial distribution of the Black population.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  BLACK   R-squared:                       0.260
Model:                            OLS   Adj. R-squared:                  0.259
Method:                 Least Squares   F-statistic:                     196.7
Date:                Tue, 28 Nov 2023   Prob (F-statistic):         1.15e-144
Time:                        02:43:37   Log-Likelihood:                -2845.7
No. Observations:                2244   AIC:                             5701.
Df Residuals:                    2239   BIC:                             5730.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2.776e-17      0.018  -1.53e-15      1.000      -0.036       0.036
NUMPOINTS     -0.0898      0.019     -4.852      0.000      -0.126      -0.054
eSHAVGINC     -0.2060      0.019    -10.877      0.000      -0.243      -0.169
POP2009        0.6073      0.028     22.034      0.000       0.553       0.661
AGE_65_UP     -0.2569      0.028     -9.156      0.000      -0.312      -0.202
==============================================================================
Omnibus:                      858.953   Durbin-Watson:                   1.046
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5372.487
Skew:                           1.676   Prob(JB):                         0.00
Kurtosis:                       9.799   Cond. No.                         2.75
==============================================================================
```

This output summarizes the results of an Ordinary Least Squares (OLS) regression analysis. The dependent variable "BLACK" has an R-squared of 0.260, indicating that approximately 26% of its variability is explained by the model, which includes four predictors (NUMPOINTS, eSHAVGINC, POP2009, AGE_65_UP). The coefficients for these predictors, along with their p-values, suggest varying levels of statistical significance, with all predictors appearing to be significantly associated with the dependent variable at conventional levels (p < 0.05).

Model 2, Spatial Lag:

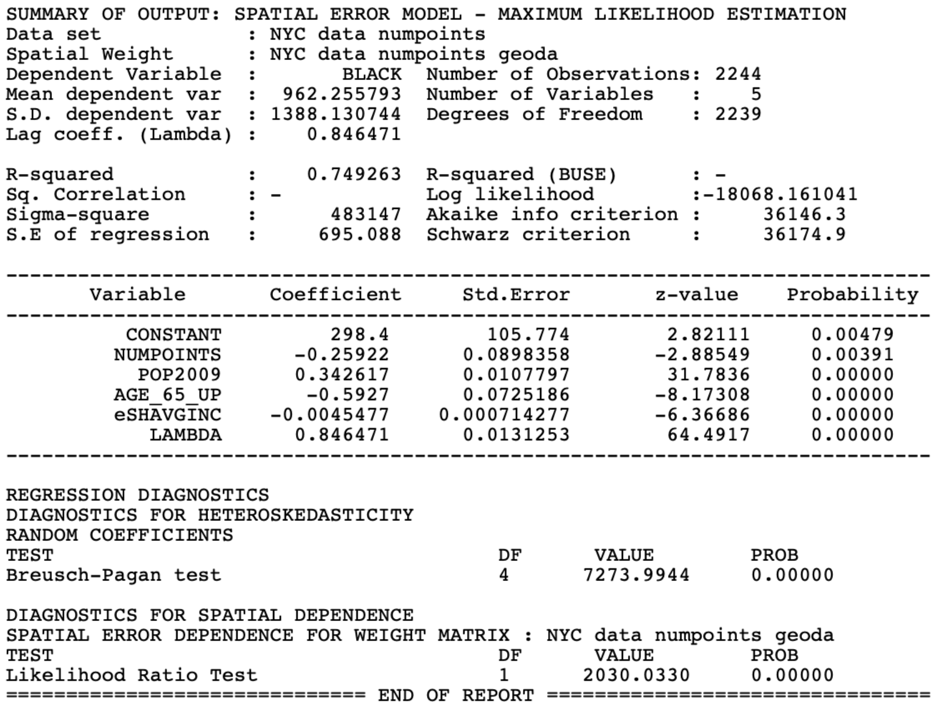


```
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : NYC data numpoints
Spatial Weight      : NYC data numpoints geoda
Dependent Variable  :        BLACK  Number of Observations: 2244
Mean dependent var  :      962.256  Number of Variables   :    6
S.D. dependent var  :      1388.13  Degrees of Freedom    : 2238
Lag coeff.   (Rho)  :     0.754741

R-squared           :     0.677999  Log likelihood        :    -18293.5
Sq. Correlation     : -             Akaike info criterion :     36598.9
Sigma-square        :       620465  Schwarz criterion     :     36633.2
S.E of regression   :      787.696

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
            W_BLACK      0.754741       0.013353        56.5222     0.00000
           CONSTANT      -245.849        42.7206       -5.75481     0.00000
          NUMPOINTS     -0.254936      0.0571961       -4.45722     0.00001
            POP2009      0.245233      0.0108268        22.6505     0.00000
          AGE_65_UP     -0.387304      0.0689537       -5.61687     0.00000
          eSHAVGINC   -0.00357377    0.000545696       -6.54901     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                      DF      VALUE        PROB
Breusch-Pagan test                        4      7050.2133     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : NYC data numpoints geoda
TEST                                      DF      VALUE        PROB
Likelihood Ratio Test                     1      1579.4504     0.00000
============================== END OF REPORT ================================
```

The model reveals a significant positive spatial autocorrelation in the distribution of the Black population (Rho ≈ 0.755). This spatial pattern indicates that areas with a higher concentration of Black residents tend to be spatially clustered, highlighting a non-random distribution. The model's R-squared value of approximately 0.678 suggests that the included variables collectively explain a substantial portion of the variance in the Black population, indicating a reasonably good fit.

Delving into specific coefficients, the negative coefficient for "NUMPOINTS" (-0.255) implies a potential relationship between the number of trees and the Black population. As the number of points increases, the Black population tends to decrease, suggesting that neighborhoods with more trees may have a lower Black population concentration. Additionally, other covariates such as "POP2009," "AGE_65_UP," and "eSHAVGINC" also exhibit significant relationships with the Black population, further underscoring the complexity of factors influencing demographic distribution. Overall, this analysis unveils insights into the spatial distribution of the Black population in New York City, highlighting the presence of spatial autocorrelation and its associations with various socioeconomic and environmental factors.

Model 3, Spatial Error:



```
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : NYC data numpoints
Spatial Weight      : NYC data numpoints geoda
Dependent Variable  :       BLACK  Number of Observations: 2244
Mean dependent var  :  962.255793  Number of Variables   :    5
S.D. dependent var  : 1388.130744  Degrees of Freedom    : 2239
Lag coeff. (Lambda) :    0.846471

R-squared           :    0.749263  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-18068.161041
Sigma-square        :      483147  Akaike info criterion :     36146.3
S.E of regression   :     695.088  Schwarz criterion     :     36174.9

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT         298.4        105.774        2.82111     0.00479
         NUMPOINTS      -0.25922      0.0898358       -2.88549     0.00391
           POP2009      0.342617      0.0107797        31.7836     0.00000
         AGE_65_UP       -0.5927      0.0725186       -8.17308     0.00000
         eSHAVGINC    -0.0045477    0.000714277       -6.36686     0.00000
            LAMBDA      0.846471      0.0131253        64.4917     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       4      7273.9944     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : NYC data numpoints geoda
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1      2030.0330     0.00000
============================== END OF REPORT ================================
```

The model indicates a strong positive spatial autocorrelation, denoted by the spatial error coefficient (Lambda) of approximately 0.846471. This suggests that there is spatial clustering in the distribution of the Black population across New York City. Areas with a high concentration of Black residents tend to be spatially adjacent to other areas with similar characteristics, revealing a non-random pattern in demographic distribution. The R-squared value of around 0.749263 implies that the model explains a significant proportion of the variance in the Black population, indicating its effectiveness in capturing spatial relationships. The covariates, including "NUMPOINTS," "POP2009," "AGE_65_UP," and "eSHAVGINC," exhibit significant associations with the Black population, underscoring the multifaceted nature of factors influencing demographic patterns. Additionally, the model reports diagnostics for heteroskedasticity, which indicate potential variability in the error terms, and spatial dependence, confirming the presence of spatial autocorrelation.

This spatial error model provides insights into the spatial distribution of the Black population in New York City. The strong spatial autocorrelation suggests that areas with similar demographic characteristics are clustered together, revealing a spatially dependent pattern. The model demonstrates the importance of considering spatial relationships when analyzing demographic distributions, highlighting the significance of various socioeconomic and environmental factors in shaping these patterns.

Model Differences:

The fundamental difference between these models lies in how they account for spatial dependence. In the spatial lag model, the emphasis is on the direct influence of neighboring observations on the dependent variable. It assumes that the value of the dependent variable in one location is influenced by the values of the same variable in adjacent areas, quantified by the spatial lag coefficient (Rho). This model is suitable when there is a theoretical basis for direct spatial interaction, such as in cases where a change in a variable in one location directly impacts nearby locations.

Conversely, the spatial error model takes a different approach by considering spatial autocorrelation as residing in the error terms of the model. It acknowledges that unobserved spatial factors or omitted variables may lead to spatial correlation in the residuals, and this spatial autocorrelation is captured by the spatial error coefficient (Lambda). The spatial error model is chosen when spatial autocorrelation is attributed to unobserved factors or measurement error, and it accounts for the indirect or unexplained spatial effects that influence the residuals.

Compared to the Ordinary Least Squares (OLS) and spatial lag models, the spatial error model appears to be the most suitable choice based on the quantifiable outputs. While OLS lacks the consideration of spatial autocorrelation, the lag model primarily focuses on direct spatial interactions. In contrast, the spatial error model accounts for unobserved spatial factors affecting the residuals, which aligns with the strong spatial autocorrelation observed in the Black population distribution. The spatial error model's Lambda coefficient of approximately 0.846 reflects this spatial autocorrelation, while also achieving a high R-squared value of around 0.749263, indicating a good fit. Therefore, the spatial error model provides a more comprehensive understanding of the complex spatial relationships affecting the Black population in New York City, making it the preferred model.

The hypothesis that the number of trees is correlated with the Black population in New York City is supported by the results. The spatial error model, which accounts for unobserved spatial factors, shows a strong spatial autocorrelation in the Black population distribution, indicating a non-random pattern that is likely influenced by various factors, including the presence of trees.