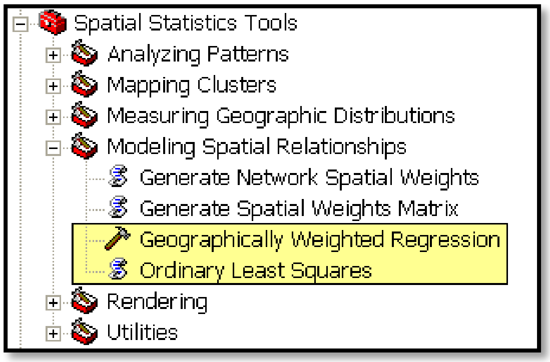
Analyzing 911 response data using Regression

This tutorial demonstrates how regression analysis has been implemented in ArcGIS, and explores some of the special considerations you'll want to think about whenever you use regression with spatial data.

Regression analysis allows you to model, examine, and explore spatial relationships, to better understand the factors behind observed spatial patterns, and to predict outcomes based on that understanding. Ordinary Least Squares regression (OLS) is a global regression method. Geographically Weighted Regression (GWR) is a local, spatial, regression method that allows the relationships you are modeling to vary across the study area. Both of these are located in the Spatial Statistics Tools -> Modeling Spatial Relationships toolset:



- Spatial Statistics Tools
  - Analyzing Patterns
  - Mapping Clusters
  - Measuring Geographic Distributions
  - Modeling Spatial Relationships
    - Generate Network Spatial Weights
    - Generate Spatial Weights Matrix
    - Geographically Weighted Regression
    - Ordinary Least Squares
  - Rendering
  - Utilities

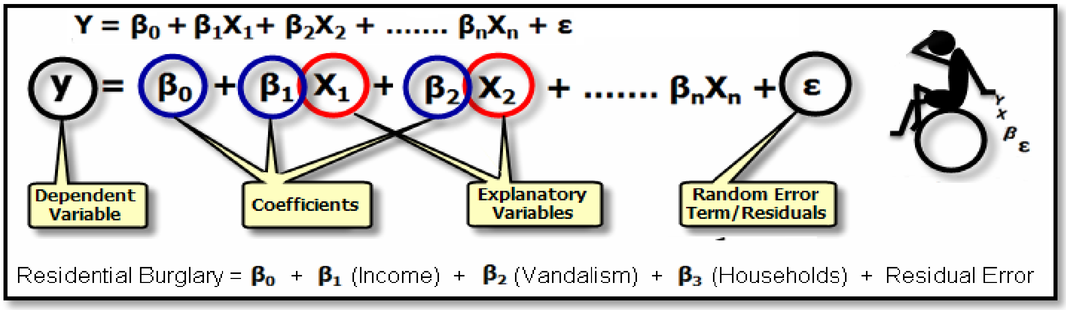
Before executing the tools and examining the results, let's review some terminology:

- **Dependent variable (Y):** what you are trying to model or predict (residential burglary incidents, for example).

- **Explanatory variables (X):** variables you believe influence or help explain the dependent variable (like: income, the number of vandalism incidents, or households).

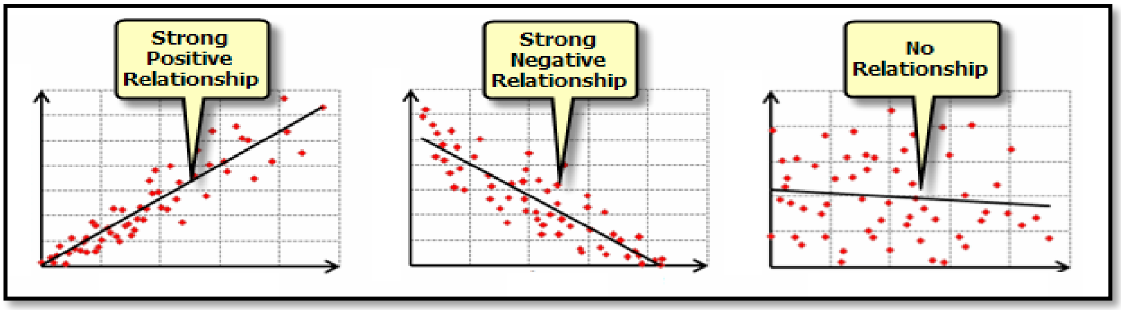
- **Coefficients (β):** values, computed by the regression tool, reflecting the relationship and strength of each explanatory variable to the dependent variable.

- **Residuals (ε):** the portion of the dependent variable that isn't explained by the model; the model under and over predictions.



$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots\dots\; \beta_n X_n + \varepsilon$
$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots\dots\; \beta_n X_n + \varepsilon$
Dependent Variable
Coefficients
Explanatory Variables
Random Error Term/Residuals
Y
x
β
ε
Residential Burglary = β_0 + β_1 (Income) + β_2 (Vandalism) + β_3 (Households) + Residual Error

The sign (+/-) associated with the coefficient (one for each explanatory variable) tells you whether the relationship is positive or negative. If you were modeling residential burglary and obtain a negative coefficient for the Income variable, for example, it would mean that as median incomes in a neighborhood go up, the number of residential burglaries goes down.



Output from regression analysis can be a little overwhelming at first. It includes diagnostics and model performance indicators. All of these numbers should seem much less daunting once you complete the tutorial below.

Important notes:

1. The steps in this tutorial document assume the data is stored at C:\SpatialStats. If a different location is used, substitute "C:\SpatialStats" with the alternate location when entering data and environment paths.
2. This tutorial was developed using ArcGIS 10.0. If you are using a different version of the software, the screenshots and how you access results, may be a bit different.

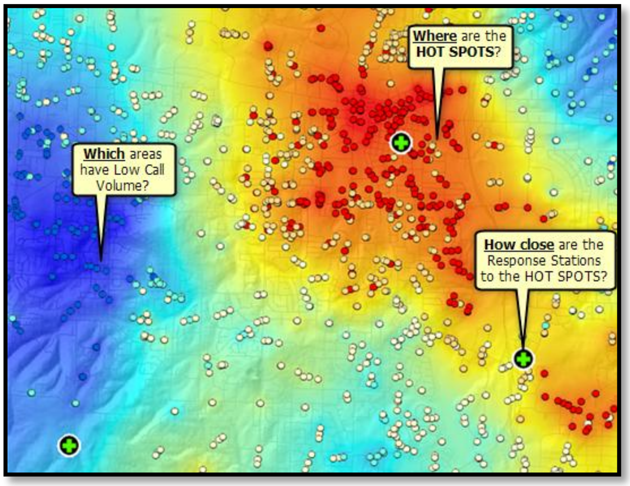
Tutorial

Estimated time: 1.5 hours

Introduction:

In order to demonstrate how the regression tools work, you will be doing an analysis of 911 Emergency call data for a portion of the Portland Oregon metropolitan area.

Suppose we have a community that is spending a large portion of its public resources responding to 911 emergency calls. Projections are telling them that their community's population is going to double in size over the next 10 years. If they can better understand some of the factors contributing to high call volumes now, perhaps they can implement strategies to help reduce 911 calls in the future.


Which areas have Low Call Volume?
Where are the HOT SPOTS?
How close are the Response Stations to the HOT SPOTS?

Step 1 Getting Started

- Open *C:\SpatialStats\RegressionExercise\RegresssionAnalysis911Calls.mxd* (the path may be different on your machine)

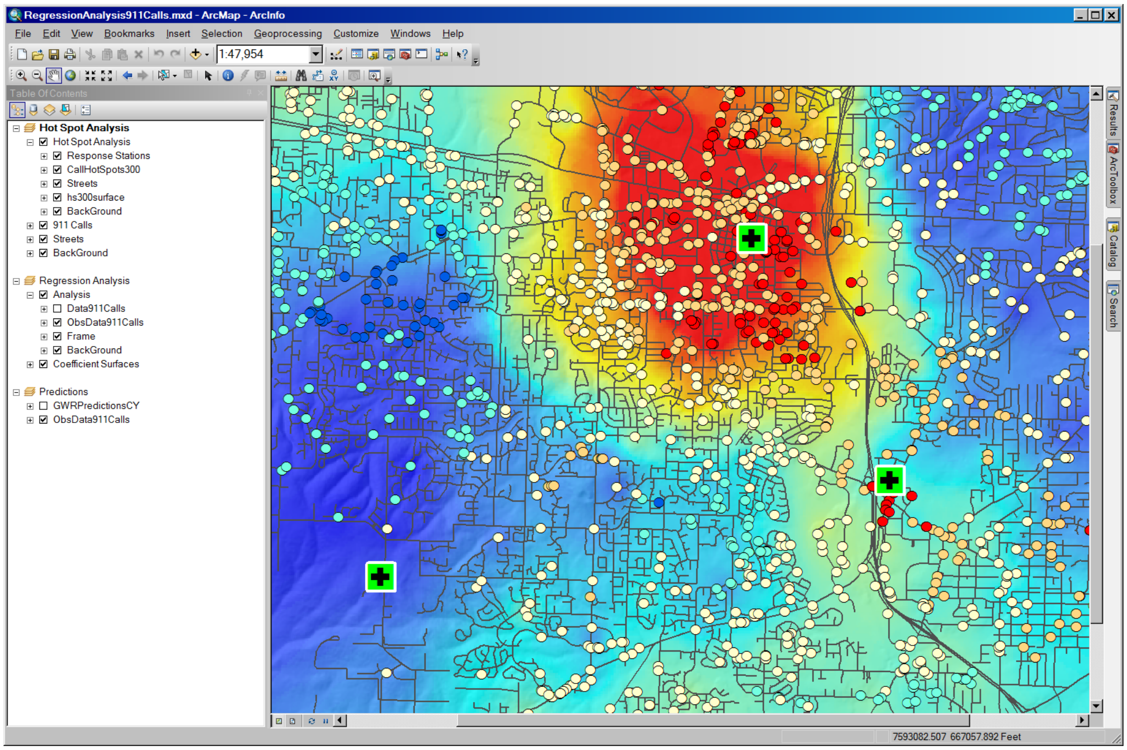
In this map document you will notice several Data frames containing layers of data for the Portland Oregon metropolitan study area.

- Ensure that the Hot Spot Analysis data frame is active

In the map, each point represents a single call into a 911 emergency call center. This is real data representing over 2000 calls.

Step 2 Examine Hotspot Analysis results

- Expand the data frame and click the + sign to the right of the Hot Spot Analysis grouped layer
- Ensure that the Response Stations layer is checked on



Results from running the Hotspot Analysis tool show us where the community is getting lots of 911 calls. We can use these results to assess whether or not the stations (fire/police/emergency medical) are optimally located.

Areas with high call volumes are shown in red (hot spots); areas getting very few calls are shown in blue (cold spots). The green crosses are the existing locations for the police and fire units tasked with responding to these 911 calls.

- Notice that the 2 stations to the right of the map appear to be located right over, or very near, call hot spots. The station in the lower left, however, is actually located over a cold spot; we may want to investigate further if this station is in the best place possible.

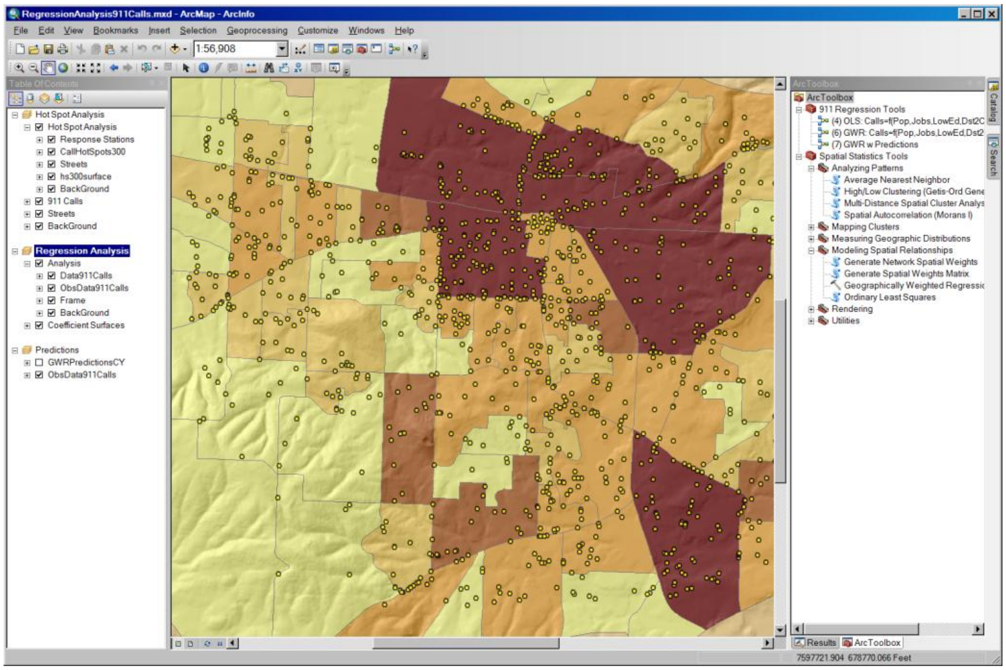
The community can use hot spot analysis to decide if adding new stations or relocating existing stations might improve 911 call response.

Step 3 Exploring OLS Regression

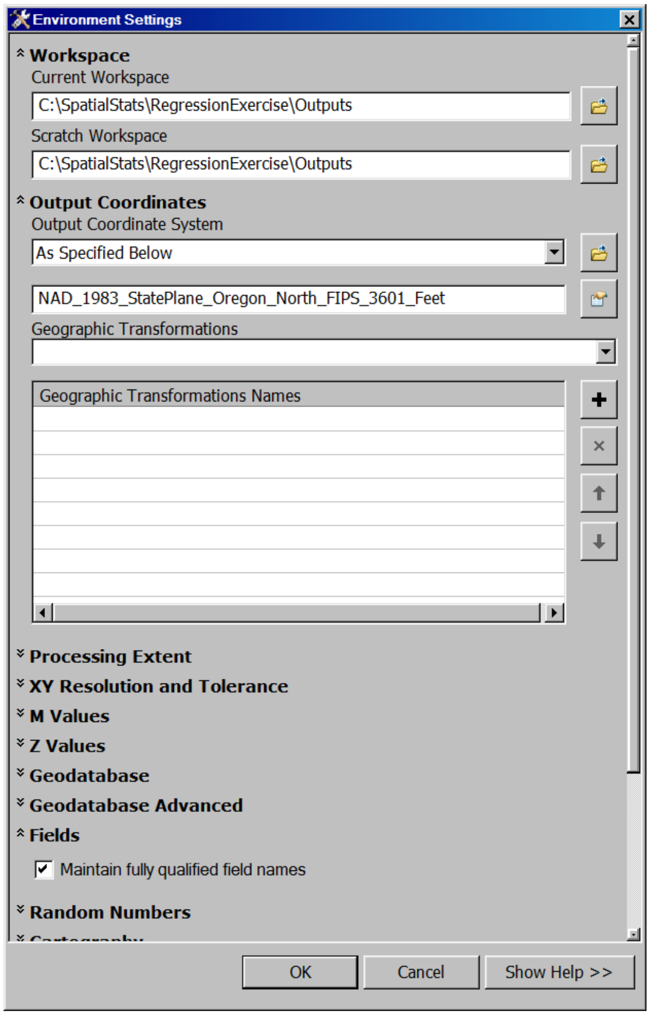
The next question our community is probably asking is, “Why are call volumes so high in those hot spot areas?” and “What are the factors that contribute to high volumes of 911 calls?” To help answer these questions, we’ll use the regression tools in ArcGIS.

- Activate the Regression Analysis data frame by right clicking and choosing Activate

- Expand the **Spatial Statistics tools** toolbox


RegressionAnalysis911Calls.mxd - ArcMap - ArcInfo
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:56,908
Table Of Contents
Hot Spot Analysis
Hot Spot Analysis
Response Stations
CallHotSpots300
Streets
hs300surface
BackGround
911 Calls
Streets
BackGround
Regression Analysis
Analysis
Data911Calls
ObsData911Calls
Frame
BackGround
Coefficient Surfaces
Predictions
GWRPredictionsCY
ObsData911Calls
ArcToolbox
ArcToolbox
911 Regression Tools
(4) OLS: Calls=f(Pop,Jobs,LowEd,Dst2C
(6) GWR: Calls=f(Pop,Jobs,LowEd,Dst2
(7) GWR w Predictions
Spatial Statistics Tools
Analyzing Patterns
Average Nearest Neighbor
High/Low Clustering (Getis-Ord Gene
Multi-Distance Spatial Cluster Analys
Spatial Autocorrelation (Morans I)
Mapping Clusters
Measuring Geographic Distributions
Modeling Spatial Relationships
Generate Network Spatial Weights
Generate Spatial Weights Matrix
Geographically Weighted Regressic
Ordinary Least Squares
Rendering
Utilities
Catalog
Search
Results
ArcToolbox
7597721.904 678770.066 Feet

- Right click in a open space in ArcToolbox and set your environment as follows:



Environment Settings

Workspace

Current Workspace

C:\SpatialStats\RegressionExercise\Outputs

Scratch Workspace

C:\SpatialStats\RegressionExercise\Outputs

Output Coordinates

Output Coordinate System

As Specified Below

NAD_1983_StatePlane_Oregon_North_FIPS_3601_Feet

Geographic Transformations

Geographic Transformations Names

Processing Extent

XY Resolution and Tolerance

M Values

Z Values

Geodatabase

Geodatabase Advanced

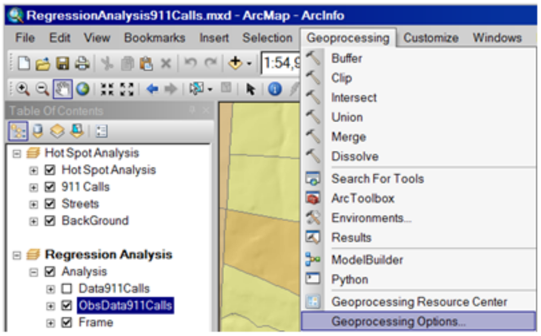
Fields

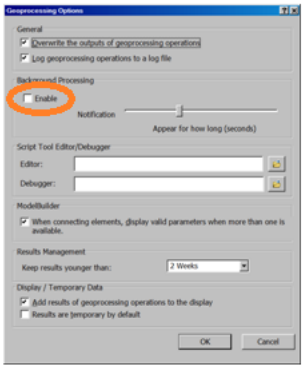
☑ Maintain fully qualified field names

Random Numbers

Cartography

- Disable background processes (Geoprocessing>Geoprocessing Options). With ArcGIS 10, geoprocessing tools can run in the background and all results are available through the Results window. By disabling background processing, we will see tool results in a progress window; this is often best when you are using the Regression tools:




Geoprocessing Options

General
☑ Overwrite the outputs of geoprocessing operations
☑ Log geoprocessing operations to a log file

Background Processing
☐ Enable
Notification
Appear for how long (seconds)

Script Tool Editor/Debugger
Editor:
Debugger:

ModelBuilder
☑ When connecting elements, display valid parameters when more than one is available.

Results Management
Keep results younger than: 2 Weeks

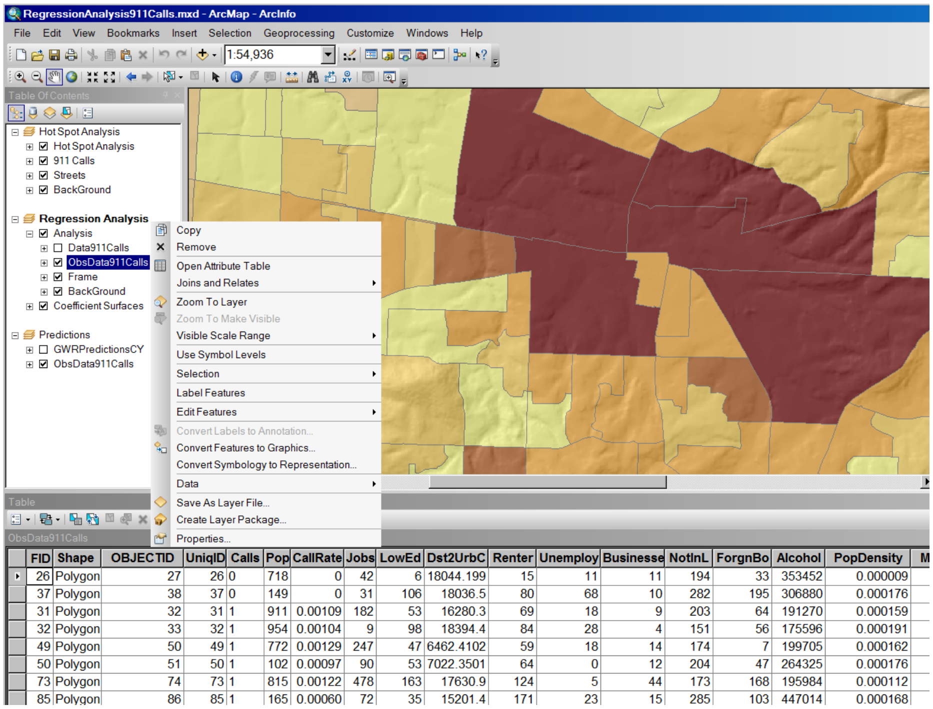
Display / Temporary Data
☑ Add results of geoprocessing operations to the display
☐ Results are temporary by default

OK Cancel

- In the data frame, check off the **Data911Calls** layer

Instead of looking at individual 911 calls as points, we have aggregated the calls to census tracts and now have a count variable (Calls) representing the number of calls in each tract.

- Right click the **ObsData911Calls** layer and choose Open Attribute Table



FID	Shape	OBJECTID	UniqID	Calls	Pop	CallRate	Jobs	LowEd	Dst2UrbC	Renter	Unemploy	Businesse	NotInL	ForgnBo	Alcohol	PopDensity	M
26	Polygon	27	26	0	718	0	42	6	18044.199	15	11	11	194	33	353452	0.000009	
37	Polygon	38	37	0	149	0	31	106	18036.5	80	68	10	282	195	306880	0.000176	
31	Polygon	32	31	1	911	0.00109	182	53	16280.3	69	18	9	203	64	191270	0.000159	
32	Polygon	33	32	1	954	0.00104	9	98	18394.4	84	28	4	151	56	175596	0.000191	
49	Polygon	50	49	1	772	0.00129	247	47	6462.4102	59	18	14	174	7	199705	0.000162	
50	Polygon	51	50	1	102	0.00097	90	53	7022.3501	64	0	12	204	47	264325	0.000176	
73	Polygon	74	73	1	815	0.00122	478	163	17630.9	124	5	44	173	168	195984	0.000112	
85	Polygon	86	85	1	165	0.00060	72	35	15201.4	171	23	15	285	103	447014	0.000168	

The reason we are using census tract level data is because this gives us access to a rich set of variables that might help explain 911 call volumes.

- Notice that the table has fields such as Educational status (LowEd), Unemployment levels (Unemploy), etc.

- When done exploring the fields, close the table

Can you think of anything … any variable… that might help explain the call volume pattern we see in the hot spot map?

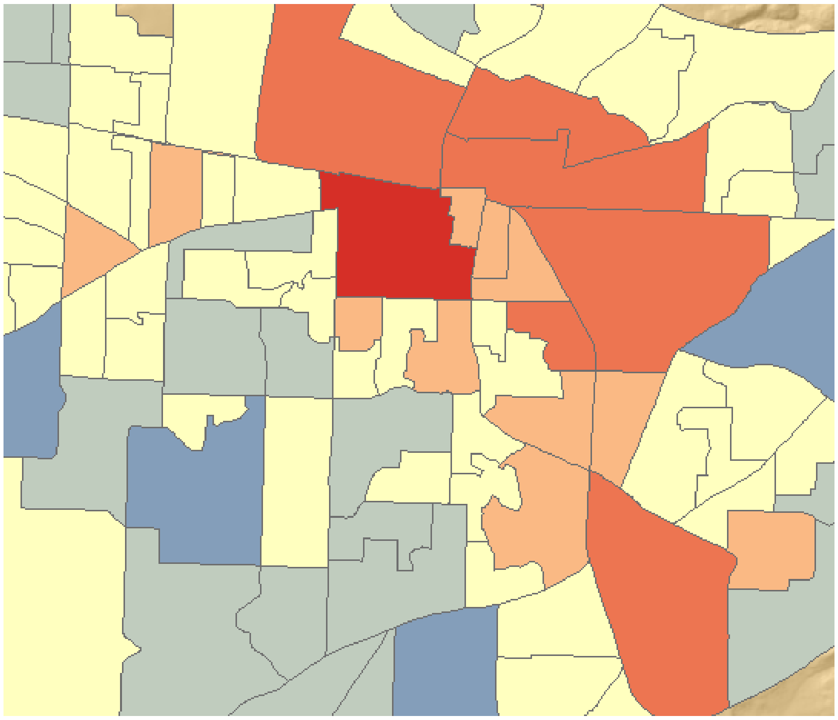
What about population? Would we expect more calls in places with more people? Let's test the hypothesis that call volume is simply a function of population. If it is, our community can use Census population projections to estimate future 911 emergency call volumes.

- Run the OLS tool with the following parameters:

  Note: once the tool starts running, make sure the "close this dialog when completed successfully box" is NOT checked

  - Input Feature Class -> ObsData911Calls
  - Unique ID Field -> UniqID
  - Output Feature Class -> C:\SpatialStats\RegressionExercise\Outputs\OLS911Calls.shp
  - Dependent Variable -> Calls
  - Explanatory Variables -> Pop

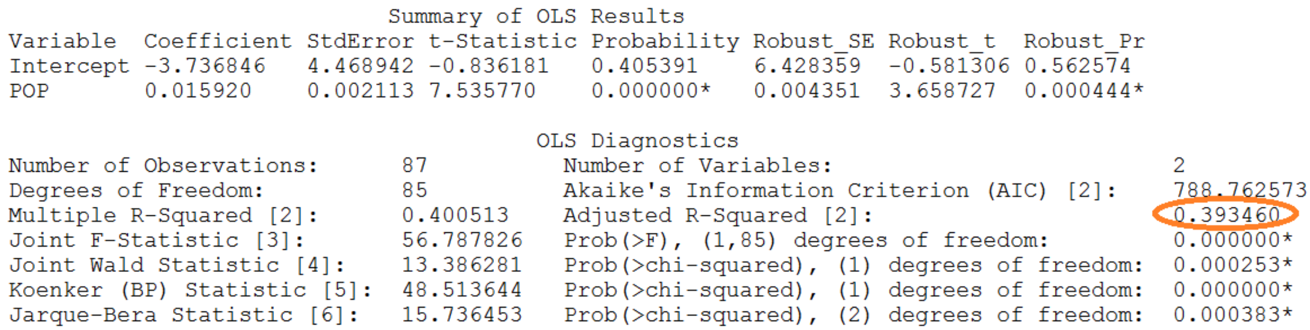
- Move the progress window to the side so you can examine the OLS911calls layer in the TOC.



The OLS default output is a map showing us how well the model performed, using only the population variable to explain 911 call volumes. The red areas are under predictions (where the actual number of calls is higher than the model predicted); the blue areas are over predictions (actual call volumes are lower than predicted). When a model is performing well, the over/under predictions reflect random noise… the model is a little high here, but a little low there… you don't see any structure at all in the over/under predictions. Do the over and under predictions in the output feature class appear to be random noise or do you see clustering? When the over (blue) and under (red) predictions cluster together spatially, you know that your model is missing one or more key explanatory variables.

- The OLS tool also produces a lot of numeric output. Expand and enlarge the progress window so you can read this output more clearly.

- Notice that the Adjusted R-Squared value is 0.393460, or 39%. This indicates that using population alone, the model is explaining 39% of the call volume story.



```
                            Summary of OLS Results
Variable  Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr
Intercept -3.736846   4.468942 -0.836181   0.405391    6.428359  -0.581306 0.562574
POP       0.015920    0.002113 7.535770    0.000000*   0.004351  3.658727  0.000444*

                                       OLS Diagnostics
Number of Observations:      87         Number of Variables:                       2
Degrees of Freedom:          85         Akaike's Information Criterion (AIC) [2]:  788.762573
Multiple R-Squared [2]:      0.400513   Adjusted R-Squared [2]:                    0.393460
Joint F-Statistic [3]:       56.787826  Prob(>F), (1,85) degrees of freedom:       0.000000*
Joint Wald Statistic [4]:    13.386281  Prob(>chi-squared), (1) degrees of freedom: 0.000253*
Koenker (BP) Statistic [5]:  48.513644  Prob(>chi-squared), (1) degrees of freedom: 0.000000*
Jarque-Bera Statistic [6]:   15.736453  Prob(>chi-squared), (2) degrees of freedom: 0.000383*
```

So looking back at our original hypothesis, is call volume simply a function of population? Might our community be able to predict future 911 call volumes from population projections alone? Probably not; if the relationship between population and 911 call volumes had been higher, say 80%, our community might not need regression at all. But with only 39% of the story, it seems other factors and other variables, are needed to effectively model 911 calls.

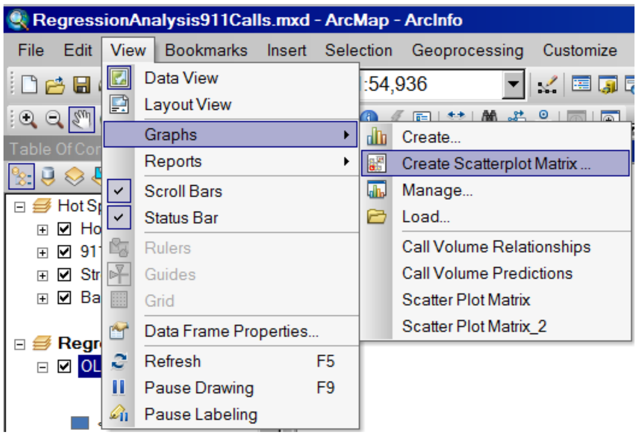
The next question that follows is what are these other variables? This, actually, is the hardest part of the regression model building process: finding *all* of the key variables that explain what we are trying to model.

- Close the Progress Window.

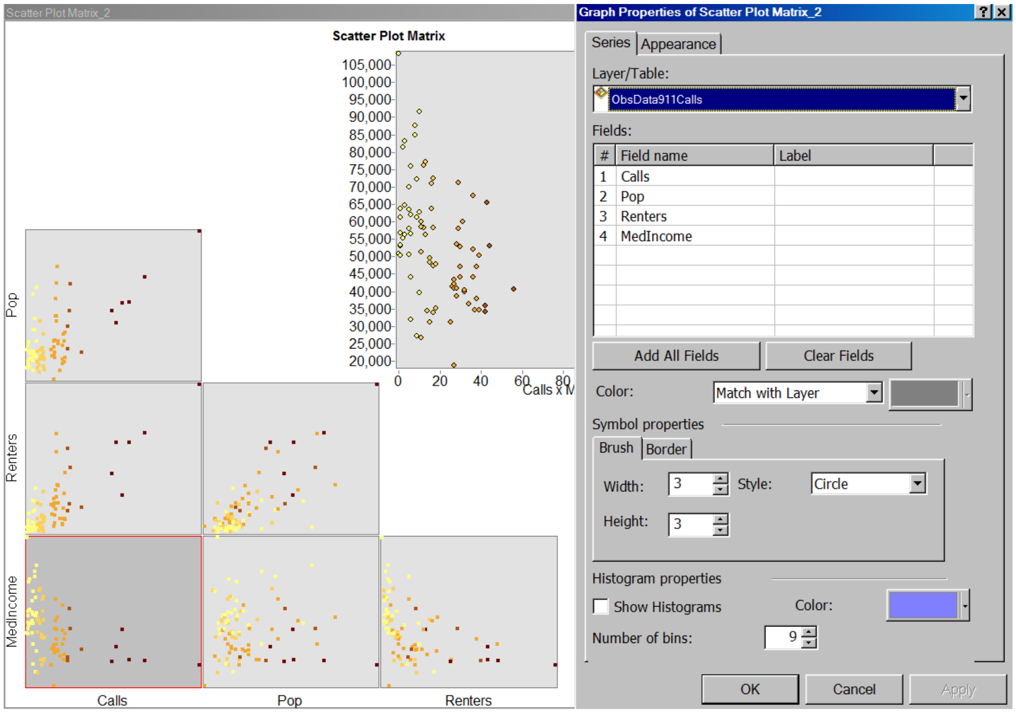
Step 4 Finding key variables

The scatterplot matrix graph can help us here by allowing us to examine the relationships between call volumes and a variety of other variables. We might guess, for example, that the number of apartment complexes, unemployment rates, income or education are also important predictors of 911 call volumes.

- Experiment with the scatterplot matrix graph to explore the relationships between call volumes and other candidate explanatory variables. If you enter the “calls” variable either first or last, it will appear as either the bottom row or the first column in the matrix.



Below is an example of scatterplot matrix parameter settings:



Layer/Table: ObsData911Calls

Fields:

#	Field name	Label
1	Calls	
2	Pop	
3	Renters	
4	MedIncome	

Color: Match with Layer

Width: 3 Style: Circle

Height: 3

☐ Show Histograms Color:

Number of bins: 9

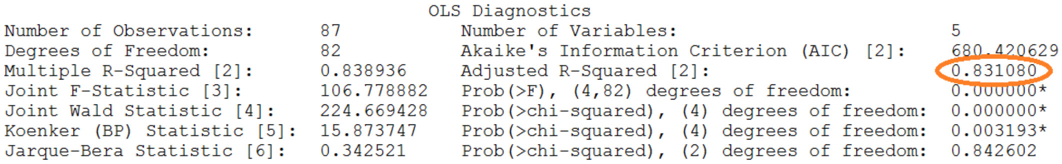
- Once you finish creating the scatterplot matrix, select features in the focus graph and notice how those features are highlighted in each scatterplot and on the map.

Step 5 A properly specified model

Now let’s try a model with 4 explanatory variables: Pop, Jobs, LowEduc, and Dst2UrbCen. The explanatory variables in this model were found by using the Scatterplot matrix and trying a number of candidate models. Finding a properly specified OLS model, is often an iterative process.

- Run OLS with the following parameters set:
  - Input Feature Class -> Analysis\ObsData911Calls
  - Unique ID Field -> UniqID
  - Output Feature Class -> C:\SpatialStats\RegressionExercise\Outputs\Data911CallsOLS.shp
  - Dependent Variable -> Calls
  - Explanatory Variables -> Pop;Jobs;LowEduc;Dst2UrbCen

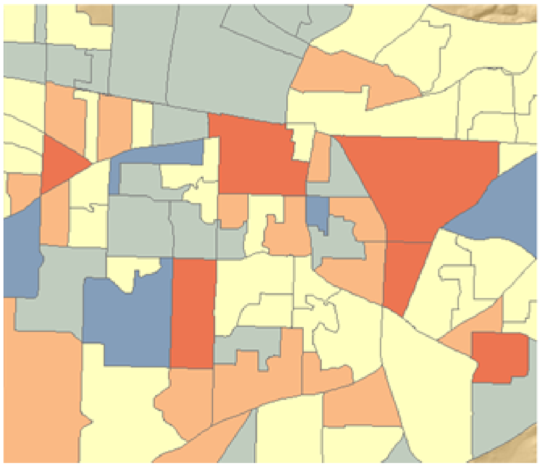
Notice that the Adjusted R2 value is much higher for this new model, 0.831080, indicating this model explains 83% of the 911 call volume story. This is a big improvement over the model that only used Population.



```
                                         OLS Diagnostics
Number of Observations:       87           Number of Variables:                       5
Degrees of Freedom:           82           Akaike's Information Criterion (AIC) [2]:  680.420629
Multiple R-Squared [2]:       0.838936     Adjusted R-Squared [2]:                    0.831080
Joint F-Statistic [3]:        106.778882   Prob(>F), (4,82) degrees of freedom:       0.000000*
Joint Wald Statistic [4]:     224.669428   Prob(>chi-squared), (4) degrees of freedom: 0.000000*
Koenker (BP) Statistic [5]:   15.873747    Prob(>chi-squared), (4) degrees of freedom: 0.003193*
Jarque-Bera Statistic [6]:    0.342521     Prob(>chi-squared), (2) degrees of freedom: 0.842602
```

- Close the Progress Window.

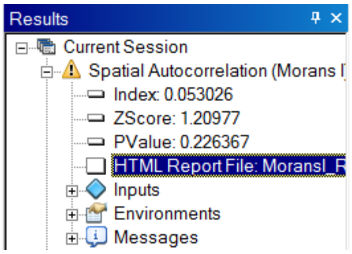
Notice, too, that the residuals (the model over/under predictions) appear to be less clustered than they were using only the Population variable.



We can check whether or not the residuals exhibit a random spatial pattern using the Spatial Autocorrelation tool.

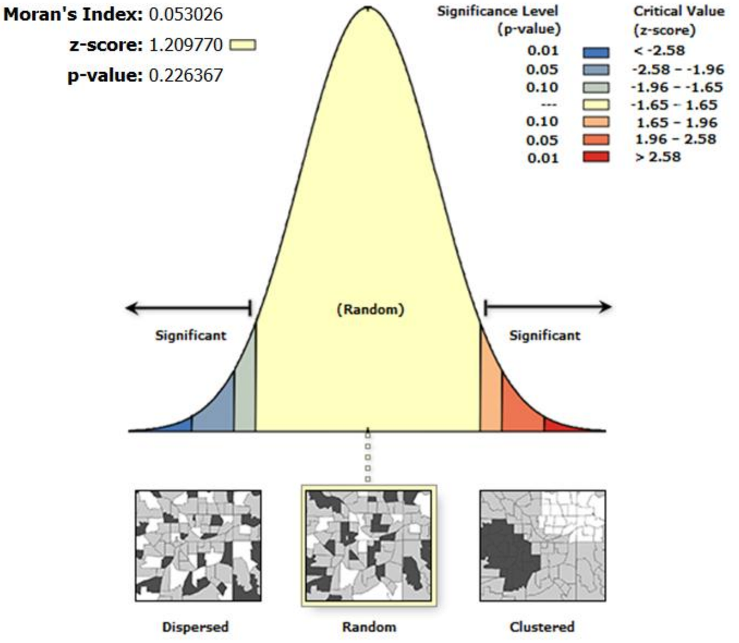
- Run the Spatial Autocorrelation tool (in the Analyzing Patterns Toolset) using the following parameters:
  - Input Feature Class → Data911CallsOLS
  - Input Field → StdResid
  - Generate Report → checked ON
  - Conceptualization of Spatial Relationships → Inverse Distance
  - Distance Method → Euclidean Distance
  - Standardization → ROW (with polygons you will almost always want to Row Standardize).
- Close the Progress Window, then open the Results Window and Expand the entry for Spatial Autocorrelation (if you don't see the Results Window, select Geoprocessing from the menu, then Results).

- Double click on the HTML Report File:



Results

Current Session
Spatial Autocorrelation (Morans I
Index: 0.053026
ZScore: 1.20977
PValue: 0.226367
HTML Report File: MoransI_R
Inputs
Environments
Messages



Moran's Index: 0.053026
z-score: 1.209770
p-value: 0.226367

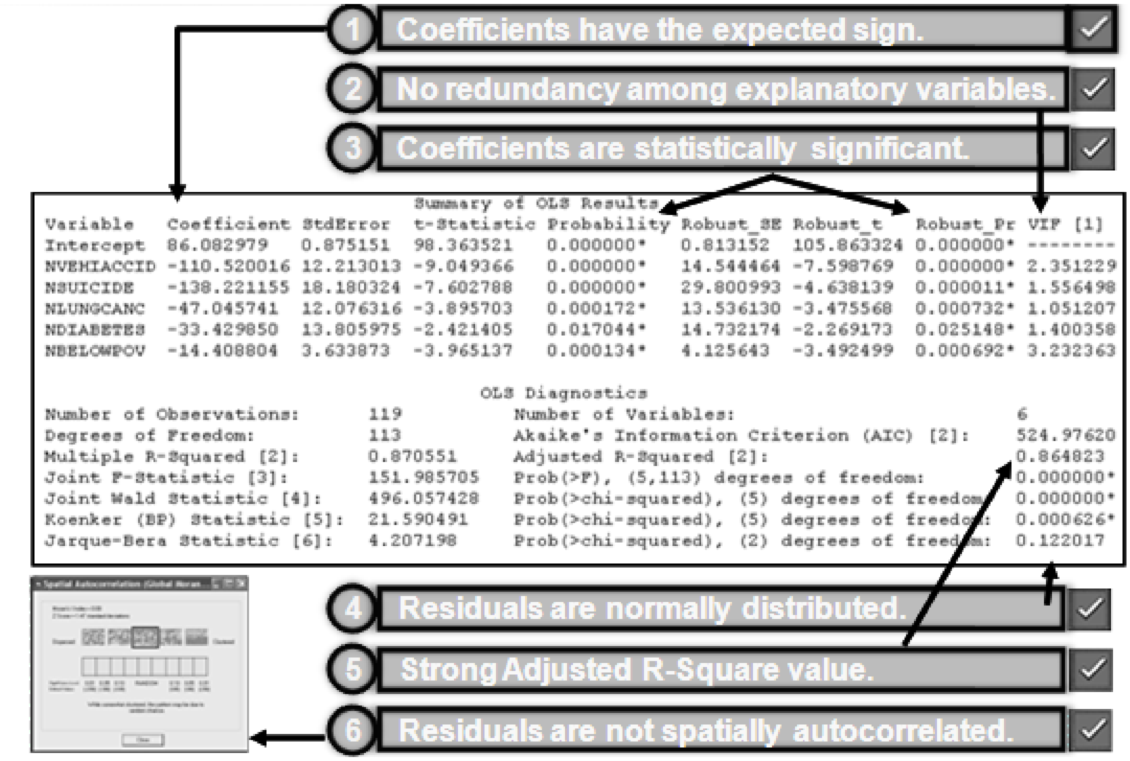
Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	▨	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟥	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of 1.21, the pattern does not appear to be significantly different than random.

Results from running the Spatial Autocorrelation tool on the regression residuals indicates they are randomly distributed; the z-score is not statistically significant so we accept the null hypothesis of complete spatial randomness. This is good news! Anytime there is structure (clustering or dispersion) of the under/over predictions, it means that your model is still missing key explanatory variables and you cannot trust your results. When you run the Spatial Autocorrelation tool on the model residuals and find a random spatial pattern (as we did here), you are on your way to a properly specified model.

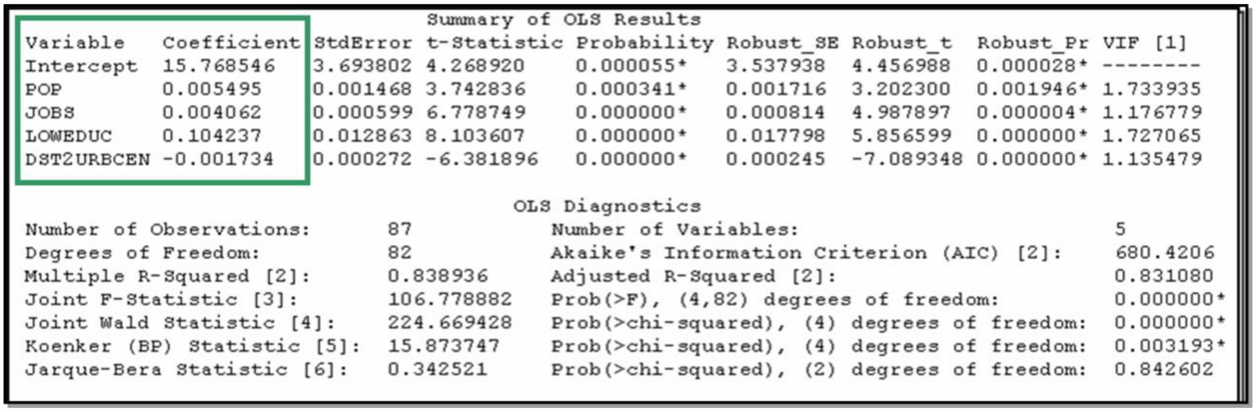
Step 6: The 6 things you gotta check!

There are 6 things you need to check before you can be sure you have a properly specified model – a model you can trust.


1 Coefficients have the expected sign. ✓
2 No redundancy among explanatory variables. ✓
3 Coefficients are statistically significant. ✓
4 Residuals are normally distributed. ✓
5 Strong Adjusted R-Square value. ✓
6 Residuals are not spatially autocorrelated. ✓

```
                                Summary of OLS Results
Variable     Coefficient StdError  t-Statistic Probability Robust_SE Robust_t   Robust_Pr VIF [1]
Intercept    86.082979   0.875151  98.363521   0.000000*   0.813152  105.863324 0.000000* --------
NVEHIACCID  -110.520016 12.213013 -9.049366    0.000000*   14.544464 -7.598769  0.000000* 2.351229
NSUICIDE    -138.221155 18.180324 -7.602788    0.000000*   29.800993 -4.638139  0.000011* 1.556498
NLUNGCANC   -47.045741  12.076316 -3.895703    0.000172*   13.536130 -3.475568  0.000732* 1.051207
NDIABETES   -33.429850  13.805975 -2.421405    0.017044*   14.732174 -2.269173  0.025148* 1.400358
NBELOWPOV   -14.408804   3.633873 -3.965137    0.000134*   4.125643  -3.492499  0.000692* 3.232363

                                   OLS Diagnostics
Number of Observations:      119          Number of Variables:                          6
Degrees of Freedom:          113          Akaike's Information Criterion (AIC) [2]:     524.97620
Multiple R-Squared [2]:      0.870551     Adjusted R-Squared [2]:                       0.864823
Joint F-Statistic [3]:       151.985705   Prob(>F), (5,113) degrees of freedom:         0.000000*
Joint Wald Statistic [4]:    496.057428   Prob(>chi-squared), (5) degrees of freedom    0.000000*
Koenker (BP) Statistic [5]:  21.590491    Prob(>chi-squared), (5) degrees of freedo:    0.000626*
Jarque-Bera Statistic [6]:   4.207198     Prob(>chi-squared), (2) degrees of freedom:   0.122017
```

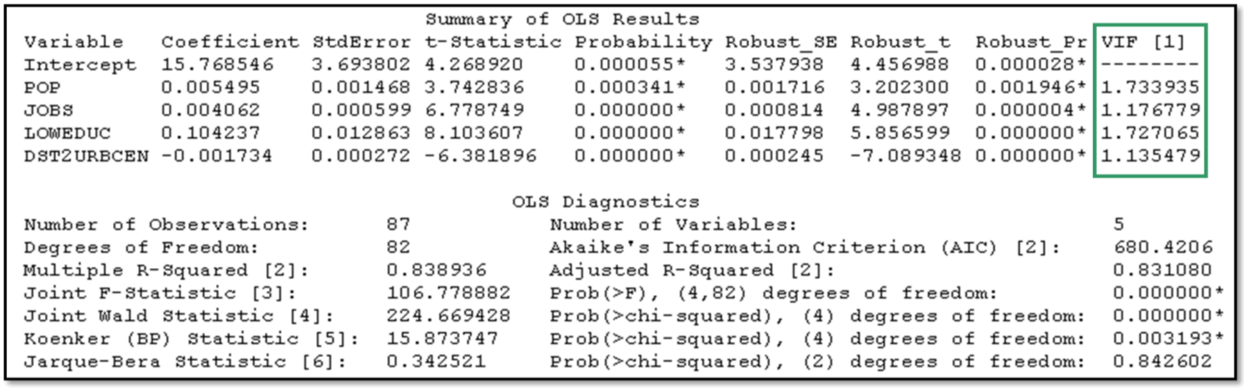
1. First check to see that each coefficient has the “expected” sign.



```
                                   Summary of OLS Results
Variable    Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr VIF [1]
Intercept   15.768546  3.693802 4.268920    0.000055*   3.537938  4.456988  0.000028* --------
POP         0.005495   0.001468 3.742836    0.000341*   0.001716  3.202300  0.001946* 1.733935
JOBS        0.004062   0.000599 6.778749    0.000000*   0.000814  4.987897  0.000004* 1.176779
LOWEDUC     0.104237   0.012863 8.103607    0.000000*   0.017798  5.856599  0.000000* 1.727065
DST2URBCEN -0.001734   0.000272 -6.381896   0.000000*   0.000245  -7.089348 0.000000* 1.135479

                                         OLS Diagnostics
Number of Observations:        87          Number of Variables:                          5
Degrees of Freedom:            82          Akaike's Information Criterion (AIC) [2]:    680.4206
Multiple R-Squared [2]:        0.838936    Adjusted R-Squared [2]:                      0.831080
Joint F-Statistic [3]:         106.778882  Prob(>F), (4,82) degrees of freedom:         0.000000*
Joint Wald Statistic [4]:      224.669428  Prob(>chi-squared), (4) degrees of freedom:  0.000000*
Koenker (BP) Statistic [5]:    15.873747   Prob(>chi-squared), (4) degrees of freedom:  0.003193*
Jarque-Bera Statistic [6]:     0.342521    Prob(>chi-squared), (2) degrees of freedom:  0.842602
```

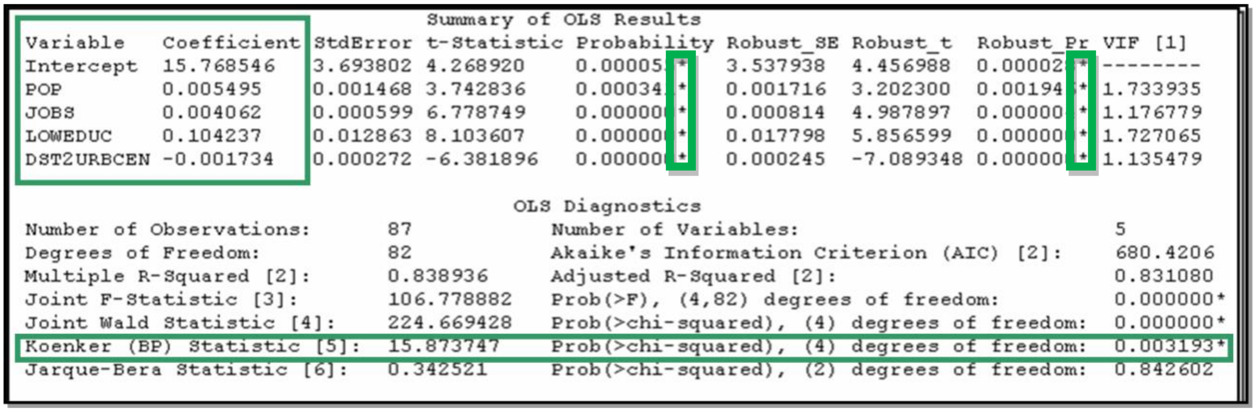
A positive coefficient means the relationship is positive; a negative coefficient means the relationship is negative. Notice that the coefficient for the Pop variable is positive. This means that as the number of people goes up, the number of 911 calls also goes up. We are expecting a positive coefficient. If the coefficient for the Population variable was negative, we would not trust our model. Checking the other coefficients, it seems that their signs do seem reasonable. Self check: the sign for Jobs (the number of job positions in a tract) is positive, this means that as the number of jobs goes <up/down>, the number of 911 calls also goes <up/down> (?).

2. Next check for redundancy among your explanatory variables. If the VIF value (variance inflation factor) for any of your variables is larger than about 7.5 (smaller is definitely better), it means you have one or more variables telling the same story. This leads to an over-count type of bias. You should remove the variables associated with large VIF values one by one until none of your variables have large VIF values. Self check: Which variable has the highest VIF value? <POP, JOBS, LOWEDUC, DST2URBCEN>



```
                                   Summary of OLS Results
Variable    Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr VIF [1]
Intercept   15.768546  3.693802 4.268920    0.000055*   3.537938  4.456988  0.000028* --------
POP         0.005495   0.001468 3.742836    0.000341*   0.001716  3.202300  0.001946* 1.733935
JOBS        0.004062   0.000599 6.778749    0.000000*   0.000814  4.987897  0.000004* 1.176779
LOWEDUC     0.104237   0.012863 8.103607    0.000000*   0.017798  5.856599  0.000000* 1.727065
DST2URBCEN -0.001734   0.000272 -6.381896   0.000000*   0.000245  -7.089348 0.000000* 1.135479

                                         OLS Diagnostics
Number of Observations:        87          Number of Variables:                          5
Degrees of Freedom:            82          Akaike's Information Criterion (AIC) [2]:    680.4206
Multiple R-Squared [2]:        0.838936    Adjusted R-Squared [2]:                      0.831080
Joint F-Statistic [3]:         106.778882  Prob(>F), (4,82) degrees of freedom:         0.000000*
Joint Wald Statistic [4]:      224.669428  Prob(>chi-squared), (4) degrees of freedom:  0.000000*
Koenker (BP) Statistic [5]:    15.873747   Prob(>chi-squared), (4) degrees of freedom:  0.003193*
Jarque-Bera Statistic [6]:     0.342521    Prob(>chi-squared), (2) degrees of freedom:  0.842602
```

3. Next, check to see that all of the explanatory variables have statistically significant coefficients.



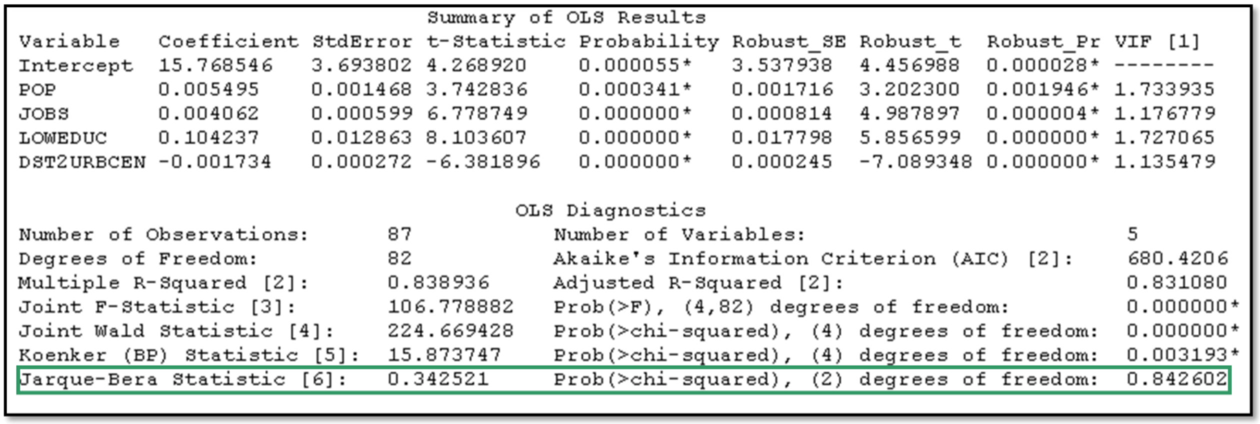
```
                                      Summary of OLS Results
Variable    Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr VIF [1]
Intercept   15.768546  3.693802 4.268920     0.00005 *  3.537938  4.456988  0.00002 * --------
POP         0.005495   0.001468 3.742836     0.00034 *  0.001716  3.202300  0.00194 * 1.733935
JOBS        0.004062   0.000599 6.778749     0.00000 *  0.000814  4.987897  0.00000 * 1.176779
LOWEDUC     0.104237   0.012863 8.103607     0.00000 *  0.017798  5.856599  0.00000 * 1.727065
DST2URBCEN -0.001734   0.000272 -6.381896    0.00000 *  0.000245  -7.089348 0.00000 * 1.135479

                                          OLS Diagnostics
Number of Observations:        87           Number of Variables:                          5
Degrees of Freedom:            82           Akaike's Information Criterion (AIC) [2]:     680.4206
Multiple R-Squared [2]:        0.838936     Adjusted R-Squared [2]:                       0.831080
Joint F-Statistic [3]:         106.778882   Prob(>F), (4,82) degrees of freedom:          0.000000*
Joint Wald Statistic [4]:      224.669428   Prob(>chi-squared), (4) degrees of freedom:   0.000000*
Koenker (BP) Statistic [5]:    15.873747    Prob(>chi-squared), (4) degrees of freedom:   0.003193*
Jarque-Bera Statistic [6]:     0.342521     Prob(>chi-squared), (2) degrees of freedom:   0.842602
```

Two columns, Probability and Robust Probability, measure coefficient statistical significance. An asterisk next to the probability tells you the coefficient is significant. If a variable is not significant, it is not helping the model, and unless theory tells us that a particular variable is critical, we should remove it. When the Koenker (BP) statistic is statistically significant, you can only trust the Robust Probability column to determine if a coefficient is significant or not. Small probabilities are "better" (more significant) than large probabilities. Self check: Which variables have the "best" statistical significance? Did you consult the Probability or Robust_Pr column? Why?

! Note: An asterisk indicates statistical significance

Three checks down!
You're ½ way there!

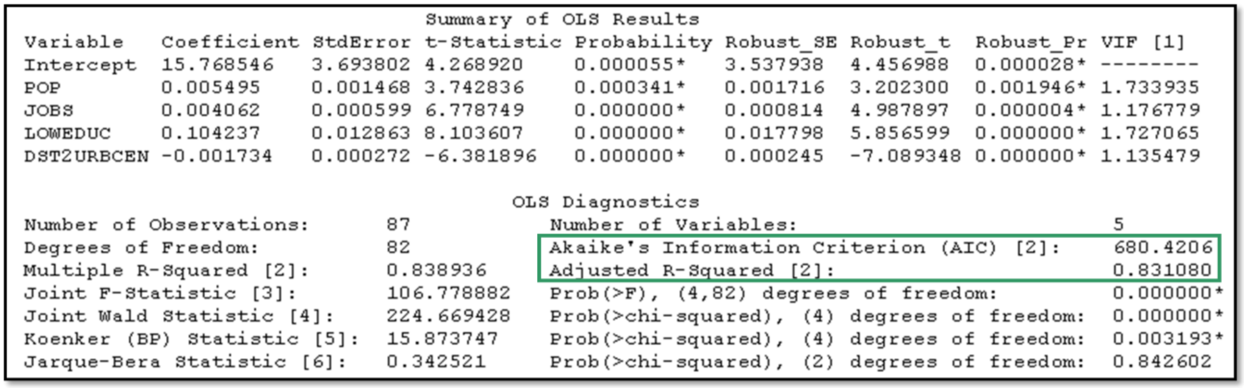
4. Make sure the Jarque-Bera test is *NOT* statistically significant:



```
                                      Summary of OLS Results
Variable    Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr VIF [1]
Intercept   15.768546   3.693802 4.268920    0.000055*   3.537938  4.456988  0.000028* --------
POP         0.005495    0.001468 3.742836    0.000341*   0.001716  3.202300  0.001946* 1.733935
JOBS        0.004062    0.000599 6.778749    0.000000*   0.000814  4.987897  0.000004* 1.176779
LOWEDUC     0.104237    0.012863 8.103607    0.000000*   0.017798  5.856599  0.000000* 1.727065
DST2URBCEN -0.001734    0.000272 -6.381896   0.000000*   0.000245  -7.089348 0.000000* 1.135479

                                          OLS Diagnostics
Number of Observations:        87           Number of Variables:                          5
Degrees of Freedom:            82           Akaike's Information Criterion (AIC) [2]:     680.4206
Multiple R-Squared [2]:        0.838936     Adjusted R-Squared [2]:                       0.831080
Joint F-Statistic [3]:         106.778882   Prob(>F), (4,82) degrees of freedom:          0.000000*
Joint Wald Statistic [4]:      224.669428   Prob(>chi-squared), (4) degrees of freedom:   0.000000*
Koenker (BP) Statistic [5]:    15.873747    Prob(>chi-squared), (4) degrees of freedom:   0.003193*
Jarque-Bera Statistic [6]:     0.342521     Prob(>chi-squared), (2) degrees of freedom:   0.842602
```

The residuals (over/under predictions) from a properly specified model will reflect random noise. Random noise has a random spatial pattern (no clustering of over/under predictions). It also has a normal histogram if you plotted the residuals. The Jarque-Bera test measures whether or not the residuals from a regression model are normally distributed (think Bell Curve). This is the one test you do NOT want to be statistically significant! When it IS statistically significant, your model is biased. This often means you are missing one or more key explanatory variables. Self check: how do you know that the Jarque-Bera Statistic is NOT statistically significant in this case?

5. Next, you want to check model performance:



```
                                   Summary of OLS Results
Variable    Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr VIF [1]
Intercept   15.768546   3.693802 4.268920    0.000055*   3.537938  4.456988  0.000028* --------
POP         0.005495    0.001468 3.742836    0.000341*   0.001716  3.202300  0.001946* 1.733935
JOBS        0.004062    0.000599 6.778749    0.000000*   0.000814  4.987897  0.000004* 1.176779
LOWEDUC     0.104237    0.012863 8.103607    0.000000*   0.017798  5.856599  0.000000* 1.727065
DST2URBCEN -0.001734    0.000272 -6.381896   0.000000*   0.000245  -7.089348 0.000000* 1.135479

                                          OLS Diagnostics
Number of Observations:       87          Number of Variables:                          5
Degrees of Freedom:           82          Akaike's Information Criterion (AIC) [2]:     680.4206
Multiple R-Squared [2]:       0.838936    Adjusted R-Squared [2]:                       0.831080
Joint F-Statistic [3]:        106.778882  Prob(>F), (4,82) degrees of freedom:          0.000000*
Joint Wald Statistic [4]:     224.669428  Prob(>chi-squared), (4) degrees of freedom:   0.000000*
Koenker (BP) Statistic [5]:   15.873747   Prob(>chi-squared), (4) degrees of freedom:   0.003193*
Jarque-Bera Statistic [6]:    0.342521    Prob(>chi-squared), (2) degrees of freedom:   0.842602
```

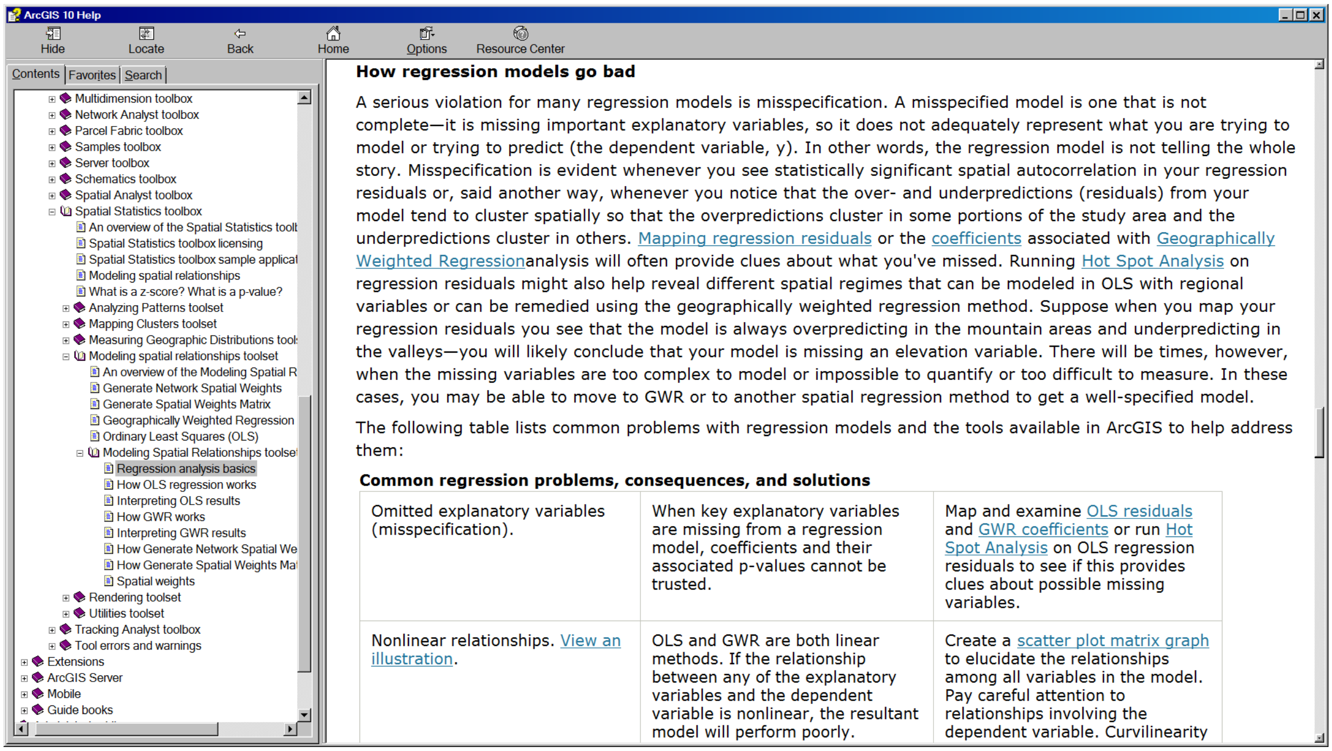
The adjusted R squared value ranges from 0 to 1.0 and tells you how much of the variation in your dependent variable has been explained by the model. Generally we are looking for values of 0.5 or higher, but a "good" R2 value depends on what we are modeling. Self Check: go back to the screen shot of the OLS model that only used Population to explain call volume. What was the Adjusted R2 value? Does the Adjusted R2 value for our new model (4 variables) indicate model performance has improved?

The AIC value can also be used to measure model performance. When we have several candidate models (all models must have the same dependent variable), we can assess which model is best by looking for the lowest AIC value. Self Check: go back to the screen shot of the OLS model that only used Population. What was the AIC value? Does the AIC value for our new model (4 variables) indicate we improved model performance?

6. Finally (but certainly NOT least important), you want to make sure your model residuals are free from spatial autocorrelation (spatial clustering of over and under predictions).

   We used the Spatial Autocorrelation tool above and found that our model passes this check too. This will not always be the case when you build your own regression models, however.

- Find the Regression Analysis Basics online documentation, and look for the table called “How Regression Models Go Bad”. In this table there are some strategies for how to deal with Spatially Autocorrelated regression residuals:


ArcGIS 10 Help
Hide
Locate
Back
Home
Options
Resource Center
Contents
Favorites
Search
Multidimension toolbox
Network Analyst toolbox
Parcel Fabric toolbox
Samples toolbox
Server toolbox
Schematics toolbox
Spatial Analyst toolbox
Spatial Statistics toolbox
An overview of the Spatial Statistics toolb
Spatial Statistics toolbox licensing
Spatial Statistics toolbox sample applicat
Modeling spatial relationships
What is a z-score? What is a p-value?
Analyzing Patterns toolset
Mapping Clusters toolset
Measuring Geographic Distributions tool
Modeling spatial relationships toolset
An overview of the Modeling Spatial R
Generate Network Spatial Weights
Generate Spatial Weights Matrix
Geographically Weighted Regression
Ordinary Least Squares (OLS)
Modeling Spatial Relationships toolse
Regression analysis basics
How OLS regression works
Interpreting OLS results
How GWR works
Interpreting GWR results
How Generate Network Spatial We
How Generate Spatial Weights Ma
Spatial weights
Rendering toolset
Utilities toolset
Tracking Analyst toolbox
Tool errors and warnings
Extensions
ArcGIS Server
Mobile
Guide books

How regression models go bad

A serious violation for many regression models is misspecification. A misspecified model is one that is not complete—it is missing important explanatory variables, so it does not adequately represent what you are trying to model or trying to predict (the dependent variable, y). In other words, the regression model is not telling the whole story. Misspecification is evident whenever you see statistically significant spatial autocorrelation in your regression residuals or, said another way, whenever you notice that the over- and underpredictions (residuals) from your model tend to cluster spatially so that the overpredictions cluster in some portions of the study area and the underpredictions cluster in others. Mapping regression residuals or the coefficients associated with Geographically Weighted Regressionanalysis will often provide clues about what you've missed. Running Hot Spot Analysis on regression residuals might also help reveal different spatial regimes that can be modeled in OLS with regional variables or can be remedied using the geographically weighted regression method. Suppose when you map your regression residuals you see that the model is always overpredicting in the mountain areas and underpredicting in the valleys—you will likely conclude that your model is missing an elevation variable. There will be times, however, when the missing variables are too complex to model or impossible to quantify or too difficult to measure. In these cases, you may be able to move to GWR or to another spatial regression method to get a well-specified model.

The following table lists common problems with regression models and the tools available in ArcGIS to help address them:

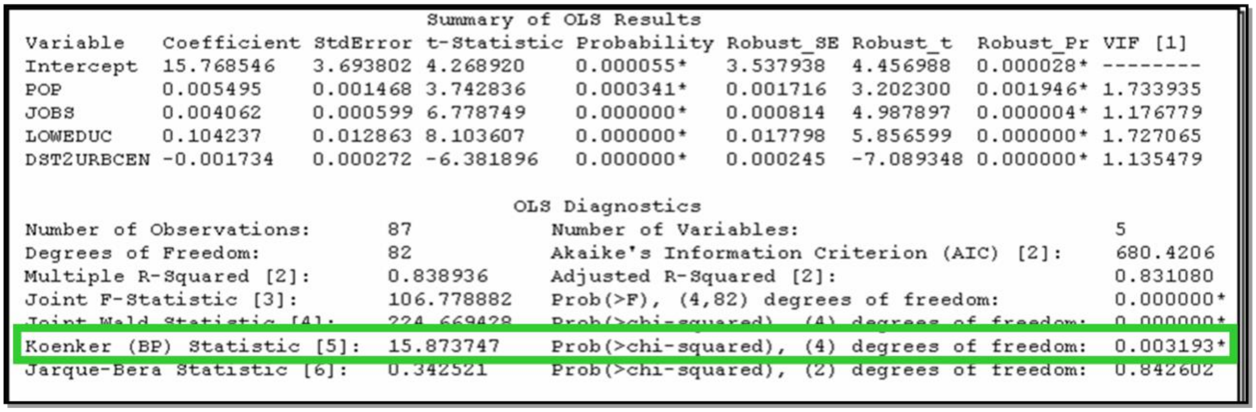
Common regression problems, consequences, and solutions

Omitted explanatory variables (misspecification).	When key explanatory variables are missing from a regression model, coefficients and their associated p-values cannot be trusted.	Map and examine OLS residuals and GWR coefficients or run Hot Spot Analysis on OLS regression residuals to see if this provides clues about possible missing variables.
Nonlinear relationships. View an illustration.	OLS and GWR are both linear methods. If the relationship between any of the explanatory variables and the dependent variable is nonlinear, the resultant model will perform poorly.	Create a scatter plot matrix graph to elucidate the relationships among all variables in the model. Pay careful attention to relationships involving the dependent variable. Curvilinearity

Self check: run OLS on alternate models. Use “Calls” for your dependent variable, with other variables in the ObsData911Calls feature class for your explanatory variables (you might select Jobs, Renters, and MedIncome, for example). For each model, go through the 6 checks above to determine if the model is properly specified. If the model fails one of the checks, look at the “Common Regression Problems, Consequences, and Solutions” table in the “Regression Analysis Basics” document shown above to determine the implications and possible solutions.

Step 7: Running GWR

One OLS diagnostic we didn't say very much about, is the Koenker test.

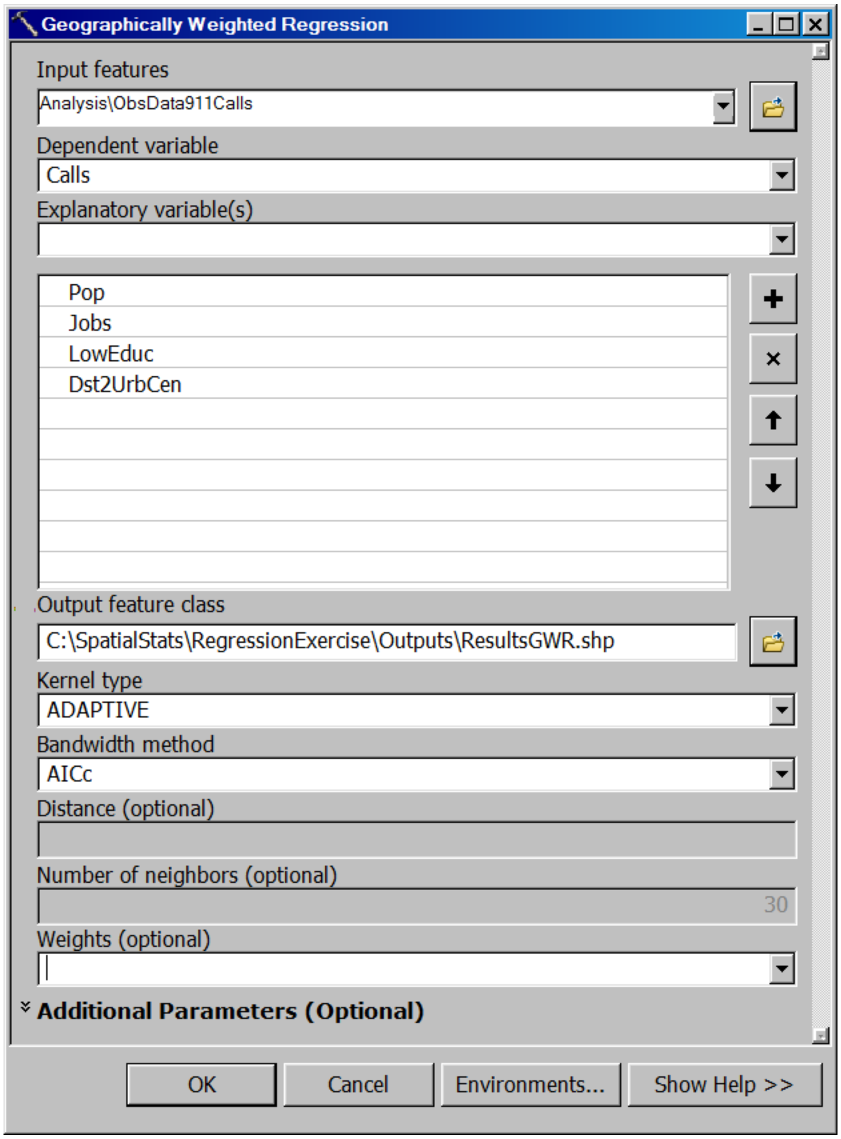


```
                              Summary of OLS Results
Variable    Coefficient StdError t-Statistic Probability Robust_SE Robust_t  Robust_Pr VIF [1]
Intercept   15.768546   3.693802 4.268920    0.000055*   3.537938  4.456988  0.000028* --------
POP         0.005495    0.001468 3.742836    0.000341*   0.001716  3.202300  0.001946* 1.733935
JOBS        0.004062    0.000599 6.778749    0.000000*   0.000814  4.987897  0.000004* 1.176779
LOWEDUC     0.104237    0.012863 8.103607    0.000000*   0.017798  5.856599  0.000000* 1.727065
DST2URBCEN -0.001734    0.000272 -6.381896   0.000000*   0.000245  -7.089348 0.000000* 1.135479

                                   OLS Diagnostics
Number of Observations:       87          Number of Variables:                      5
Degrees of Freedom:           82          Akaike's Information Criterion (AIC) [2]: 680.4206
Multiple R-Squared [2]:       0.838936    Adjusted R-Squared [2]:                   0.831080
Joint F-Statistic [3]:        106.778882  Prob(>F), (4,82) degrees of freedom:      0.000000*
Joint Wald Statistic [4]:     224.669428  Prob(>chi-squared), (4) degrees of freedom: 0.000000*
Koenker (BP) Statistic [5]:   15.873747   Prob(>chi-squared), (4) degrees of freedom: 0.003193*
Jarque-Bera Statistic [6]:    0.342521    Prob(>chi-squared), (2) degrees of freedom: 0.842602
```

When the Koeker test is statistically significant, as it is here, it indicates relationships between some or all of your explanatory variables and your dependent variable are non-stationary. This means, for example, that the population variable might be an important predictor of 911 call volumes in some locations of your study, but perhaps a weak predictor in other locations. Whenever you notice that the Koenker test is statistically significant, it indicates you will likely improve model results by moving to Geographically Weighted Regression.

The good news is that once you've found your key explanatory variables using OLS, running GWR is actually quite simple. In most cases, GWR will use the same dependent and explanatory variables you used in OLS.

- Run the Geographically Weighted Regression tool with the following parameters (open the side panel help and review the parameter descriptions):
    - Input feature class: ObsData911Calls
    - Dependent variable: Calls
    - Explanatory variables: Pop, Jobs, LowEduc, Dst2UrbCen
    - Output feature class: C:\SpatialStats\RegressionExercise\Outputs\ResultsGWR.shp
    - Kernal type: ADAPTIVE
    - Bandwidth method: AICs (you will let the tool find the optimal number of neighbors)



Geographically Weighted Regression

Input features: Analysis\ObsData911Calls

Dependent variable: Calls

Explanatory variable(s):

- Pop
- Jobs
- LowEduc
- Dst2UrbCen

Output feature class: C:\SpatialStats\RegressionExercise\Outputs\ResultsGWR.shp

Kernel type: ADAPTIVE

Bandwidth method: AICc

Distance (optional):

Number of neighbors (optional): 30

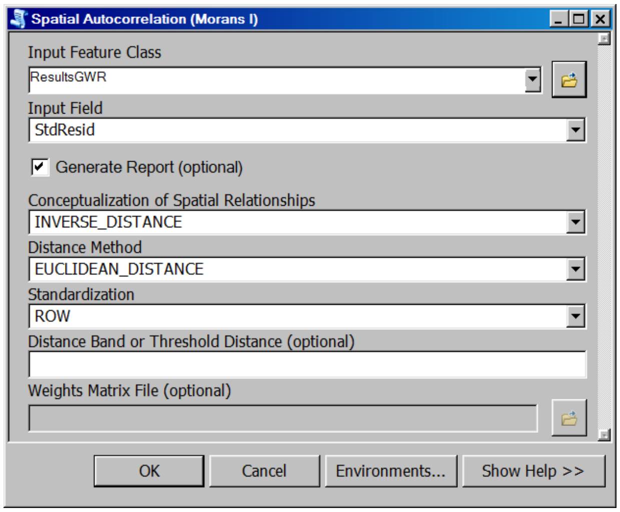
Weights (optional):

- Notice the output from GWR:

```
Neighbors              : 50
ResidualSquares        : 7326.2793171502362
EffectiveNumber        : 19.863531396247254
Sigma                  : 10.44629989196762
AICc                   : 674.6519110481853
R2                     : 0.89572753438054042
R2Adjusted             : 0.86642979248431506
```

GWR found, applying the AICc method, that using 50 neighbors to calibrate each local regression equation yields optimal results (minimized bias and maximized model fit). Notice that the Adjusted R2 value is higher for GWR than it was for our best OLS model (OLS was 83%; GWR is almost 86.6%). The AICc value is lower for the GWR model. A decrease of more than even 3 points indicates a real improvement in model performance (OLS was 680; GWR is 674).

- Close the progress window. Notice that, like the OLS tool, GWR default output is a map of model residuals. Do the over and under predictions appear random? It's a bit difficult to tell. Run the Spatial Autocorrelation tool on the Standardized Residuals in the Output Feature Class:



Spatial Autocorrelation (Morans I)

Input Feature Class
ResultsGWR

Input Field
StdResid

☑ Generate Report (optional)

Conceptualization of Spatial Relationships
INVERSE_DISTANCE

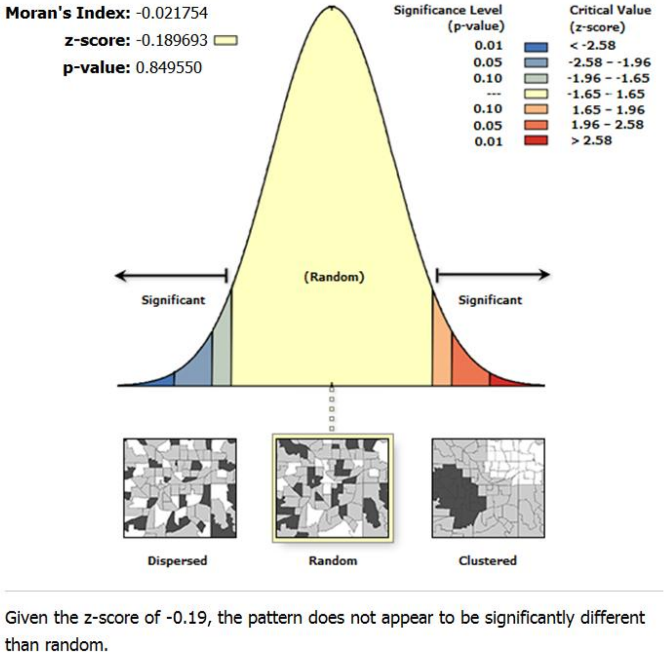
Distance Method
EUCLIDEAN_DISTANCE

Standardization
ROW

Distance Band or Threshold Distance (optional)

Weights Matrix File (optional)

- Close the Progress Window, then double click on the HTML Report in the Results Window to see that the residuals do, in fact, reflect a random spatial pattern.



Moran's Index: -0.021754
z-score: -0.189693 🟨
p-value: 0.849550

Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	🟩	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

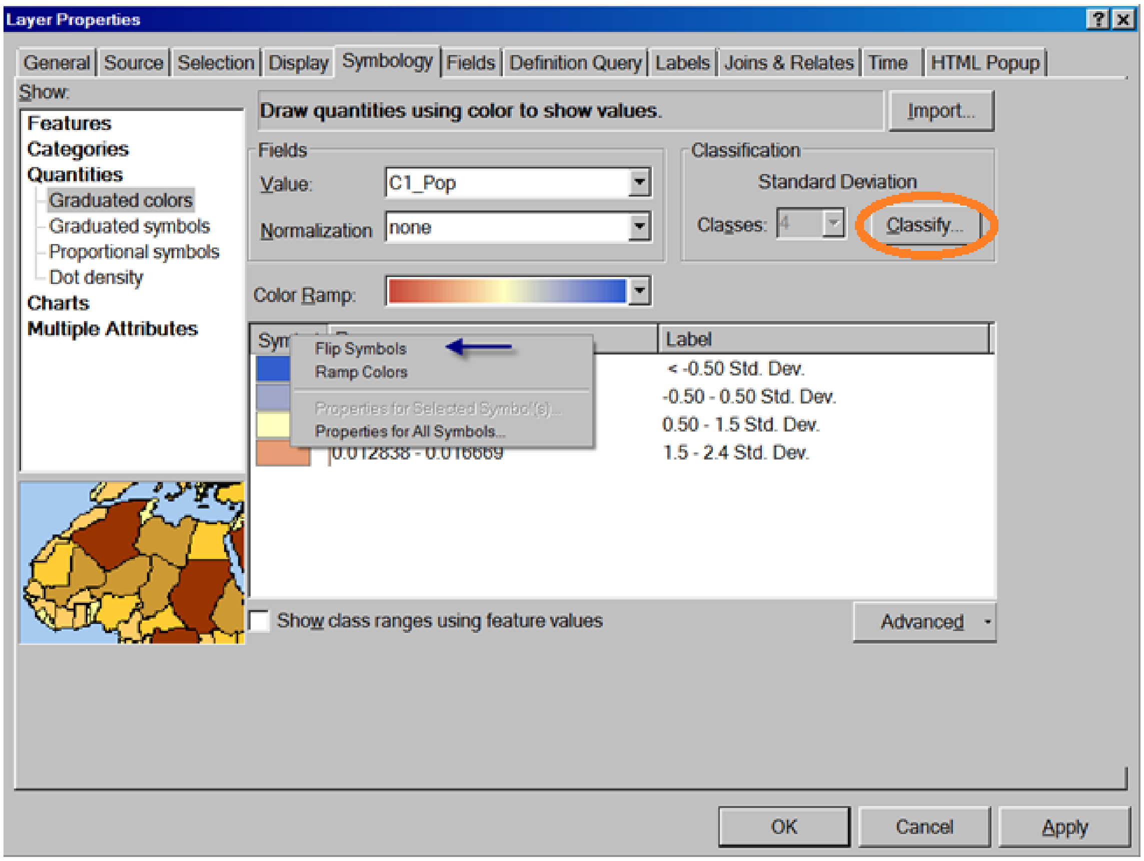
Given the z-score of -0.19, the pattern does not appear to be significantly different than random.

- Open the table for the ResultsGWR output feature class and notice several fields with names beginning with "C". These are the coefficient values for each explanatory variable, for each feature.

Mapping these coefficients shows you how the relationship between each explanatory variable and the dependent variable changes across the study area.

- Right click on ResultsGWR, select Properties and then click on the Symbology tab.

- Render the coefficients for C1_Pop, C2_Jobs, and C3_LowEduc. Use Standard Deviation class breaks and a cold to hot rendering scheme. You may need to Flip the color ramp. When the relationship is positive (as it is for Pop, Jobs, and LowEduc), the red areas show where the coefficients are large; these are the locations where the explanatory variable you are mapping is a strong predictor.



Draw quantities using color to show values.

Fields

Value: C1_Pop

Normalization: none

Classification

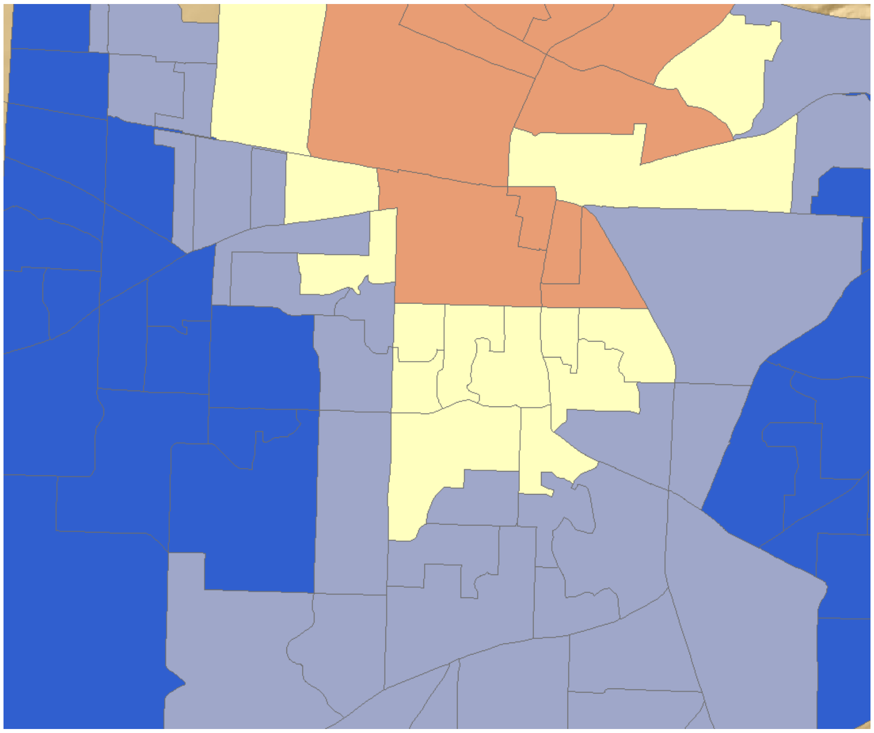
Standard Deviation

Classes: 4

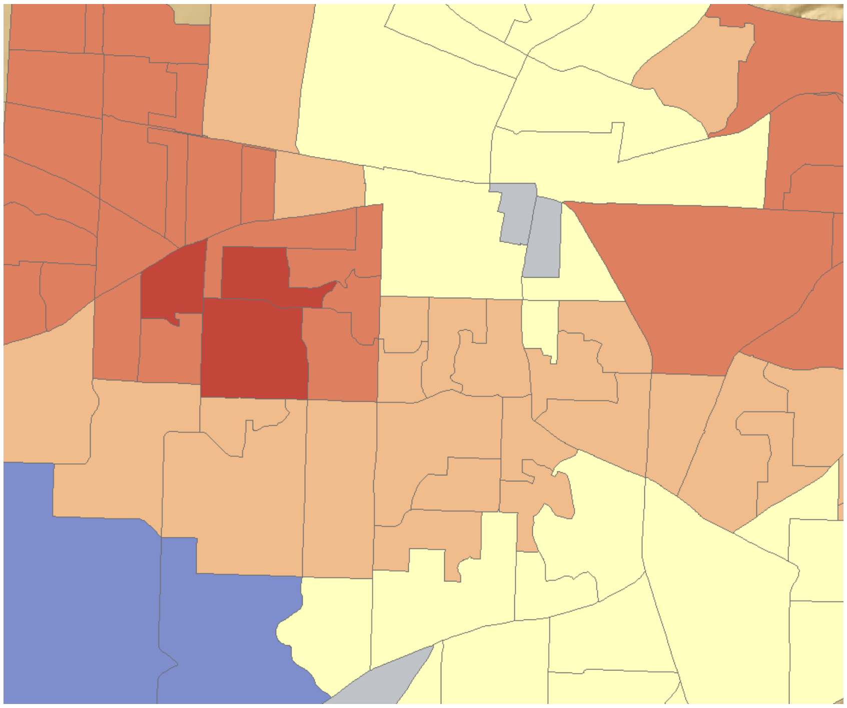
Sym	R	Label
🟦		< -0.50 Std. Dev.
🟪		-0.50 - 0.50 Std. Dev.
🟨		0.50 - 1.5 Std. Dev.
🟧	0.012838 - 0.016669	1.5 - 2.4 Std. Dev.

□ Show class ranges using feature values

This is a map of the coefficients for the Population (Pop) variable:



This is the map of the coefficients for Low Education (LowEdu):



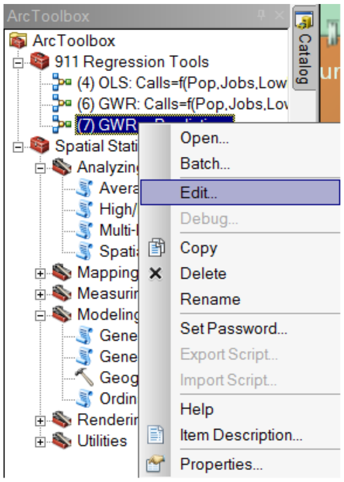
In the map above, the red areas are locations where the low education variable is a strong predictor of the number of 911 calls. The blue areas are locations where low education isn't very strong. Suppose your community decides that, as a way to reduce 911 calls, as well as to promote overall community benefits, they are going to implement a program designed to encourage kids to say in school. They certainly could apply this new program to the entire community. But if resources are limited (and resources are always limited), they might elect to begin a rollout for this new program in those areas where low education has a strong relationship to 911 call volumes (the red areas).

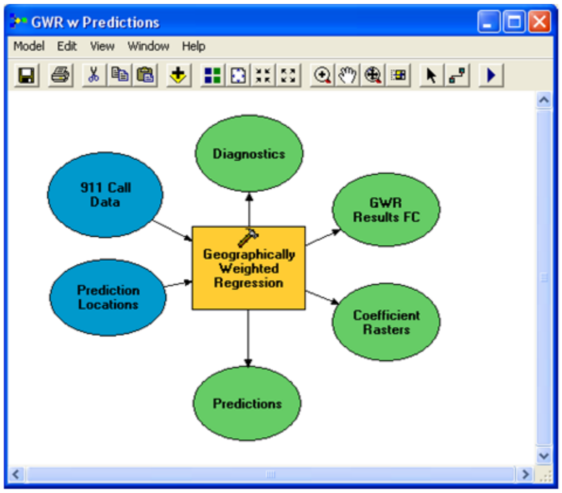
Anytime the variables in your model have policy implications or are associated with particular remediation strategies, you can use GWR to better target where those polices and projects are likely to have the biggest impact.

Step 8: GWR Predictions

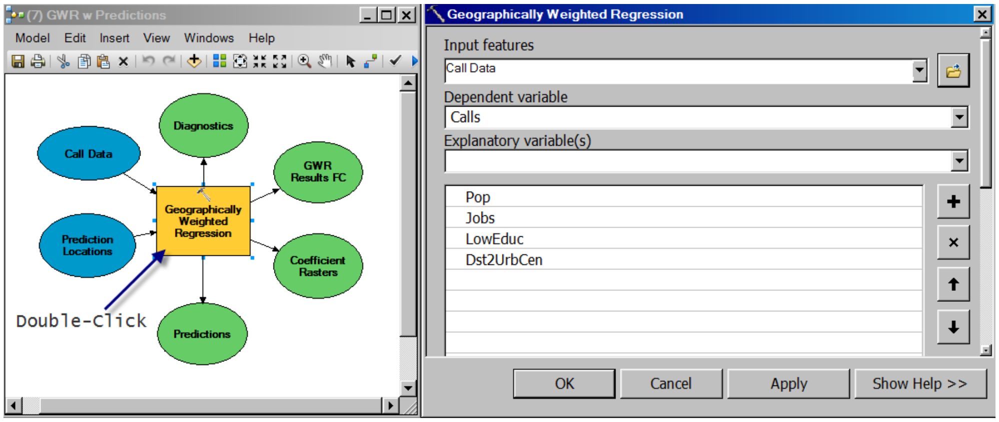
GWR may also be used to predict values for a future time or for locations within the study area where you have X values (known values for each explanatory variable), but don't know what the Y values (your dependent variable) are. In this next step we will explore using GWR to predict future 911 call volumes. We will run GWR using a model that has already been created for you in the 911 Regression Tools toolbox.

- In the TOC, Activate the Predictions View (right click and select Activate)
- Edit the **GWR w Predictions** model





- Double click the GWR tool in the model to see the parameter settings.



Input features: Call Data

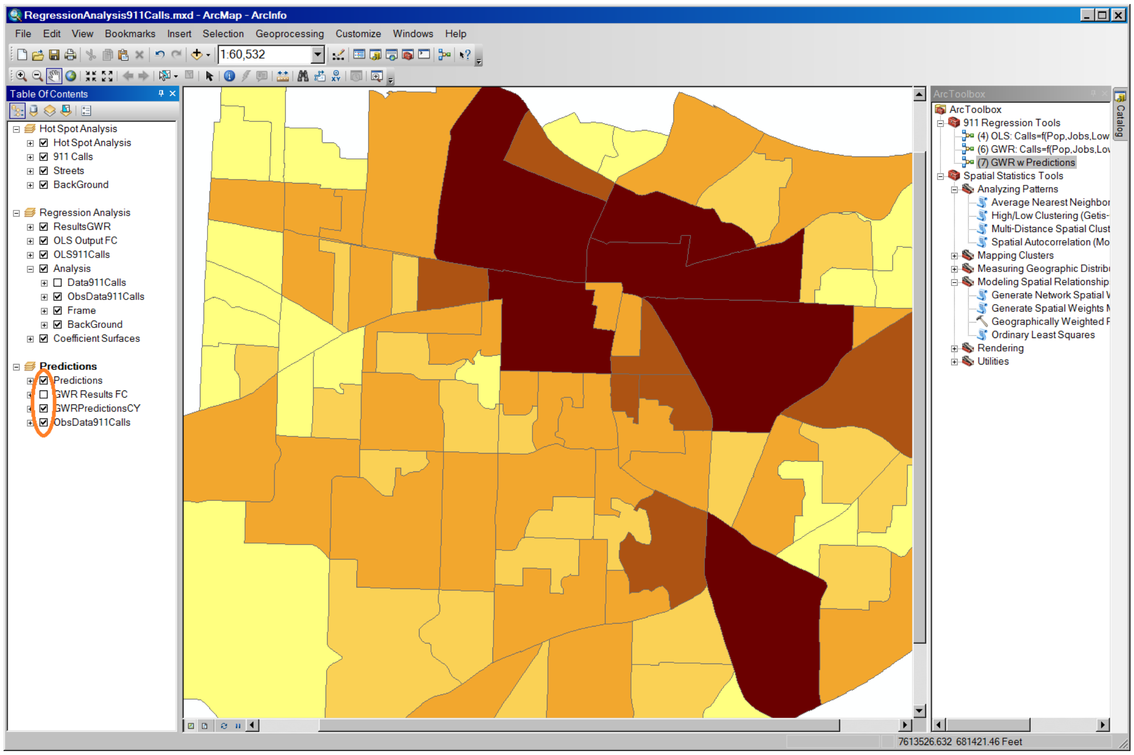
Dependent variable: Calls

Explanatory variable(s): Pop, Jobs, LowEduc, Dst2UrbCen

- Notice that the model is calibrated using the variables we've been using all along, but that the explanatory variables for the predictions are new. The new variables represent projected population, job, and education variables for some time in the future.
  - Input feature class: Call Data
  - Dependent variable: Calls
  - Explanatory variable(s): Pop, Jobs, LowEduc, Dst2UrbCen
  - Output feature class: C:\SpatialStats\RegressionExercise\Outputs\ResultsGWR_CY.shp
  - Kernel type: ADAPTIVE
  - Bandwidth method: BANDWIDTH PARAMETER
  - Number of neighbors: 50
  - Prediction locations: Prediction Locations
  - Prediction explanatory variable(s): PopFY, JobsFY, LowEducFY, Dst2UrbCen (note: the order must match the Explanatory variable(s) list exactly)
  - Output prediction feature class: C:\SpatialStats\RegressionExercise\Outputs\GWRCallPredictionsFY.shp

- Run the model ▶

- When the model finishes, toggle on and off the layers representing the actually 911 call data (obsData911Calls), the model predictions for the current year (GWRPredictionsCY), and the future 911 call volume predictions (GWRPredictionsFY). Notice the differences.



Conclusion:

You used Ordinary Least Squares (OLS) regression to see if population alone would explain 911 emergency call volumes. The scatterplot matrix tool allowed you to explore other candidate explanatory variables that might allow you to improve your model. You checked the results from OLS to determine whether or not you had a properly specified OLS model. Noting that the Koenker test was statistically significant, indicating non-stationarity among variable relationships, you moved to GWR to see if you could improve your regression model. Your analysis provided your community with a number of important insights:

- Hot Spot Analysis allowed them to evaluate how well the fire and police units are currently located in relation to 911 call demand.
- OLS helped them identify the key factors contributing to 911 call volumes.
- Where those factors suggested remediation or policy changes, GWR helped them identify the neighborhoods where those remediation projects might be most effective.
- GWR also predicted 911 call volumes for the future, allowing them not only to anticipate future demand, but also providing a yard stick to assess the effectiveness of implemented remediation.

To learn more about OLS and GWR see the Spatial Statistics Resources at www.bit.ly/spatialstats