

Answering **Why** Questions

An introduction to using regression analysis with spatial data

By Lauren Scott, *ESRI Geoprocessing Spatial Statistics Product Engineer*, and
Monica Pratt, *ArcUser Editor*

Regression analysis allows you to model, examine, and explore spatial relationships and can help explain the factors behind observed spatial patterns. Regression analysis is also used for prediction. Tools included in the Modeling Spatial Relationships toolset, found in ArcToolbox, help answer *why* questions such as

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

You may want to understand why people are persistently dying young in certain regions, for example, or predict rainfall where there are no rain gauges. The tools in this toolset include Ordinary Least Squares (OLS) Regression and Geographically Weighted Regression (GWR).

OLS, the best known of all regression techniques, is the proper starting point for all spatial regression analyses. It provides a global model of the variable or process you are trying

to understand or predict (early death/rainfall) and creates a single regression equation to represent that process.

GWR is one of several spatial regression techniques increasingly used in geography and other disciplines. GWR provides a local model of the variable or process you are trying to understand/predict by fitting a regression equation to every feature in the dataset. When used properly, these methods are powerful and reliable statistics for examining/estimating linear relationships.

Linear relationships are either positive or negative. If you find that the number of search and rescue events increases when daytime temperatures rise, the relationship is said to be positive; there is a positive correlation. Another way to express this positive relationship is to say that search and rescue events decrease as daytime temperatures decrease.

Conversely, if you find that the number of crimes goes down as the number of police officers patrolling an area goes up, the relationship is said to be negative. You can also express this negative relationship by stating that the number of crimes increases as the number of patrolling officers decreases. The illustration at the top of the next page depicts both positive and negative relationships as well as the case where there is no relationship between two variables.

Correlation analyses and their associated graphics, depicted in this illustration, test the strength of the relationship between two variables. Regression analyses, on the other hand, make a stronger claim. These analyses attempt to demonstrate the degree to which one or more variables potentially promote positive or negative change in another variable.

Using Regression Analysis

Regression analysis can be used for many types of applications such as modeling fire frequency to determine high-risk areas and better understand the factors that contribute to high-risk areas. It can be used to model property loss from fire as a function of variables such as degree of fire department involvement, response time, or property value. If you find that response time is the key factor, you may need to build more fire stations. If you find that involvement is the key factor, you may need to increase equipment/officers dispatched.

Regression analysis can help you better understand phenomena to make better decisions, predict values for phenomena at other locations or times, and test hypotheses.

Modeling a phenomenon can yield a better understanding that can affect policy or provide input for deciding which actions are most appropriate. The basic objective is to measure

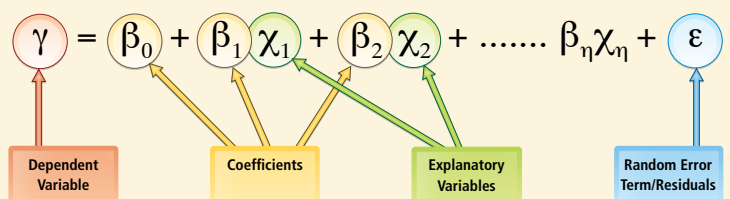
Regression Analysis Components

Terms and basic concepts

It is impossible to discuss regression analysis without first becoming familiar with a few of the terms and basic concepts specific to regression statistics.

Regression equation is the mathematical formula is applied to the explanatory variables to best predict the dependent variable you are trying to model. Although those in the geosciences think of X and Y as coordinates, the notation in regression equations uses X and y. The dependent variable is always y, and independent or explanatory variables are always X. Each independent variable is associated with a regression coefficient describing the strength and the sign of that variable's relationship to the dependent variable. A regression equation might look like the accompanying illustration where y is the dependent variable, the Xs are the explanatory variables, and the β s are regression coefficients.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n + \epsilon$$



Elements of an OLS regression equation

Suppose you want to both model and predict residential burglary (RES_BURG) for the census tracts in your community. You've identified median income (MED_INC), the number of vandalism incidents (VAND), and the number of household units (HH_UNITS) to be key explanatory variables.

The equation would have the elements shown here:

$$\text{RES_BURG} = \beta_0 + \beta_1 * (\text{MED_INC}) + \beta_2 * (\text{VAND}) + \beta_3 * (\text{HH_UNITS}) + \epsilon$$



These graphs show relationships of two variables when that relationship is positive and negative as well as when no relationship is indicated.

the extent that changes in one or more variables jointly affect changes in another. For example, understanding the key characteristics of the habitat of an endangered species of bird (e.g., precipitation, food sources, vegetation, or predators) may help design legislation that will more effectively protect that species.

Modeling a phenomenon to predict values for that phenomenon at other places or other times is another valid and valuable use of regression analysis. In this case, the basic objective is to build a prediction model that is consistent and accurate. You could use this model to predict rainfall in places (such as peaks or valleys)

where there are no gauges, based on a set of variables that explain observed precipitation values. Regression may be used in cases where interpolation is not effective because of insufficient sampling.

You can also use regression analysis to test hypotheses. Suppose you are modeling residential crime to better understand it and implement policy to prevent it. At the outset, you probably have questions or hypotheses you want to test. Some of these might include

- Will there be a positive relationship between vandalism incidents and residential burglary (i.e., Broken Window Theory)?

- Is there a relationship between illegal drug use and burglary? Might drug addicts steal to support their habits?
- Are burglars predatory? Might there be more incidents in residential neighborhoods with higher proportions of elderly or female-headed households?
- Are people at greater risk for burglary if they live in a rich or a poor neighborhood?

Using regression analysis, you can explore these relationships to answer these questions.

Building a regression model is an iterative process that involves finding effective independent variables to explain the process

Continued on page 42

Dependent variable (y) is the variable representing the process you are trying to predict or understand (e.g., residential burglary, foreclosure, rainfall). In the regression equation, it appears on the left side of the equal sign. While you can use regression to predict the dependent variable, you always start with a set of known y values and use these to build (or to calibrate) the regression model. The known y values are often referred to as *observed values*.

Independent/Explanatory variables (X) are used to model or predict the dependent variable values. In the regression equation, they appear on the right side of the equal sign. We say that the dependent variable is a function of the independent (or explanatory) variables. If you are interested in predicting annual purchases for a proposed store, you might include in your model explanatory variables representing the number of potential customers, distance to competition, store visibility, and local spending patterns.

Regression coefficients (β) are computed by the regression tool. They are values, one for each explanatory variable, that represent the strength and type of relationship the explanatory variable has to the dependent variable. Suppose you are modeling fire frequency as a function of solar radiation, vegetation, precipitation, and aspect. You might expect a positive relationship between fire frequency and solar radiation (i.e., the more sun, the more frequent is occurrence of fire incidents). When the relationship is positive, the sign for the associated coefficient is also positive. You might expect a negative relationship between fire frequency and precipitation (places with more rain have fewer fires). Coefficients for negative relationships have negative signs. When the relationship is a strong one, the coefficient is large. Weak relationships are associated with coefficients near zero. β_0 is the regression *intercept*. It represents the expected value for the dependent variable if all of the independent variables are zero.

Continued on page 42

Answering *Why* Questions

Continued from page 41

you are trying to model/understand, running the regression tool to determine which variables are effective predictors, then removing/adding variables until you find the best model possible. The accompanying article, “Regression Analysis Components: An introduction to terms and basic concepts,” defines the terms used when discussing this type of analysis.

Regression Analysis Issues

OLS regression is a straightforward method that has both well-developed theory behind it and effective diagnostics to assist with interpretation and troubleshooting. OLS is only effective and reliable, however, if your data and regression model meet/satisfy all the assumptions inherently required by this method. Be sure to visit the ArcGIS Resource Center for ArcGIS Desktop (resources.esri.com/arcgisdesktop) and read “How Regression Models Go Bad.” This article supplies an excellent discussion of this topic with examples.

Spatial data often violates the assumptions/requirements of OLS regression, so it is important to use regression tools in conjunction with appropriate diagnostic tools that can assess whether or not regression is an appropriate method for your analysis, the structure of the data, and the model being implemented.

Spatial Regression

Spatial data exhibits two properties that make it difficult (but not impossible) to meet the assumptions and requirements of traditional

(nonspatial) statistical methods such as OLS regression.

First, geographic features are more often than not spatially autocorrelated. This means that features near each other tend to be more similar than features that are farther away. This creates an overcount type of bias when using traditional (nonspatial) regression methods.

Second, geography *is* important. Often, the processes most important to the model are nonstationary; these processes behave differently in different parts of the study area. This characteristic of spatial data can be referred to as regional variation or spatial drift.

True spatial regression methods were developed to robustly deal with these two characteristics of spatial data and even incorporate the special qualities of spatial data to better model data relationships. Some spatial regression methods deal effectively with spatial autocorrelation, while others accommodate nonstationarity processes well. At present, no spatial regression methods are effective for both characteristics. However, for a properly specified GWR model, spatial autocorrelation is typically not a problem.

There seems to be a big difference between how a traditional statistician and a spatial statistician view spatial autocorrelation. The traditional statistician sees spatial autocorrelation as a bad thing that needs to be removed from the data (through resampling, for example) because it violates the underlying assumptions of many traditional (nonspatial) statistical methods.

For the geographer or GIS analyst, spatial autocorrelation is evidence of important underlying spatial processes at work. It is an integral component of the data. Removing space removes data from the spatial context—it is like getting only half the story. The spatial processes and spatial relationships evident in the data are a primary interest and are one of the reasons geographers get so excited about spatial data analysis.

However, to avoid an overcounting type of bias in your model, you must identify the full set of explanatory variables that will effectively capture the inherent spatial structure in your dependent variable. If you cannot identify all these variables, you will very likely see statistically significant spatial autocorrelation in the model residuals. Unfortunately, you cannot trust your regression results until this is remedied. Use the Spatial Autocorrelation tool in the Spatial Statistics toolbox to test for statistically significant spatial autocorrelation in your regression residuals.

There are at least three strategies for dealing with spatial autocorrelation in regression model residuals: resampling input variables, isolating spatial and nonspatial components, and incorporating spatial autocorrelation into the regression model.

■ Resample until the input variables no longer exhibit statistically significant spatial autocorrelation. While this does not ensure the analysis is free of spatial autocorrelation problems, these problems are far less likely

Regression Analysis Components

Continued from page 41

P-values are generated by a statistical test that is performed by most regression methods to compute a probability for the coefficients associated with each independent variable. The null hypothesis for this statistical test states that a coefficient is not significantly different from zero (in other words, for all intents and purposes, the coefficient is zero and the associated explanatory variable is not helping your model). Small p-values reflect small probabilities and suggest that the coefficient is, indeed, important to your model with a value that is significantly different from zero (the coefficient is *not* zero). For example, you would say that a coefficient with a p-value of 0.01 is statistically significant at the 99 percent confidence level; the associated variable is an effective predictor. Variables with coefficients near zero do not help predict or model the dependent variable; they are almost always removed from the regression equation unless there are strong theoretical reasons to keep them.

R² / R-Squared values, which range from 0 to 100 percent, are a measure of model performance. Multiple R-squared and Adjusted R-Squared are both statistics derived from the regression equation to quantify model performance. If your model fits the observed dependent variable values perfectly, R-Squared is 1.0 and you (no doubt) have made an error. Perhaps you've used a form of *y* to predict *y*. More likely, you will see R-Squared values such as 0.49, which you can interpret in the following manner: this model explains 49 percent of the variation in the dependent variable. To understand what the R-squared value is indicating, create a bar graph showing both the estimated and observed *y* values sorted by the estimated values. Notice how much overlap there is. This graphic provides a visual representation of how well the model's predicted values explain the variation in the observed dependent variable values. The Adjusted R-Squared value is always a bit lower than the Multiple R-Squared value because it reflects model complexity (the number of variables) as it relates to the data.

to occur when spatial autocorrelation is removed from the dependent and explanatory variables. This is the approach of a traditional statistician for dealing with spatial autocorrelation. It is only appropriate if spatial autocorrelation is the result of data redundancy (i.e., the sampling scheme is too fine).

- Isolate the spatial and nonspatial components of each input variable using a spatial filtering regression method. Space is removed from each variable, but then it is put back into the regression model as a new variable to account for spatial effects/spatial structure. Spatial filtering regression methods will be added to ArcGIS in a future release.
- Incorporate spatial autocorrelation into the regression model using spatial econometric regression methods. Econometric spatial regression methods will be added to ArcGIS in a future release.

Regional Variation

Global models, such as OLS regression, create equations that best describe the overall data relationships in a study area. When those relationships are consistent across the study area, the OLS regression equation models those relationships well. However, when those relationships behave differently in different parts of the study area, the regression equation produces more of an average of the mix of relationships present. When those relationships represent two extremes, the global average will

not model either extreme well.

When your explanatory variables exhibit nonstationary relationships (i.e., regional variation), global models tend to fall apart unless robust methods are used to compute regression results. Ideally, you will be able to identify a full set of explanatory variables to capture the regional variation inherent in your dependent variable. However, if you cannot identify all these spatial variables, you will again notice statistically significant spatial autocorrelation in your model residuals and/or lower-than-expected R-squared values. (R-squared values are a measure of model performance. Values vary from 0.0 to 1.0, with higher values being preferable.) Unfortunately, you cannot trust your regression results until this is remedied.

There are at least four ways to deal with regional variation in OLS regression models:

- Include a variable in the model that explains the regional variation. If you see that your model is always overpredicting in the north and underpredicting in the south, for example, add a regional variable set to 1 for northern features and 0 for southern features.
- Use methods that incorporate regional variation into the regression model such as GWR.
- Consult robust regression standard errors and probabilities to determine if variable coefficients are statistically significant. In the ArcGIS Desktop Help Online, see the topic "Interpreting OLS regression results." GWR is still the recommended tool.

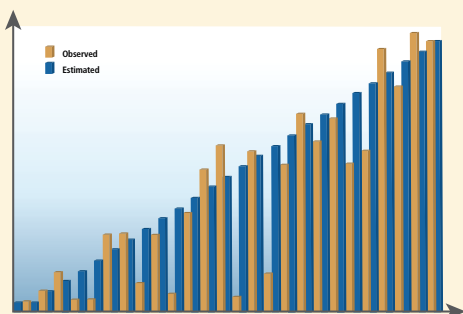
- Redefine/reduce the size of the study area so processes within it are all stationary and no longer exhibit regional variation.

Learning More about Using These Tools

This article provides an introduction to the OLS and GWR tools that were released in ArcGIS Desktop 9.3. ESRI provides many resources for understanding how to intelligently use regression analysis and other spatial statistics tools. A great place to start learning about spatial statistics in general—and these tools specifically—is the ArcGIS Desktop Web help available through the ArcGIS Desktop and Geoprocessing Resource Centers (resources.esri.com) as well as the knowledge bases, communities, and blogs on those sites.

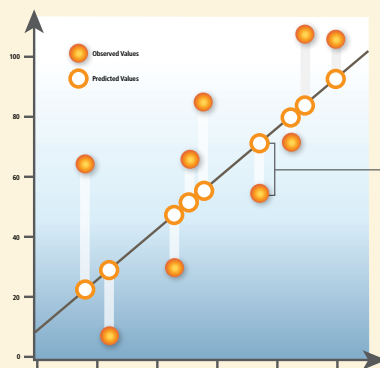
Several training courses offered by ESRI cover spatial statistics. *Understanding Spatial Statistics in ArcGIS 9*, a training seminar, is available at no charge from www.esri.com/training. *Advanced Analysis with ArcGIS*, an instructor-led course, includes an introduction to the Spatial Statistics toolbox, analyzing patterns, and measuring geographic distributions.

A book from ESRI Press, *The ESRI Guide to GIS Analysis, Volume 2: Spatial Measurements and Statistics*, by Andy Mitchell explains how these analyses are performed and used effectively.

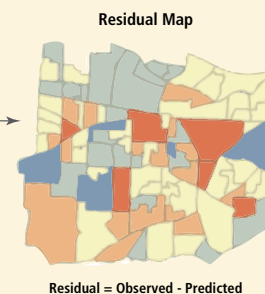


R-Squared is a measure of model performance, summarizing how well the estimated y values match the observed y values.

Residuals are the unexplained portion of the dependent variable, represented in the regression equation as the random error term (ϵ). Known values for the dependent variable are used to build and calibrate the regression model. Using known values for the dependent variable (y) and known values for all of the explanatory variables (the X s), the regression tool constructs an equation that will predict those known y values as well as possible. The predicted values will rarely match the observed values exactly. The differences between the observed y values and the predicted y values are called the residuals. The magnitude of the residuals from a regression equation is one measure of model fit. Large residuals indicate poor model fit.



Regression residuals



Residual = Observed - Predicted