Regression Analysis Using ArcMap

By Jennie Murack

Regression Basics

How is Regression Different from other Spatial Statistical Analyses?

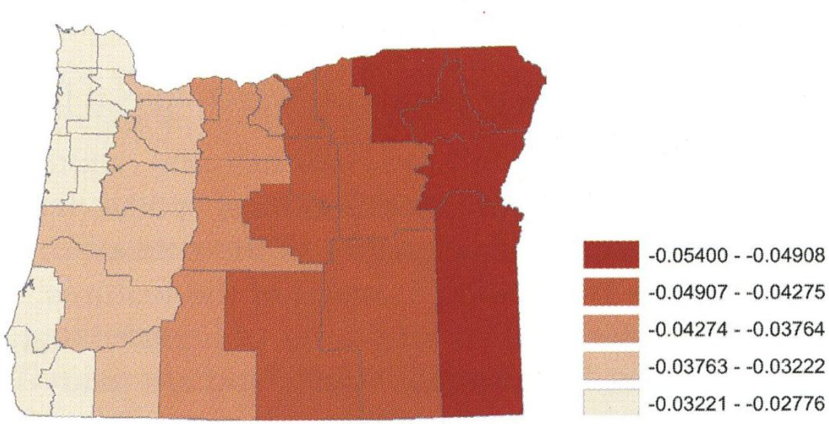
- With other tools you ask **WHERE** something is happening?

  - 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?

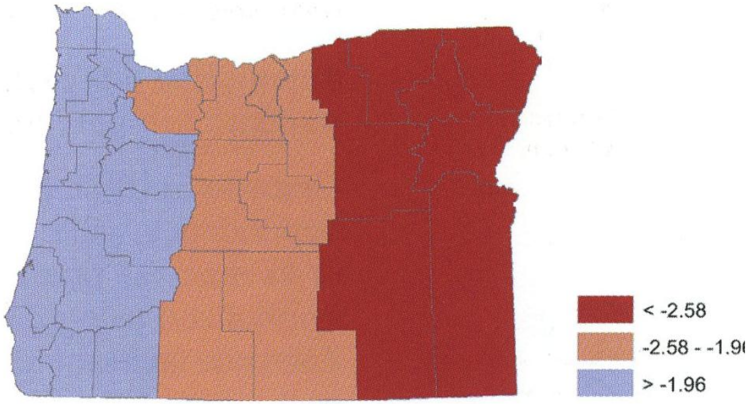
- With Regression Analyses, you ask **WHY** something is happening.
  - Why are there places in the United States where people persistently die young? What might be causing this?
  - Can we model the characteristics of places that experience a lot of crime, 911 calls, or fire events to help reduce these 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 analysis allows you to…

- Model, examine and explore spatial relationships
- Predict



Coefficients for percent rural and low-weight births



T-scores show where this relationship is significant

Reasons to Use Regression Analysis

- To model phenomenon to better understand it and possibly make decisions
- To model phenomenon to predict values at other places or times
- To explore hypotheses

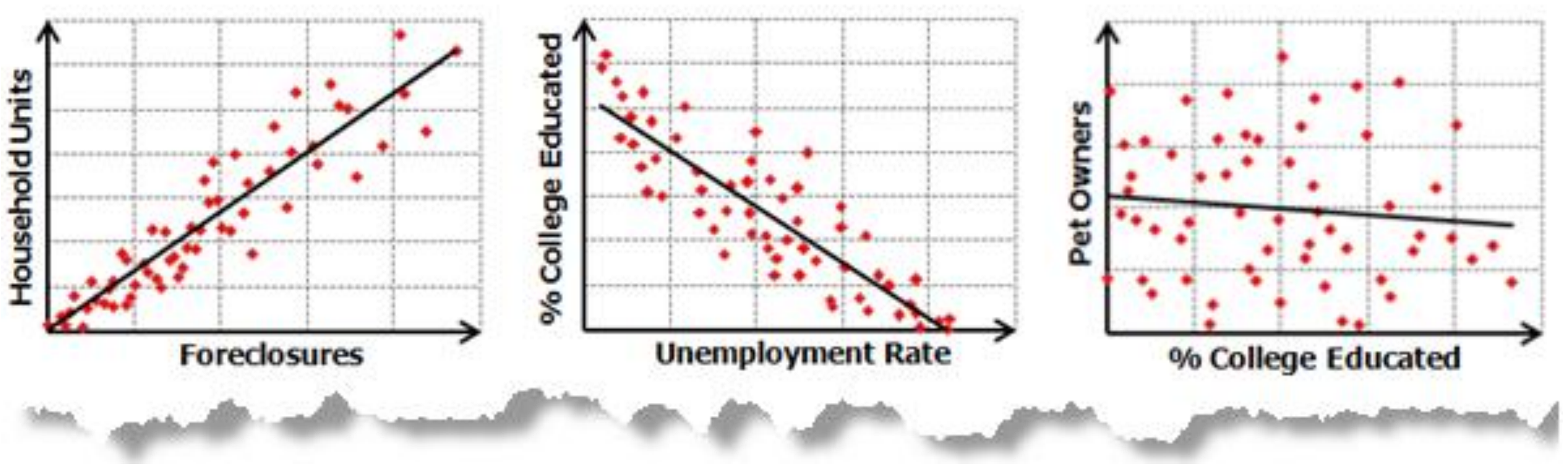
Types of Regression

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.

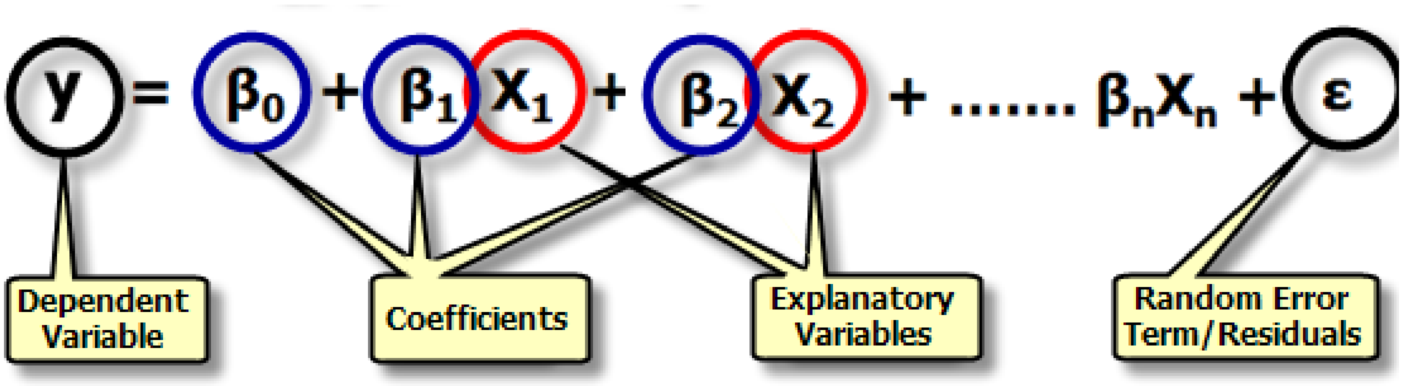


Linear Regression Techniques

- Ordinary Least Squares (OLS) is the best known technique and a good starting point for all spatial regression analyses.
  - Global model = provides 1 equation to represent the entire dataset
- Geographically Weighted Regression (GWR)
  - Local Model = fits a regression equation to every feature in the dataset
  - Regional variation incorporated into the regression model

The Equation

Regression Equation



$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$
Dependent Variable
Coefficients
Explanatory Variables
Random Error Term/Residuals

y = process you are trying to predict or understand

X = used to model or predict the dependent variable

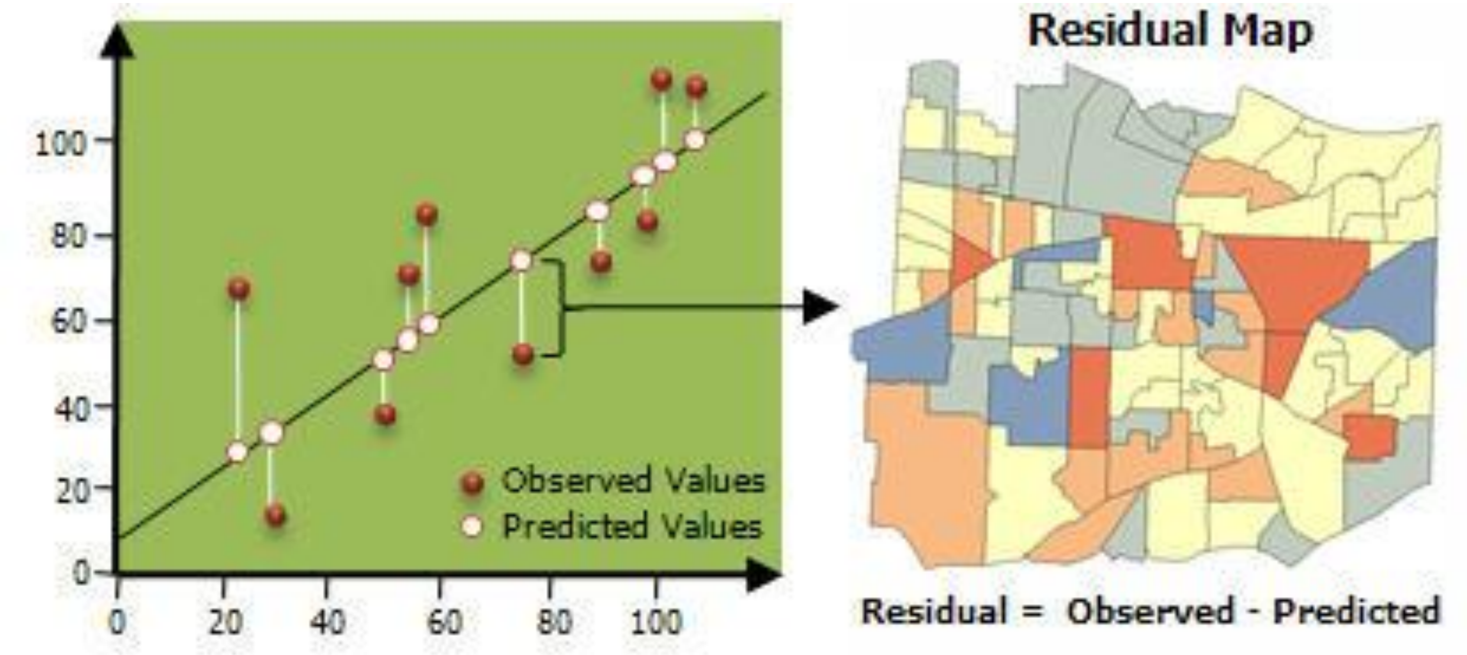
B = coefficients computed by the regression tool, represent the strength and type of relationship X has to Y

Regression Equation

- p-values = result of a statistical test
  - low p-values suggest that the coefficient is important to your model
- R^2 = statistics derived from the regression equation to quantity the performance of the model
  - The closer r^2 is to 1, the more dependence there is among variables.
- Residuals = the unexplained portion of the dependent variable
  - large residuals = a poor model fit

Residuals

Difference between the observed and predicted values



Potential Regression Problems

Omitted explanatory variables (misspecification)

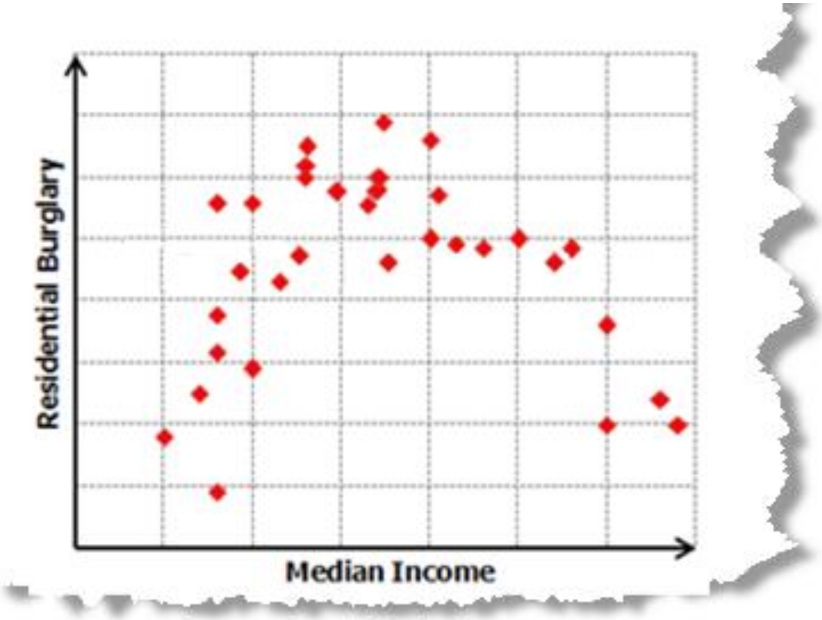
Solution:

- Map and examine OLS residuals and GWR coefficients
- Run Hot Spot Analysis on OLS residuals

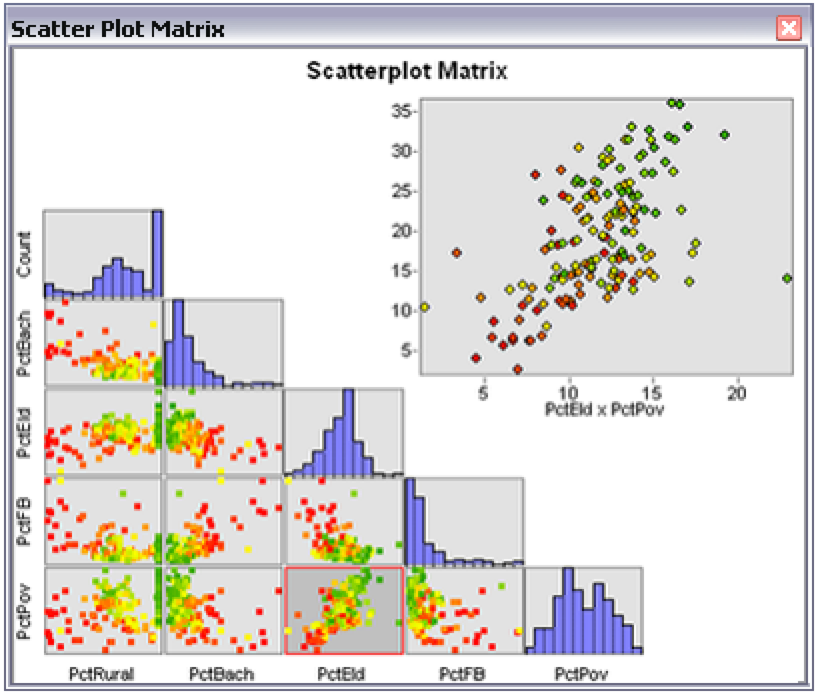
Nonlinear Relationships

Solutions:

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



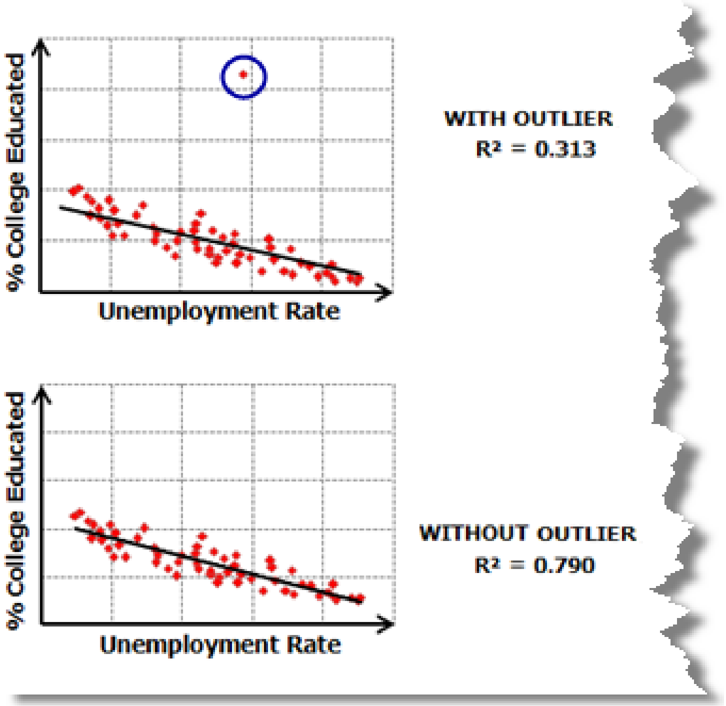
Data Outliers



- Create a scatter plot to examine extreme values and correct or remove outliers if possible.

Solutions:

- Run regression with and without outliers to see their effect on the analysis



Nonstationarity

- Definition: *The relationship among the data changes based on location.*

Solutions:

- OLS automatically tests for problems with nonstationarity.
- GWR may be a more appropriate analysis.

Multicollinearity

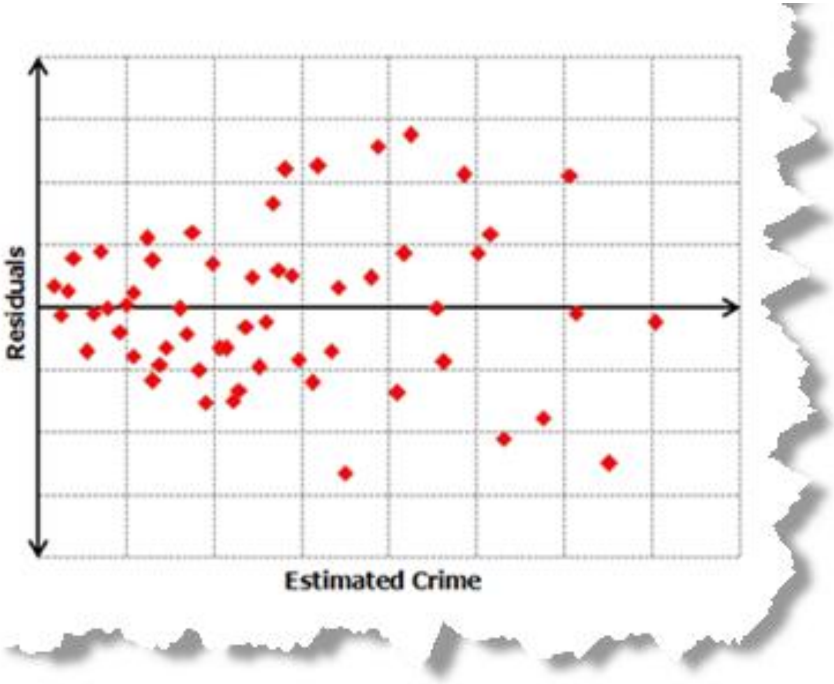
- Definition: *One or a combination of explanatory variables is redundant.*

Solutions:

- OLS tool automatically checks for this.
- Remove or modify the variable(s).

Inconsistent variance in residuals

- Definition: *Model may predict well for small values of the dependent variable, but become unreliable for large values.*



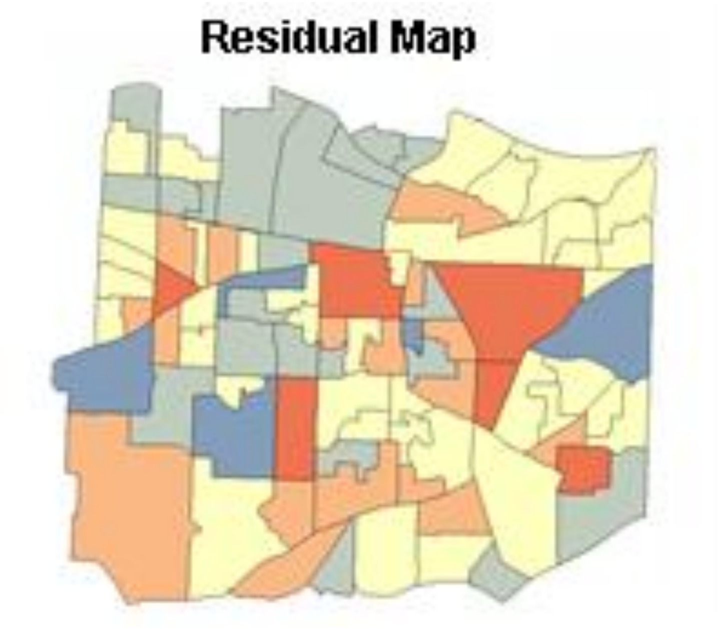
Solutions:

- OLS tests for inconsistent residuals.
- Consult the robust probabilities from the output.

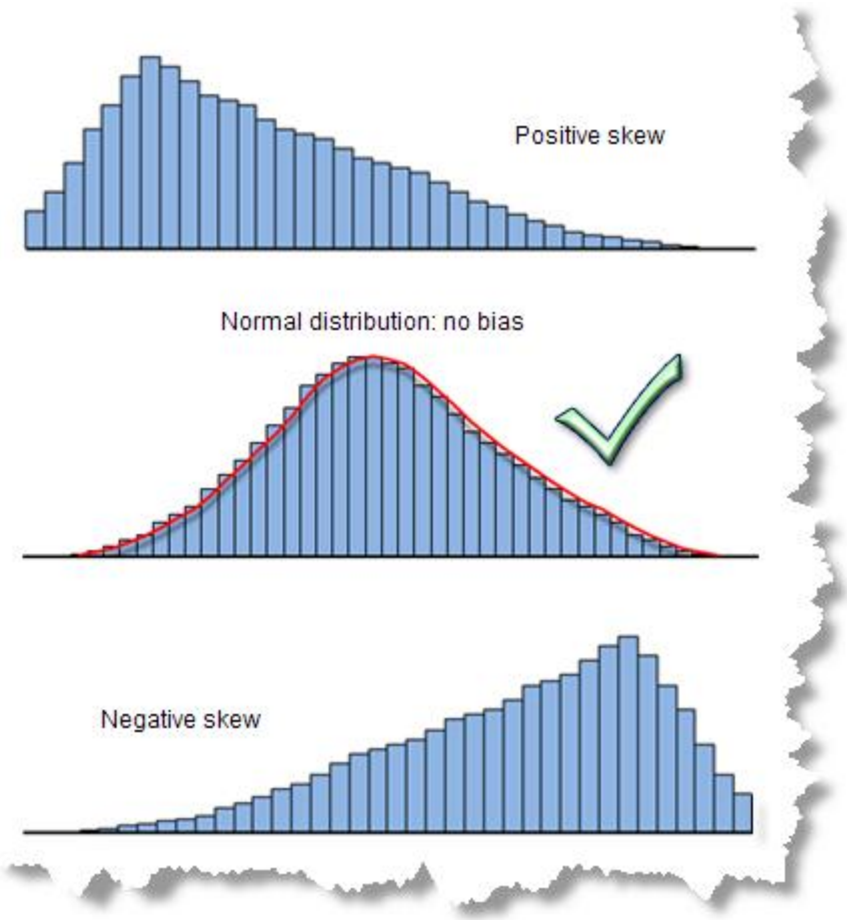
Spatially autocorrelated residuals

Solutions:

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



Normal Distribution Bias



Solutions:

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

Steps of Regression

- Determine what you are trying to predict or examine (dependent variable)
- Identify key explanatory variables
- Examine the distribution to determine the type of regression to conduct
- Run the regression
- Examine the coefficients
- Examine the residuals
  - The mean should equal 0.
  - Overestimates and underestimates should create a random pattern.
  - They should create a normal distribution.
  - Problems could indicate missing variables.
- Remove or add variables and repeat regression