Introduction to Spatial Statistics GIS

Jennie Murack, MIT Libraries, 2015

What is spatial statistics?

- methods for analyzing spatial distributions, patterns, processes, and relationships.
- developed specifically for use with geographic data
- unlike traditional non-spatial statistical methods, they incorporate space (proximity, area, connectivity, and/or other spatial relationships) directly into their mathematics.

Spatial Autocorrelation

Objectives

- Understand the concept of spatial autocorrelation
- Learn which tools to use in Geoda and ArcMap to test for autocorrelation
- Interpret output from spatial autocorrelation tests

Software

- ArcGIS
  - Complete GIS software with hundreds of tools
  - Can work with several datasets (layers) at once.
  - www.arcgis.com

- GeoDa – open source
  - Solely for spatial statistics
  - Use one dataset (layer) at a time.
  - Simple, easy-to-use, interface
  - Available with at: http://geodacenter.asu.edu/

What is spatial autocorrelation?

What is spatial autocorrelation?

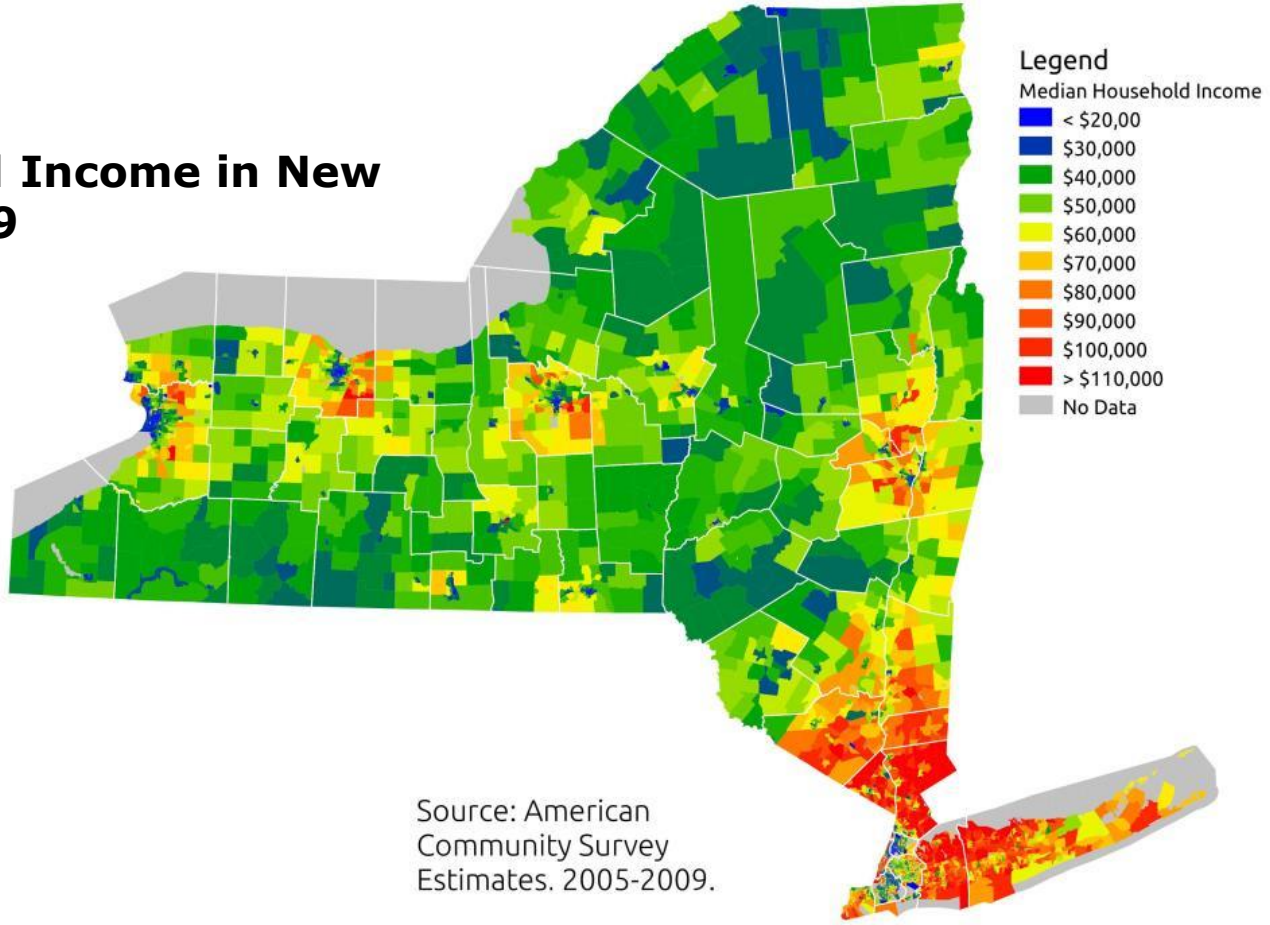
- Based on Tobler's first law of geography, "Everything is related to everything else, but near things are more related than distant things."
- It's the correlation of a variable with itself through space (only 1 variable is involved).
- Patterns may indicate that data are not independent of one another, violating the assumption of independence for some statistical tests.

Tests for spatial autocorrelation will allow you to answer the following questions about your data:

- How are the features distributed?
- What is the pattern created by the features?
- Where are the clusters?
- How do patterns and clusters of different variables compare to one another?

You can measure the pattern formed by the **location of features** or patterns of **attribute values** associated with features (ex. median home value, percent female, etc.).

Median Household Income in New York State, 2005-9



Conceptual Models

Spatial Neighborhoods and Weights

- Neighborhood = area in which the GIS will compare the target values to neighboring values
- Neighborhoods are most often defined based on adjacency or distance, but can be defined based on travel time, travel cost, etc.
- You can also define a cutoff distance, the amount of adjacency (borders vs. corners), or the amount of influence at different distances
- A table of spatial weights is used to incorporate these definitions into statistical analysis.

Distance Models

- Inverse distance – all features influence all other features, but the closer something is, the more influence it has
- Distance band – features outside a specified distance do not influence the features within the area
- Zone of indifference – combines inverse distance and distance band

Adjacency Models

- K Nearest Neighbors – a specified number of neighboring features are included in calculations
- Polygon Contiguity – polygons that share an edge or node influence each other
- Spatial weights – specified by user (ex. Travel times or distances)

Types of Contiguity

- Rook = share edges
- Bishop = share corners
- Queen = share edges or corners
- Secondary order contiguity = neighbor of neighbor

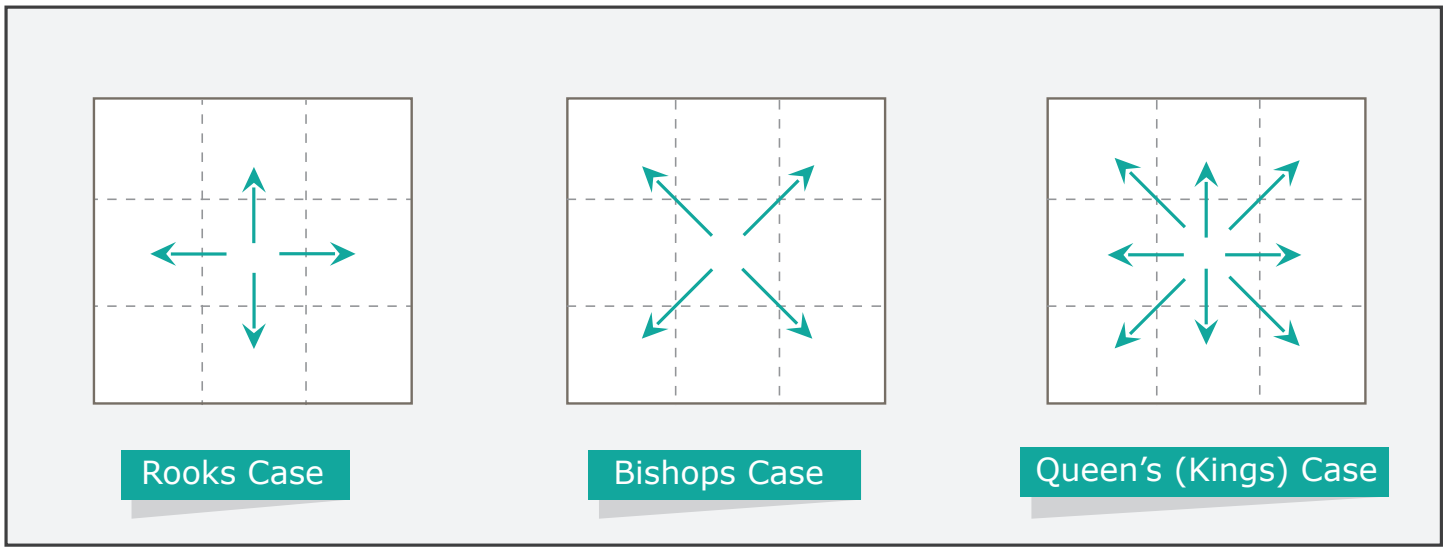


Image by MIT OpenCourseWare.

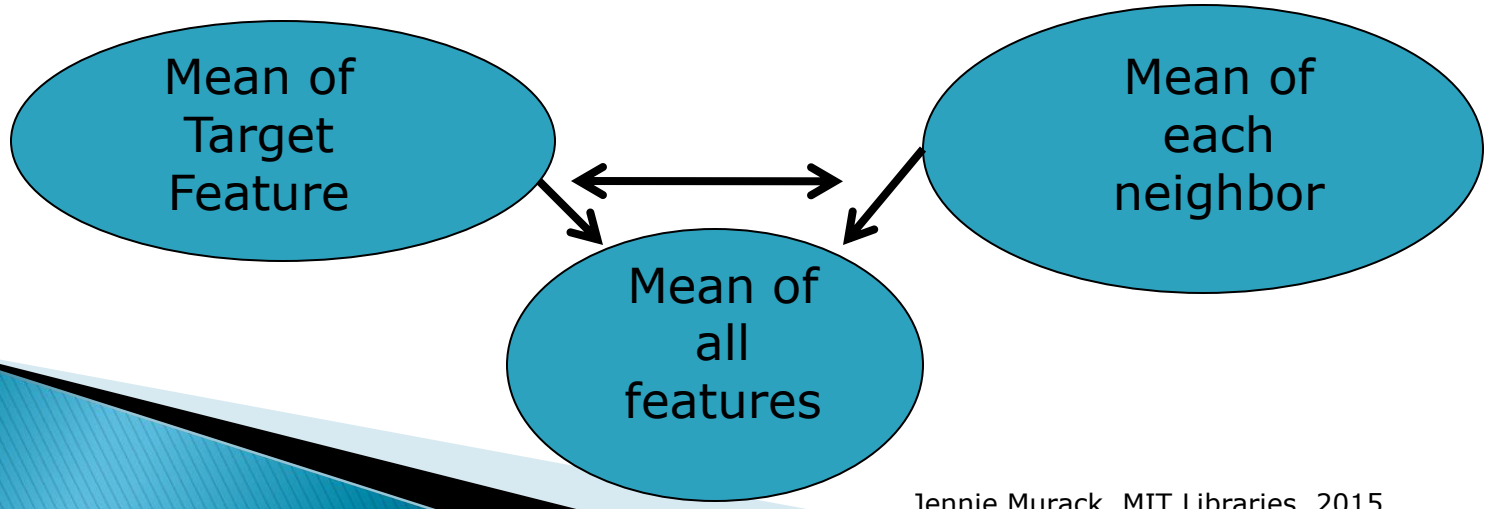
Measuring Data Values

Global vs. Local Statistics

- Global statistics – identify and measure the pattern of the entire study area
  - Do not indicate where specific patterns occur
- Local Statistics – identify variation across the study area, focusing on individual features and their relationships to nearby features (i.e. specific areas of clustering)

Spatial Autocorrelation (Moran's I)

- **Global** statistic
- Measures whether the pattern of feature ***values*** is clustered, dispersed, or random.
- Compares the difference between the mean of the target feature and the mean for all features to the difference between the mean for each neighbor and the mean for all features.
- For more information on the equation, see ESRI online help



Anselin Local Moran's I

- **Local** statistic
- Measures the strength of patterns for **each specific feature**.
- Compares the value of each feature in a pair to the mean value for all features in the study area.
- See ESRI online help for the equation.

Anselin Local Moran's I

- Positive I value:
  - Feature is surrounded by features with similar values, either high or low.
  - Feature is part of a cluster.
  - Statistically significant clusters can consist of high values (HH) or low values (LL)
- Negative I value:
  - Feature is surrounded by features with dissimilar values.
  - Feature is an outlier.
  - Statistically significant outliers can be a feature with a high value surrounded by features with low values (HL) or a feature with a low value surrounded by features with high values (LH).

Getis-Ord General G

- **Global** statistic
- Indicates that high or low ***values*** are clustered
- The value of the target feature itself is not included in the equation so it is useful to see the effect of the target feature on the surrounding area, such as for the dispersion of a disease.
- Works best when either high or low values are clustered (but not both).
- See ESRI online help for the equation.

Hot Spot Analysis (Getis-Ord Gi*)

- **Local** version of the G statistic
- The value of the target feature is included in analysis, which shows where hot spots (clusters of high values) or cold spots (clusters of low values) exist in the area.
- To be statistically significant, a hot spot has a higher value than its neighbors, which also have high values. The opposite is true of cold spots.
- See ESRI online help for the equation.

Hot Spot Analysis vs. Local Moran's I

- G statistics are useful when negative spatial autocorrelation (outliers) is negligible.

- Local Moran's I calculates spatial outliers.

Resources

- ESRI Spatial Statistics Website:
  http://blogs.esri.com/esri/arcgis/2010/07/13/spatial-statistics-resources/
- Geoda Workbook:
  https://geodacenter.asu.edu/og_tutorials
- ESRI Spatial Statistics Tool help:
  http://resources.arcgis.com/EN/HELP/MAIN/10.2/index.html#/An_overview_of_the_Spatial_Statistics_toolbox/005p00000002000000/

ESRI Spatial Statistics Links

- Spatial Autocorrelation (Moran's I)
  - http://resources.arcgis.com/EN/HELP/MAIN/10.2/index.html#/How_Spatial_Autocorrelation_Global_Moran_s_I_works/005p0000000t000000/
- Anselin Local Moran's I
  - http://resources.arcgis.com/en/help/main/10.1/index.html#/How_Cluster_and_Outlier_Analysis_Anselin_Local_Moran_s_I_works/005p00000012000000/
- Getis Ord General G
  - http://resources.arcgis.com/en/help/main/10.1/index.html#/How_High_Low_Clustering_Getis_Ord_General_G_works/005p0000000q000000/
- Hot Spot Analysis (Getis Ord Gi*)
  - http://resources.arcgis.com/en/help/main/10.1/index.html#/How_Hot_Spot_Analysis_Getis_Ord_Gi_works/005p00000011000000/

Regression

How is Regression Different from other Spatial Statistical Analyses?

With other spatial statistics tools you ask **WHERE** something is happening?

- Are there places in the United States where people are persistently dying young?
- Where are the hot spots for crime, 911 emergency calls, or fires?
- Where do we find a higher than expected proportion of traffic accidents in a city?

With Regression Analyses, you ask **WHY** something is happening.

- Why are there places in the United States where people persistently die young? What might be causing this?
- Can we model the characteristics of places that experience a lot of crime, 911 calls, or fire events to help reduce these incidents?
- What are the factors contributing to higher than expected traffic accidents? Are there policy implications or mitigating actions that might reduce traffic accidents across the city and/or in particular high accident areas?

Regression Models

Spatial Regression

- Spatial data often do not fit traditional, non-spatial regression requirements because they are:
  - spatially autocorrelated (features near each other are more similar than those further away)
  - nonstationary (features behave differently based on their location/regional variation)
- No spatial regression method is effective for both characteristics.

Linear Regression

- Used to analyze linear relationships among variables.
- Linear relationships are positive or negative
- Regression analyses attempt to demonstrate the degree to which one or more variables potentially promote positive or negative change in another variable.

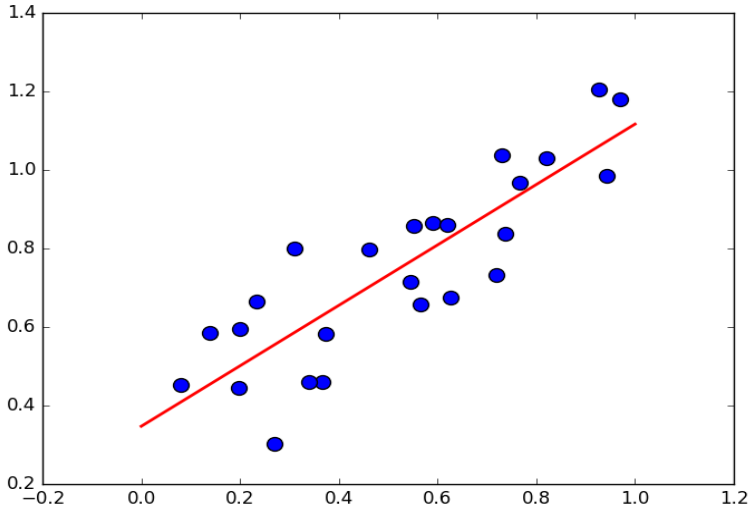


Image by MIT OpenCourseWare.

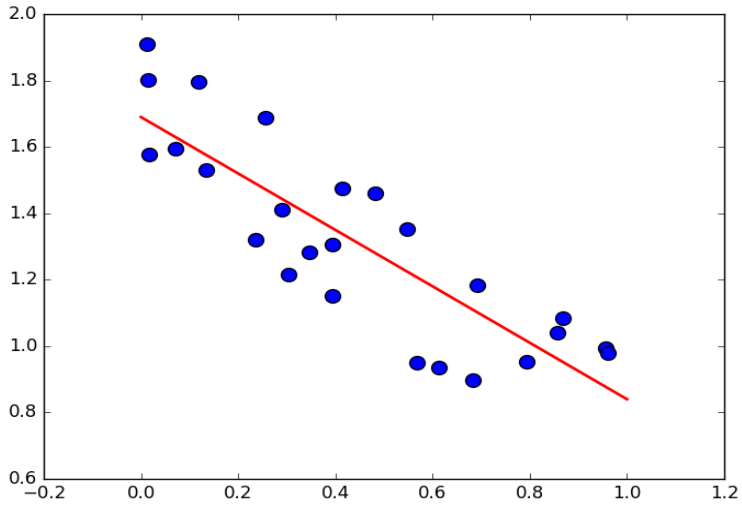


Image by MIT OpenCourseWare.

Linear Regression Equation

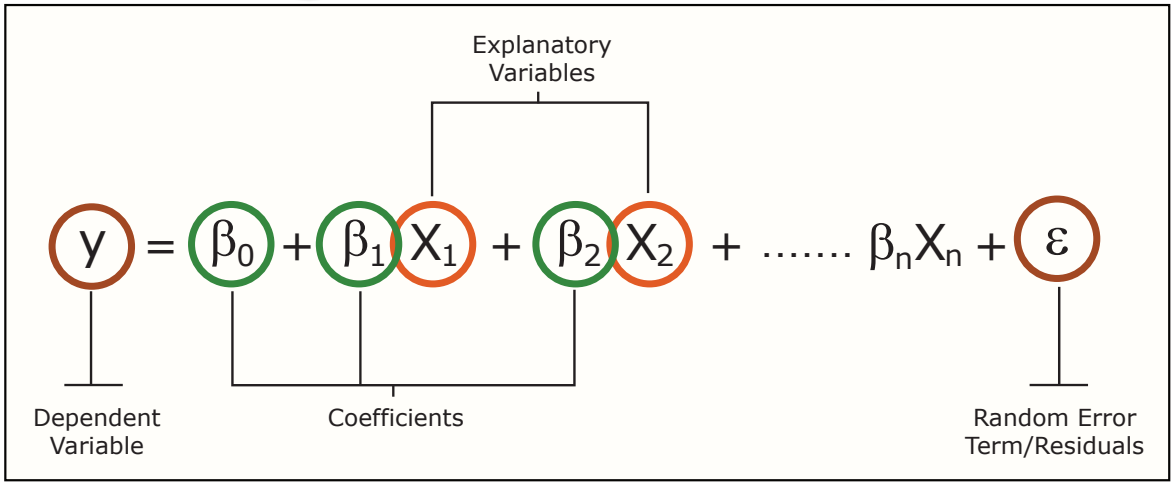


Image by MIT OpenCourseWare.

Y = variable you are trying to predict or understand

X = value of the dependent variables

β = coefficients computed by the regression tool that represent the strength and type of relationship X has to Y

Residuals = the unexplained portion of the dependent variable

- large residuals = a poor model fit

Residuals

Difference between the observed and predicted values

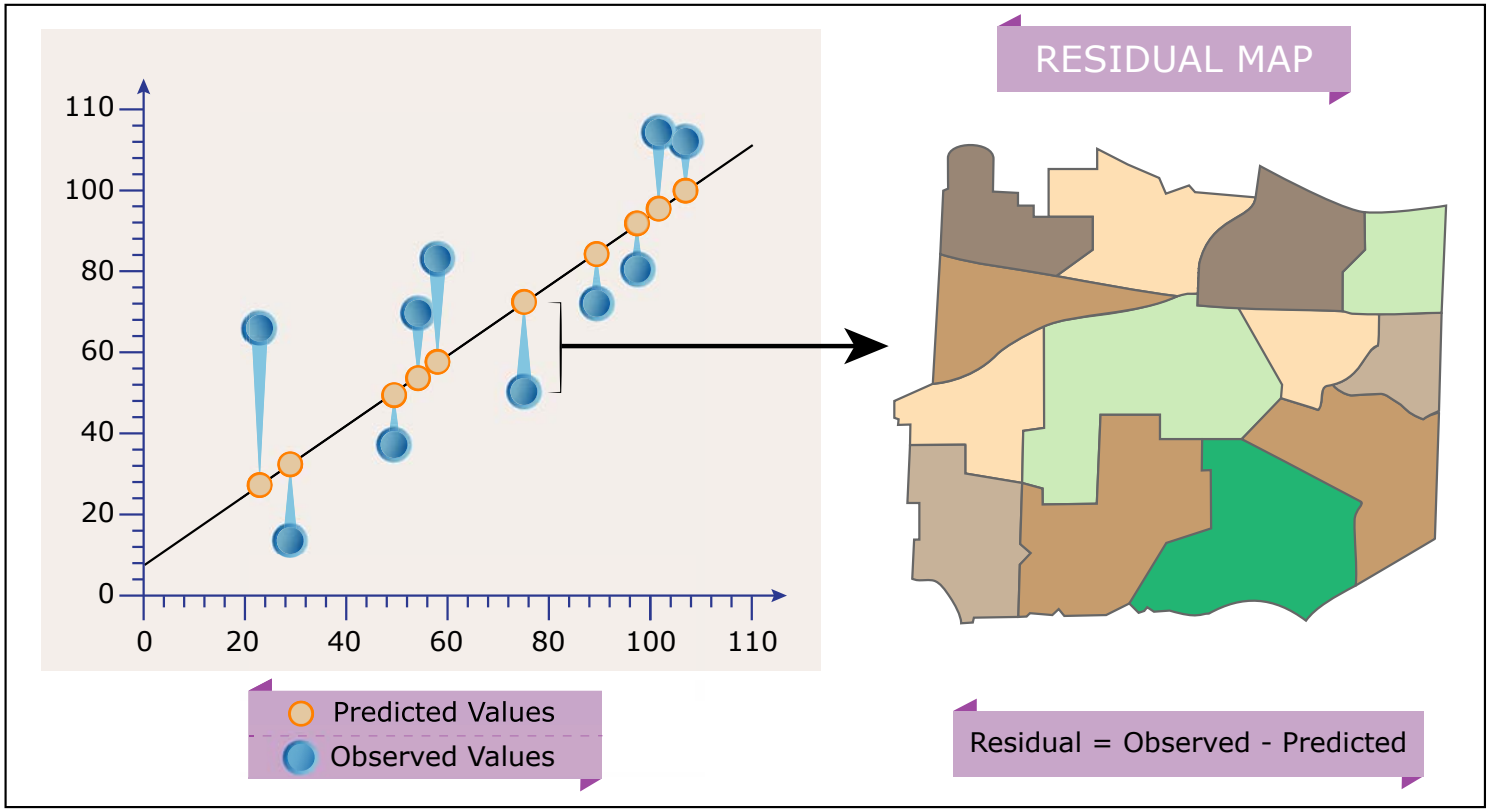


Image by MIT OpenCourseWare.

Ordinary Least Squares Regression (ArcMap and Geoda)

- Best known linear regression technique and a good starting point for all spatial regression analyses.
- As a global model, it provides 1 equation to represent the entire dataset

Geographically Weighted Regression (GWR) (ArcMap)

- Provides a local model of the variable by fitting a regression equation to every feature in the dataset.
- The equations incorporate the dependent and explanatory variables of features falling within the bandwidth of each target feature.

Spatial Lag Model (Geoda)

- Includes a spatially lagged dependent variable: $y=(\rho)Wy + X(\beta) + \varepsilon$
  - Wy = spatially lagged dependent variable for weights matrix W
  - X = matrix of observations on the explanatory variable
  - ε = vector of error terms
  - ρ and β are parameters
- A spatial lag is a variable that averages the neighboring values of a location.
- A weights matrix specifies how features are related to each other (are they neighbors or not?)
- Accounts for autocorrelation in the model with the weights matrix
  - y is dependent on its neighbors (through the weights matrix)

Spatial Error Model (Geoda)

- Assumes the errors of a model are spatially correlated.
- Includes a spatial autoregressive error term:
  $y = X(\beta) + \varepsilon, \varepsilon = \lambda(W)\varepsilon + u$
  - X is a matrix of observations on the explanatory variables
  - W is the spatial weights matrix
  - ε is a vector of spatially autocorrelated error terms
  - $(W)\varepsilon$ is the extent to which the spatial component of the errors are correlated with one another for nearby observations
  - u is a vector of i.i.d. (independent identically distributed) errors
  - λ and β are parameters.
- Similar to spatial lag model: accounts for autocorrelation in the error with the weights matrix.

Interpreting Results

Measure of Regression Fit

R^2

- How well the regression line fits the data
- The proportion of variability in the dataset that is accounted for by the regression equation.
- Ranges from 0 to 1
- Outliers or non-linear data could decrease R^2.

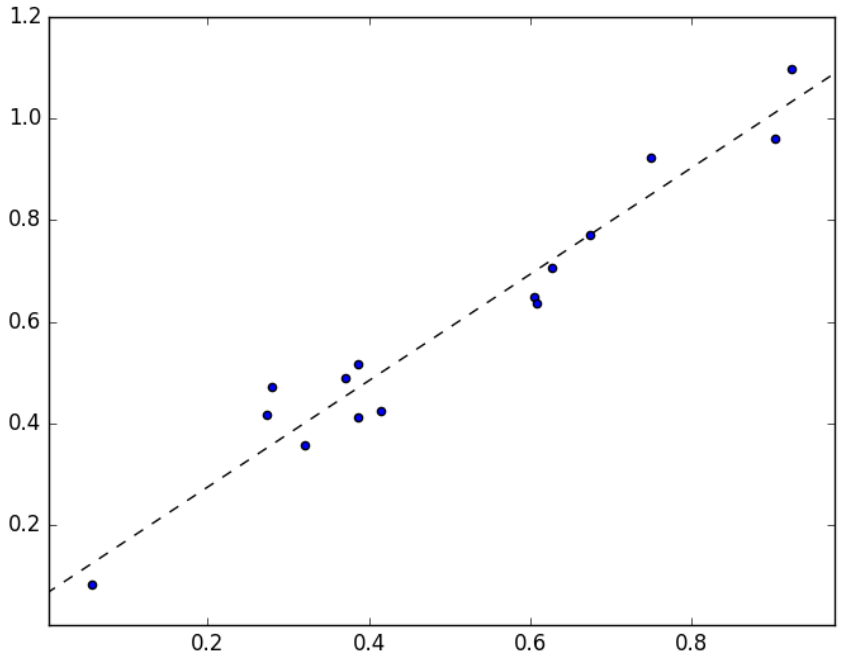


Image by MIT OpenCourseWare.

Variable Coefficients

- The sign shows whether the relationship is positive or negative
- The coefficient shows the strength of the relationship.
- P-values indicate whether the variable is a significant predictor of the independent variable.
- Use the coefficients to form a regression equation: $y = 10 + .5a - 6b + 8c$
- Remove variables with high p-values to see if R^2 increases.

Comparability

- Use Akaike information Criterion (AICc) value when comparing models.
- AICc is a measure of the relative goodness of fit of a statistical model.
- It assists with model selection, but does not test the null hypothesis.
- A lower AICc value means the model is a better fit for the data.

Multicollinearity

- Two or more variables may be highly correlated with one another
- Variance Inflation Factor (VIF)
  - Larger than 7.5 could indicate redundancy among variables.
- Multicollinearity Condition Number
  - Values over 30 indicate a problem

Tests for Residuals/Errors

- **Jarque-Bera Test**: Tests the normality of errors. If it is significant, you may be missing an explanatory variable.
- **Breusch-Pagan**, **Koenker-Bassett**, **White**: Test for heteroskedasticity (non-constant variance). If these are significant, the relationships between some or all of the explanatory variables and the dependent variable are non-stationary (a strong predictor in one area, but weak in others). Try other regression models (GWR, etc.)
- **Spatial Autocorrelation**: Autocorrelated residuals could indicate missing variables or the need for alternative regression models.

Plotting Residuals

- Residuals vs. ID (or any unique identifier)
  - Should not display any pattern
  - Examine large residuals and look for systematic relationships to improve upon the model
- Residuals vs. Predicted
  - Detects heteroskedasticity, or unequal variances
  - Funnel-like patterns indicate relationships between the residuals and predicted values

Maps

- Predicted Value Map
  - The value of the dependent variable, based on the regression equation
  - A smoothed map
  - Random variability, due to factors other than those included in the model, have been smoothed out
- Residual Map
  - Indicates systematic over or under prediction in regions, which could be evidence of spatial autocorrelation

Potential Problems and Solutions

Data Outliers

- Create a scatter plot to examine extreme values and correct or remove outliers if possible.
- Run regression with and without outliers to see their effect on the analysis

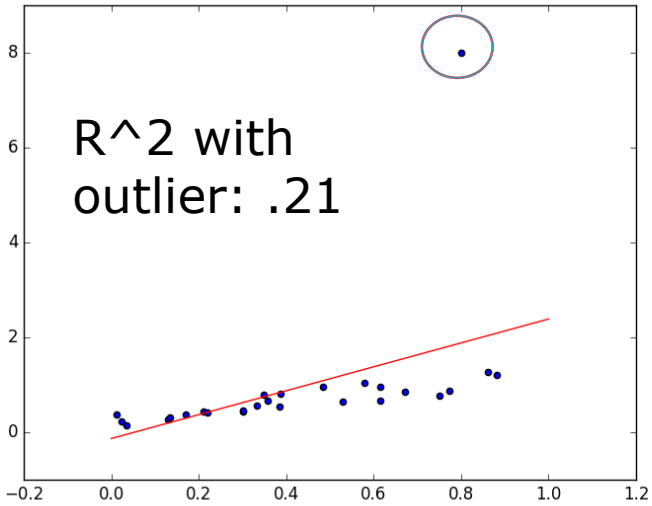


Image by MIT OpenCourseWare.

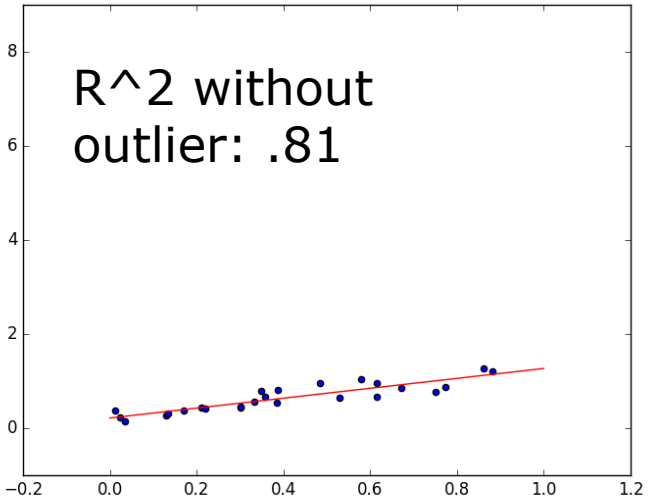


Image by MIT OpenCourseWare.

Nonlinear Relationships

- Create a scatter plot matrix graph and transform variables
- Use a non-linear regression model

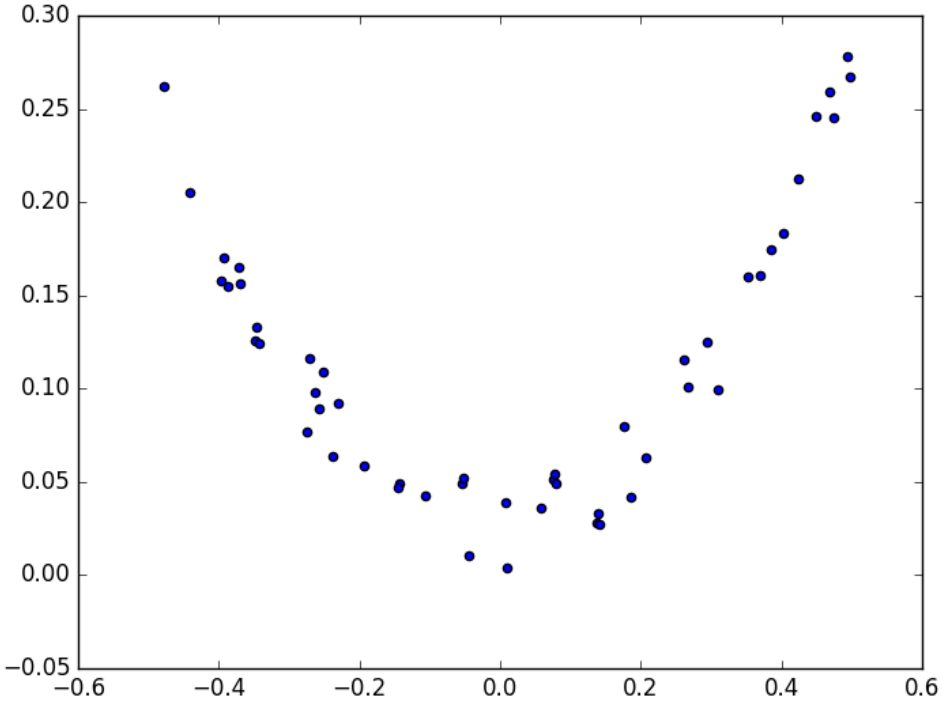
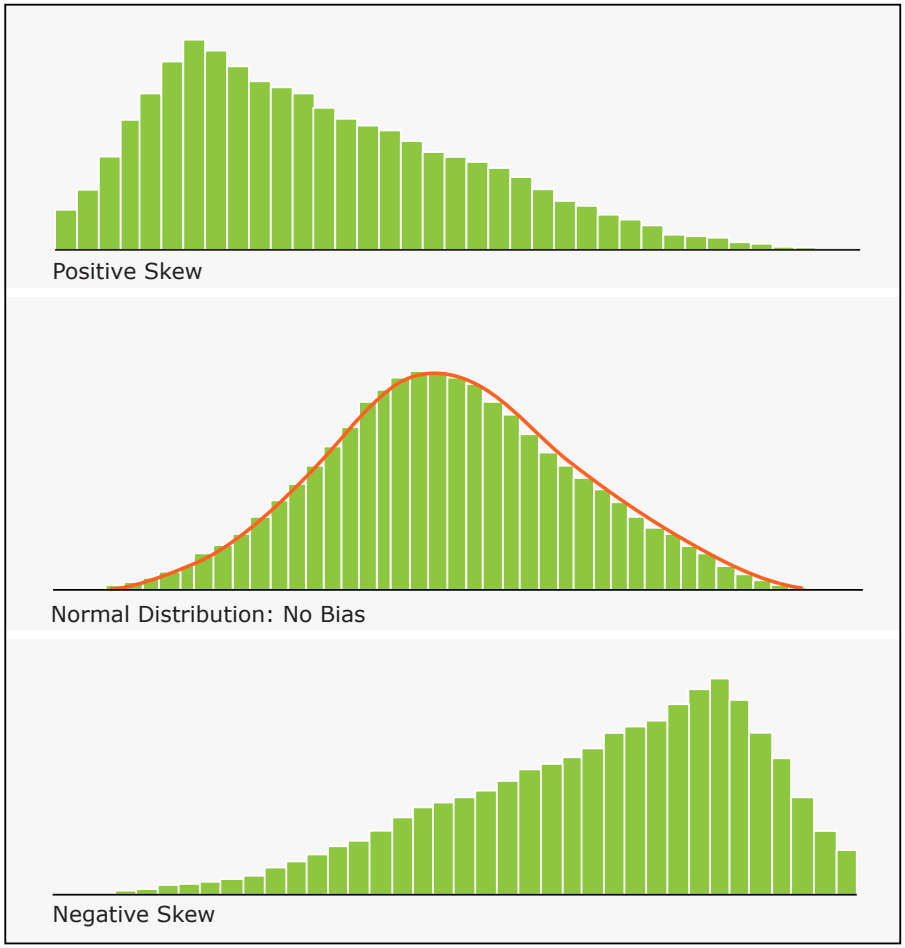


Image by MIT OpenCourseWare.

Normal Distribution Bias



- Jarque-Bera tests whether residuals are normally distributed.
- Model may be misspecified or nonlinear.

Spatially autocorrelated residuals

- Run the spatial autocorrelation tool on the residuals.
- If there is significant clustering, there could be misspecification (a variable is missing from the model).

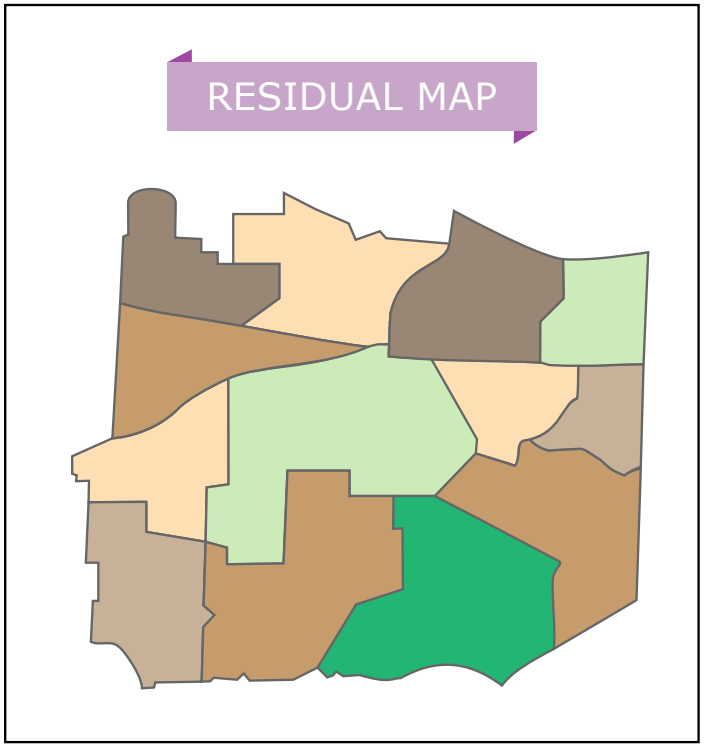
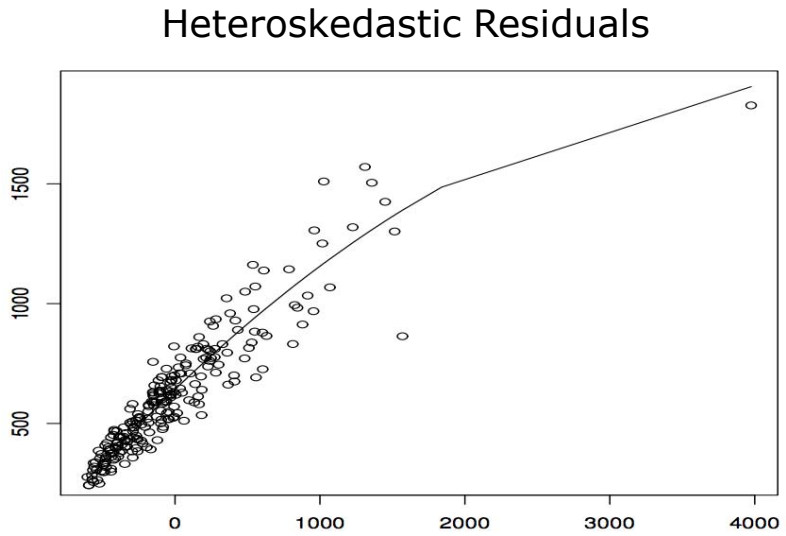


Image by MIT OpenCourseWare.

Heteroskedasticity



- If heteroskedasticity exists, variability differs across sub-populations.
- Variables could be strong predictors in some areas, but weak predictors in others.
- Run a GWR, modify the study area, or re-define the model

Regression Resources

- ESRI Spatial Statistics Website:

  http://blogs.esri.com/Dev/blogs/geoprocessing/archive/2010/07/13/Spatial-Statistics-Resources.aspx
- Geoda Workbook:

  https://geodacenter.asu.edu/og_tutorials
- ESRI Regression Tool Help:

  https://desktop.arcgis.com/en/desktop/latest/tools/spatial-statistics-toolbox/modeling-spatial-relationships.htm
- Video lecture on Spatial Lag and Error:

  https://geodacenter.asu.edu/spatial-lag-and
- An Introduction to Spatial Regression Models in the Social Sciences:

  http://dces.wisc.edu/wp-content/uploads/sites/30/2013/08/W4_W7_WardGleditsch.pdf

RES.STR-001 Geographic Information System (GIS)Tutorial

January IAP 2016