X

Regression Analysis Basics

Exercise 1: Analyze 911 response data using regression analysis
Estimated time: 50 minutes

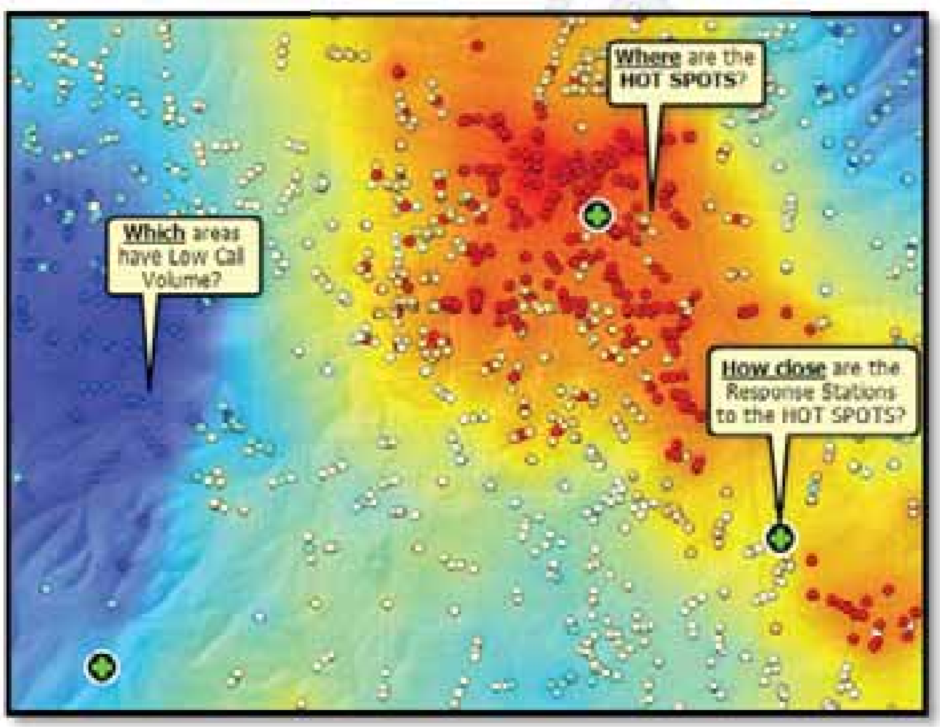
Exercise 1: Analyze 911 response data using regression analysis

Estimated time: 50 minutes

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.

In order to illustrate 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 you 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.

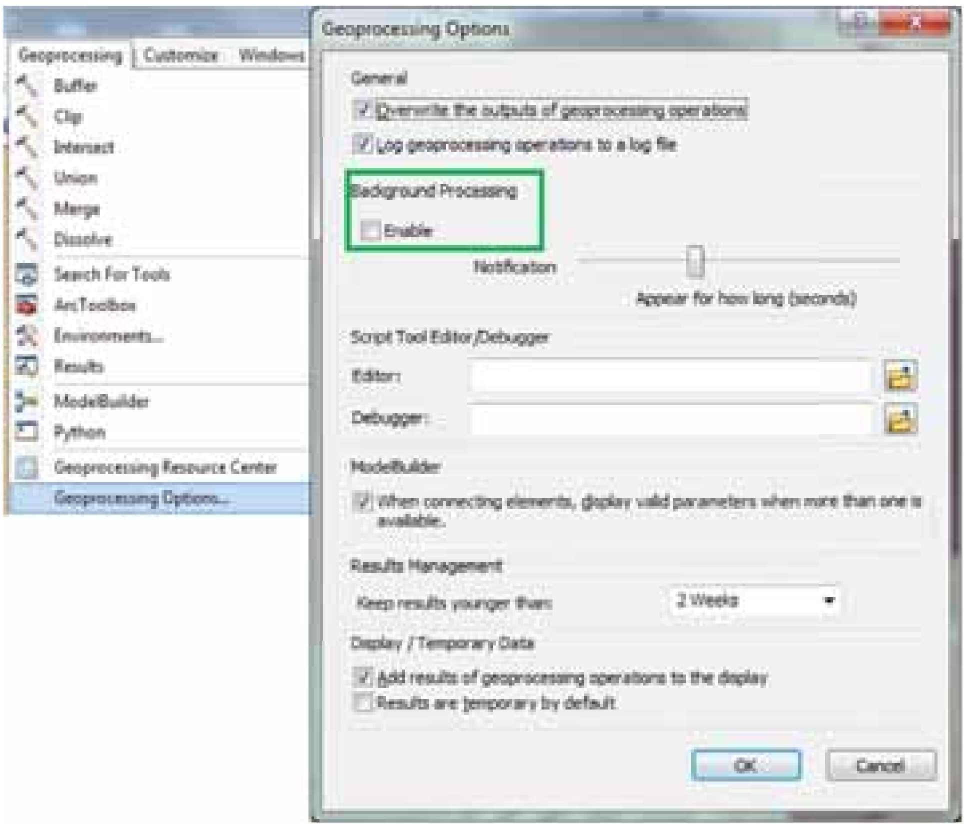


Step 1: Getting started

The steps in this tutorial document assume the tutorial 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.

☐ Open …\RegressionExercise\RegresssionAnalysis911Calls.mxd map document.

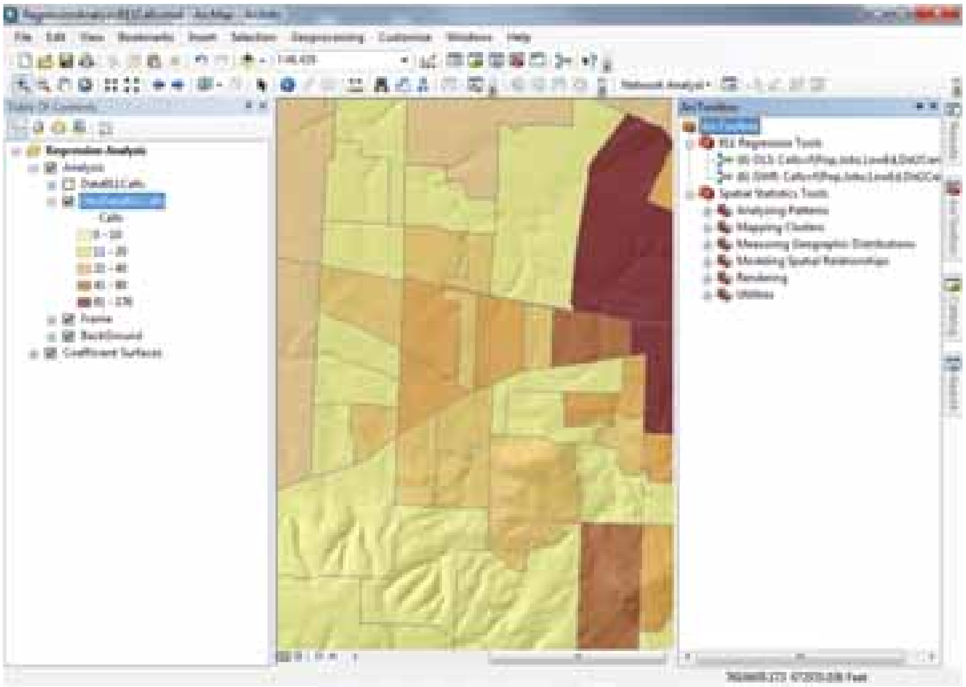
☐ Make sure that Background Processing is NOT enabled (Geoprocessing menu > Geoprocessing Options).



Step 2: Explore OLS Regression

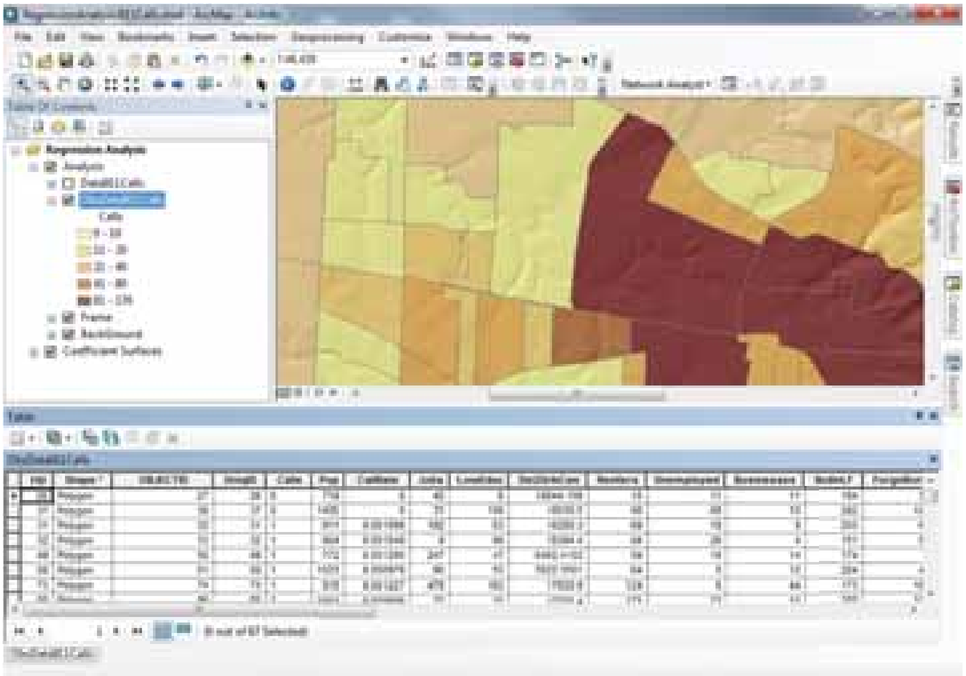
The next question your community is probably asking is, "Why are call volumes so high in those hot spot areas?" and "What are the factors that contribute to high volumes of 911 calls?" To help answer these questions, you'll use the regression tools in ArcGIS.

☐ Expand the 911 Regression tools and Spatial Statistics tools toolboxes.



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

☐ Right-click the ObsData911Calls layer and choose Open Attribute Table.



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

☐ Notice that the table has fields such as Educational status, Unemployment levels, etc.

☐ When done exploring the fields, close the table.

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

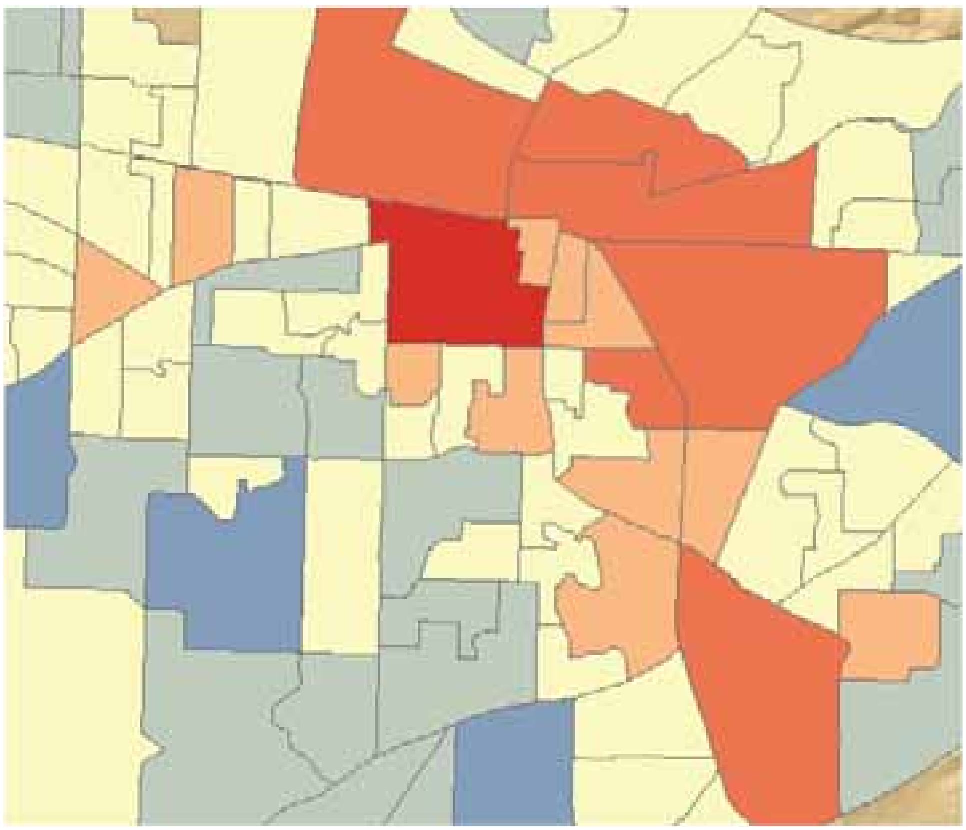
What about population? Would you expect more calls in places with more people? You will test the hypothesis that call volume is simply a function of population. If it is, your community can use Census population projections to estimate future 911 emergency call volumes.

☐ Navigate to the Modeling Spatial Relationships toolset in the Spatial Statistics toolbox, and run the Ordinary Least Squares (OLS) tool with the following parameters:

> **!** **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: …\RegressionExercise\Outputs\OLS911calls
- Dependent Variable: Calls
- Explanatory Variables: Pop

☐ Move the progress window to the side so you can examine the OLS911calls layer in the table of contents.

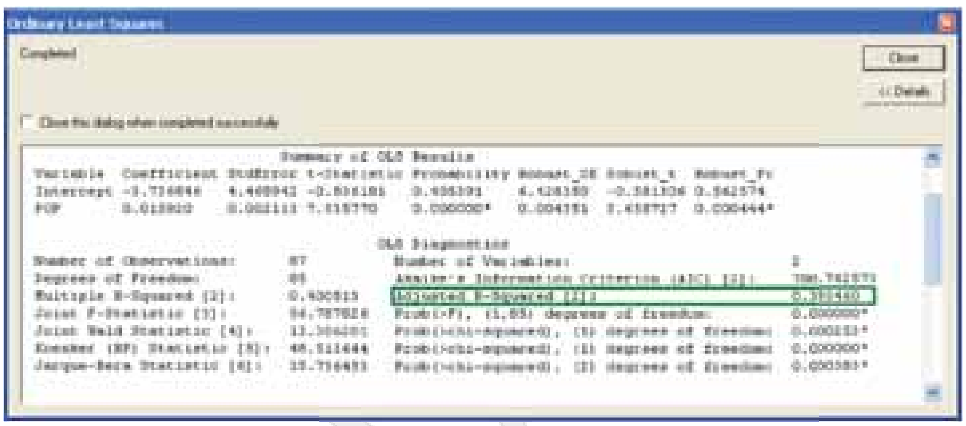


The OLS default output is a map showing you how well the model performed, using only the population variable to explain 911 call volumes. The red areas are under predictions (where the actual number of calls is higher than the model predicted); the blue areas are over predictions (actual call volumes 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.



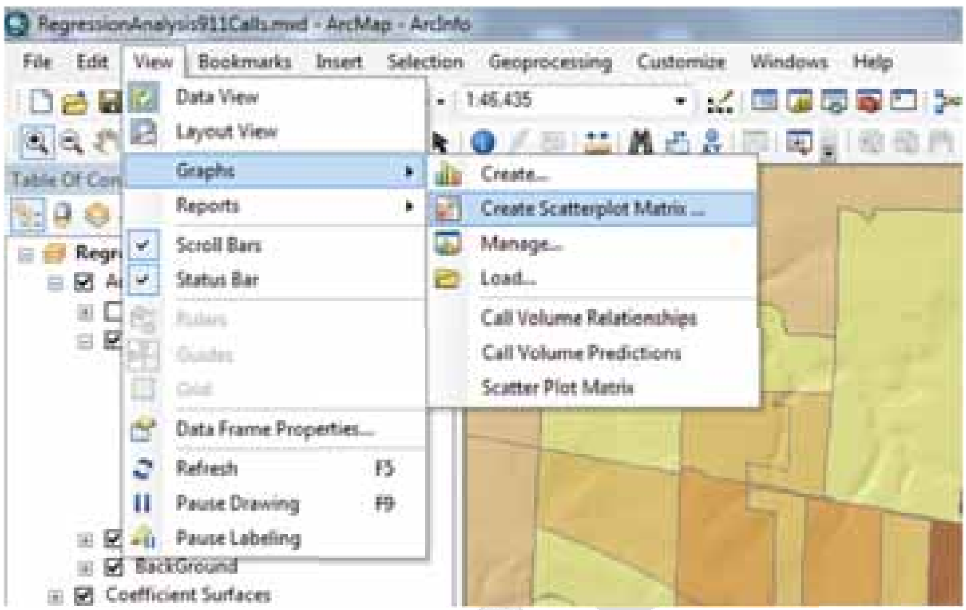
So looking back at the original hypothesis, is call volume simply a function of population? Might your community be able to predict future 911 call volumes from population projections alone? Probably not; if the relationship between population and 911 call volumes had been higher, say 80%, your community 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 the key variables that explain what you are trying to model.

Step 3: Find key variables

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

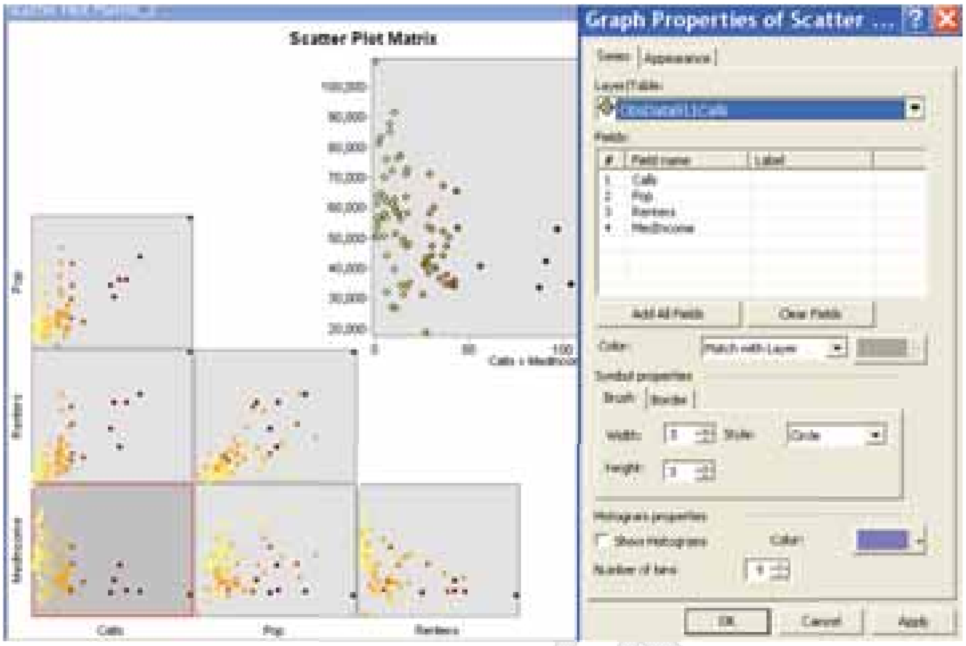
☐ From the View menu, select Graphs and choose Create Scatterplot Matrix (as shown below).



☐ 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.
- Select features in the focus graph and notice how those features are highlighted in each scatterplot and on the map.

☐ Review the following example scatterplot matrix parameter settings (be sure to set the Layer/Table parameter to ObsData911Calls):



☐ Clear selected features, if necessary.

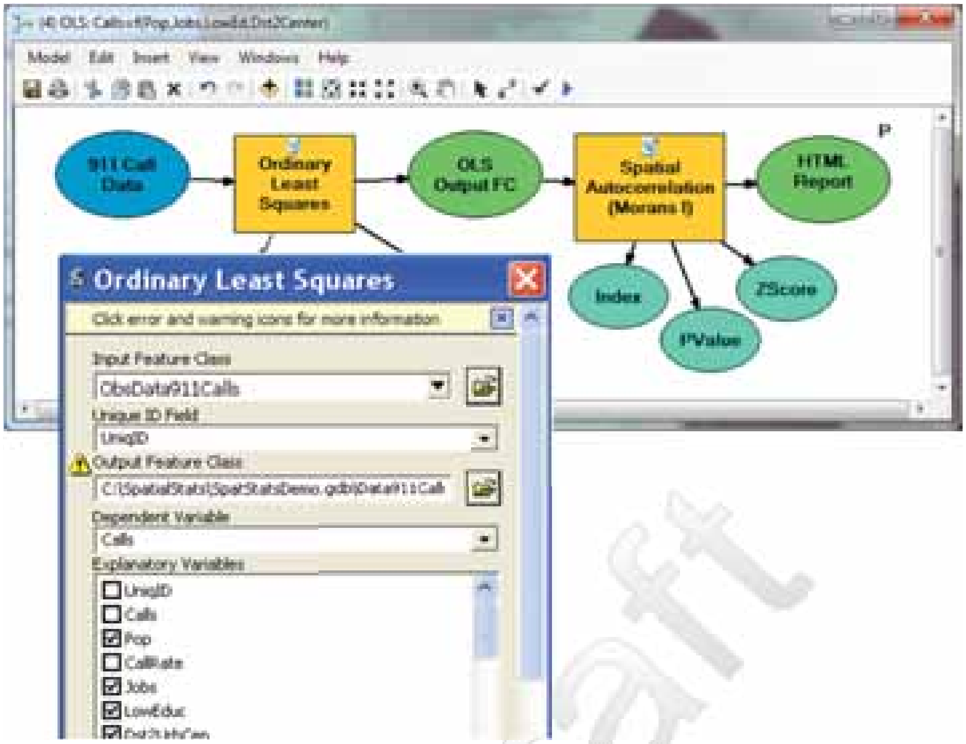
Step 4: A properly-specified model

This time you are going to execute the OLS tool from a model.

☐ In the ArcToolbox window, expand the 911 Regression Tools toolbox and find the model called: OLS: Calls=f(Pop,Jobs,LowEd,Dst2Center).

☐ Right-click this model and select Edit. Notice that there are two tools: OLS and Spatial Autocorrelation.

☐ Double-click the Ordinary Least Squares tool to open it.



☐ Check that the tool has the following parameters set:

- Input Feature Class: Analysis\ObsData911Calls
- Unique ID Field: UniqID
- Output Feature Class: …..\RegressionExercise\Outputs\Data911CallsOLS
- Dependent Variable: Calls
- Explanatory Variables: Pop;Jobs;LowEduc;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. Modeling call volumes as a function of population, jobs, adults who didn't complete high school, and distance to the urban center, produces a properly specified model. You will investigate this.

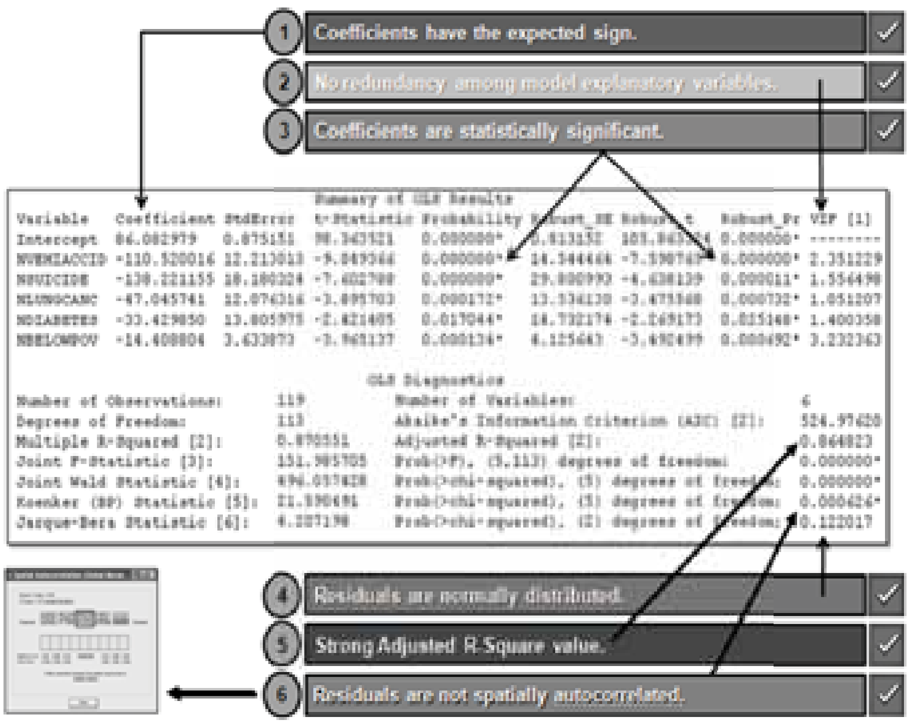
☐ Close the model.

☐ In ArcToolbox, double-click the model to run the model tool (running it this way will ensure that the results from the analysis are in the Results Window).

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.

Step 5: The six things you must check!

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



☐ 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	4.456980	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	8.103407	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]:	0.342321	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. You are expecting a positive coefficient.

If the coefficient for the Population variable was negative, you would not trust your 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.

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	4.456980	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	8.103407	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]:	0.342321	Prob(>chi-squared), (2) degrees of freedom:	0.842602

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>`

☐ 3.) Next, check to see that all 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	0.000053*	3.537930	4.456909	0.000026*	--------
POP	0.005495	0.001468	3.742836	0.000342*	0.001716	3.202300	0.001946*	1.733935
JOBS	0.004042	0.000599	6.778749	0.000000*	0.000814	4.987897	0.000004*	3.174779
LOWEDUC	0.104237	0.012863	8.103407	0.000000*	0.017798	5.856599	0.000000*	1.727065
DSTTURBCEN	-0.001734	0.000272	-6.381896	0.000000*	0.000245	-7.089348	0.000000*	1.135479

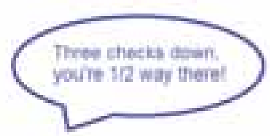
OLS Diagnostics

Number of Observations:	87	Number of Variables:	5
Degrees of Freedom:	82	Akaike's Information Criterion (AICc) [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]:	0.342321	Prob(>chi-squared), (2) degrees of freedom:	0.842802

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 you that a particular variable is critical, you 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!



☐ 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	4.454988	0.000028*	--------
POP	0.005495	0.001468	3.742834	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	8.103407	0.000000*	0.017798	5.856599	0.000000*	1.727065
DST2URBCEN	-0.001734	0.000272	-6.381836	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 (AICc) [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	4.454988	0.000028*	--------
POP	0.005495	0.001468	3.742834	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	8.103407	0.000000*	0.017798	5.856599	0.000000*	1.727065
DST2URBCEN	-0.001734	0.000272	-6.381836	0.000000*	0.000245	-7.089348	0.000000*	1.135479

OLS Diagnostics

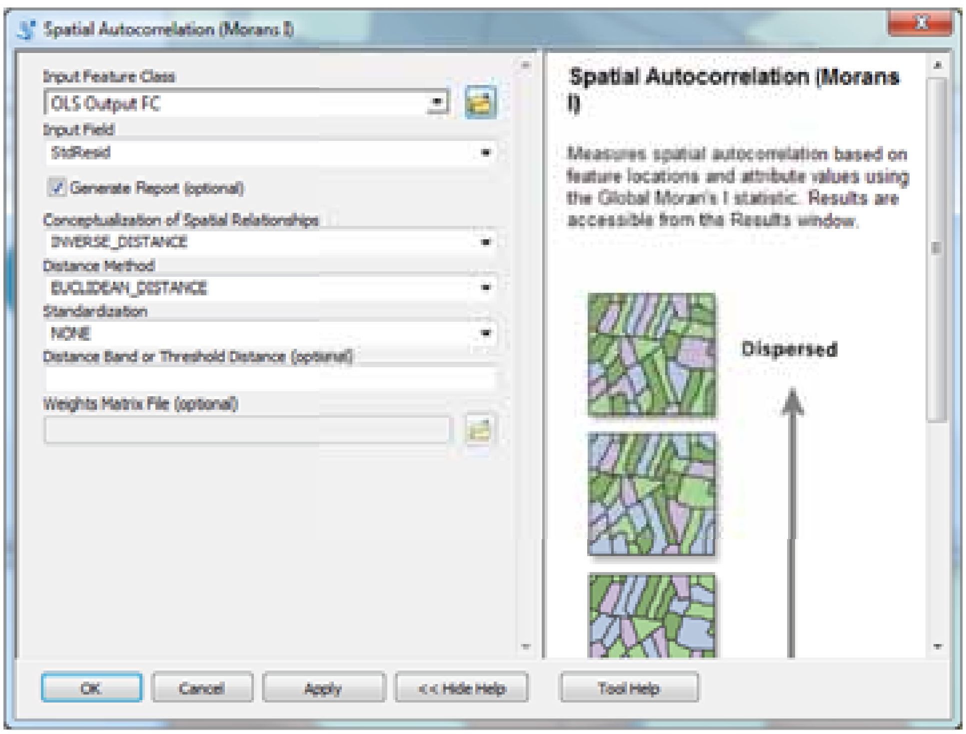
Number of Observations:	87	Number of Variables:	5
Degrees of Freedom:	82	Akaike's Information Criterion (AICc) [2]:	680.4206
Multiple R-Squared [2]:	0.838936	Adjusted R-Squared [2]:	0.831080
Joint F-Statistic [3]:	106.778882	Prob(>F), (4,82) degrees of freedom:	0.000000*
Joint Wald Statistic [4]:	224.669428	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, you are looking for values of 0.5 or higher, but a "good" R2 value depends on what you are modeling. *Self check*: Return to the screenshot of the OLS model that only used Population to explain call volume. What was the Adjusted R2 value? Does the Adjusted R2 value for your new model (four variables) indicate model performance has improved?

The AIC value can also be used to measure model performance. When you have several candidate models (all models must have the same dependent variable), you 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 your new model (4 variables) indicate you improved model performance?

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

- Right-click the OLS: Calls=f(Pop,Jobs,LowEd,Dst2Center) model and choose Edit.
- Double-click the Spatial Autocorrelation tool to see the parameters:

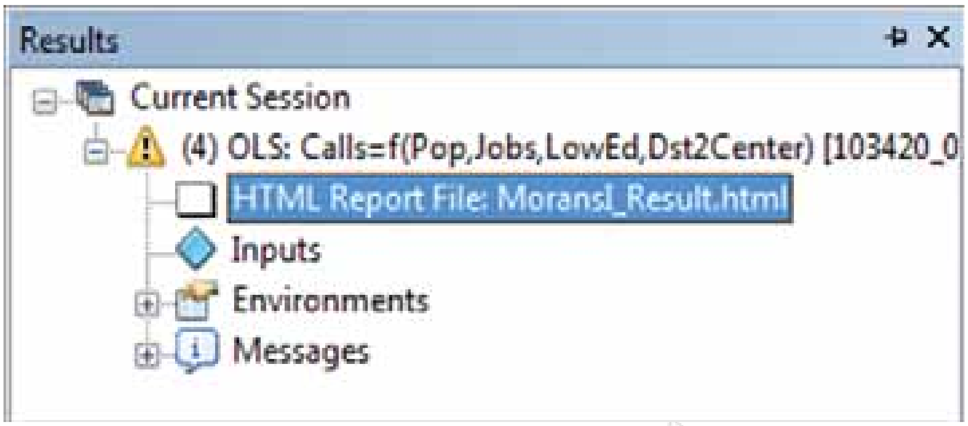


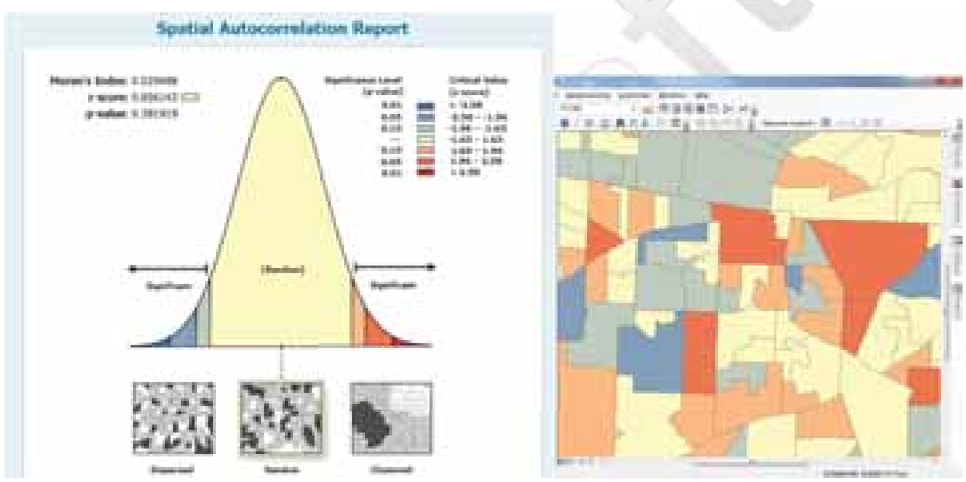
The Output Feature Class from the OLS tool contains a field called Std Resid (Standardized Residuals) that measures the under and over predictions from the model. The Spatial Autocorrelation tool tests to see if the residuals exhibit spatial clustering, or if they reflect random noise (a random pattern).

To see the HTML output from the Spatial Autocorrelation tool, you have to go to the results window.

☐ From ArcMap's Geoprocessing menu, choose Results.

☐ Expand the results, and double-click the HTML output (as indicated below).





The graphic tells you that the residuals (over/under predictions) are random. 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, you are on your way to a properly-specified model. Notice, too, that the map of residuals appears much less clustered than it did with only the Population variable.

☐ Find the Regression Analysis Basics online documentation (search for "Regression Analysis Basics" at http://help.arcgis.com/en/arcgisdesktop/10.0/help), and look for the table called "How Regression Models Go Bad". In this table there are some strategies for how to deal with spatially auto-correlated regression residuals:

Optional work to be completed after class

- 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 six checks covered in this step 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 in the previous graphic to determine the implications and possible solutions.

Step 6: Run the Geographically Weighted Regression tool

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	13.768546	3.693802	4.248920	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	0.004042	0.000599	6.778749	0.000000*	0.000814	4.987897	0.000004*	1.176779
LOWEDUC	0.104237	0.012843	8.103407	0.000000*	0.017798	5.856599	0.000000*	1.727045
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]:	[illegible]	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]:	0.342321	Prob(>chi-squared), (2) degrees of freedom:	0.842802

When the Koenker 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 (GWR).

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):

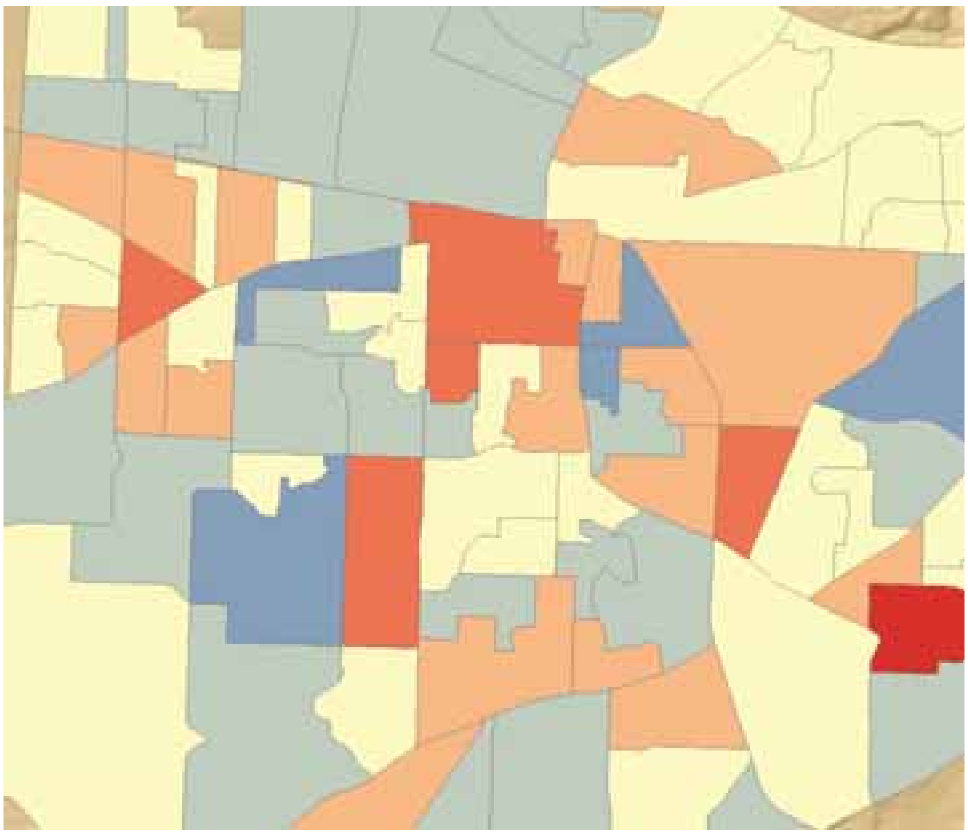
- Input feature class: ObsData911Calls
- Dependent variable: Calls
- Explanatory variables: Pop, Jobs, LowEduc, Dst2UrbCen
- Output feature class: …\RegressionExercise\Outputs\GWRResults
- Kernel type: ADAPTIVE
- Bandwidth method: AICc (you will let the tool find the optimal number of neighbors)

☐ Notice the output from GWR:

```
Neighbors          : 50
ResidualSquares    : 7326.27931715036
EffectiveNumber    : 19.863531396247385
Sigma              : 10.446299891967717
AICc               : 674.6519110481873
R2                 : 0.8957275343805387
R2Adjusted         : 0.8664297924843125
```

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 your 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 three 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.



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

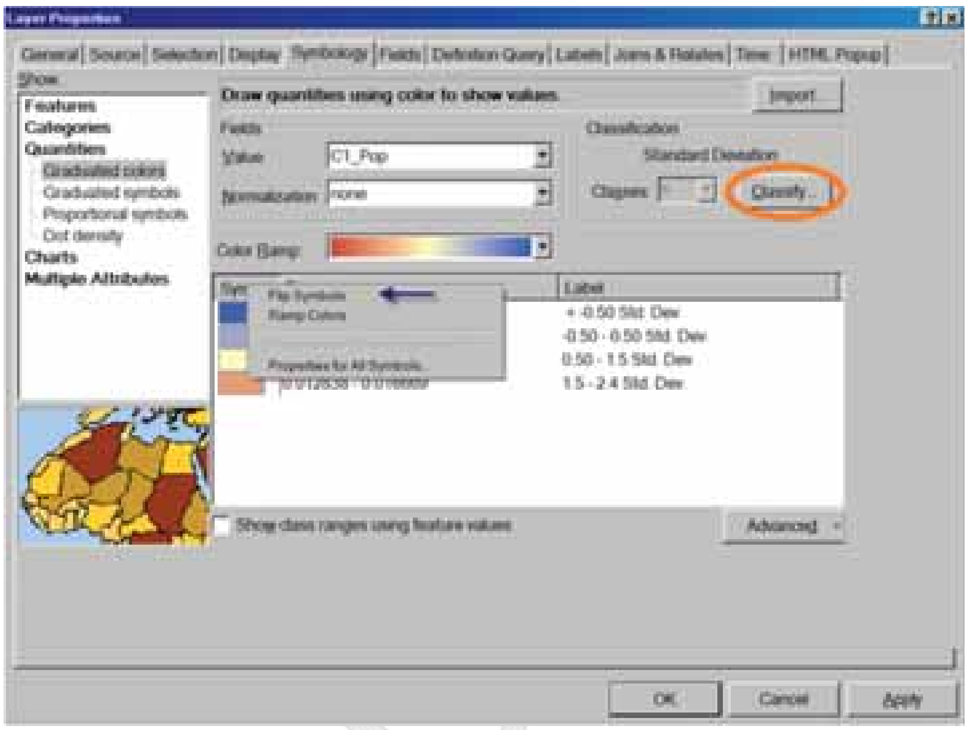
Mapping these coefficients shows you how the relationship between each explanatory variable and the dependent variable changes across the study area.

☐ Right-click ResultsGWR, select Properties.

☐ Click the Symbology tab.

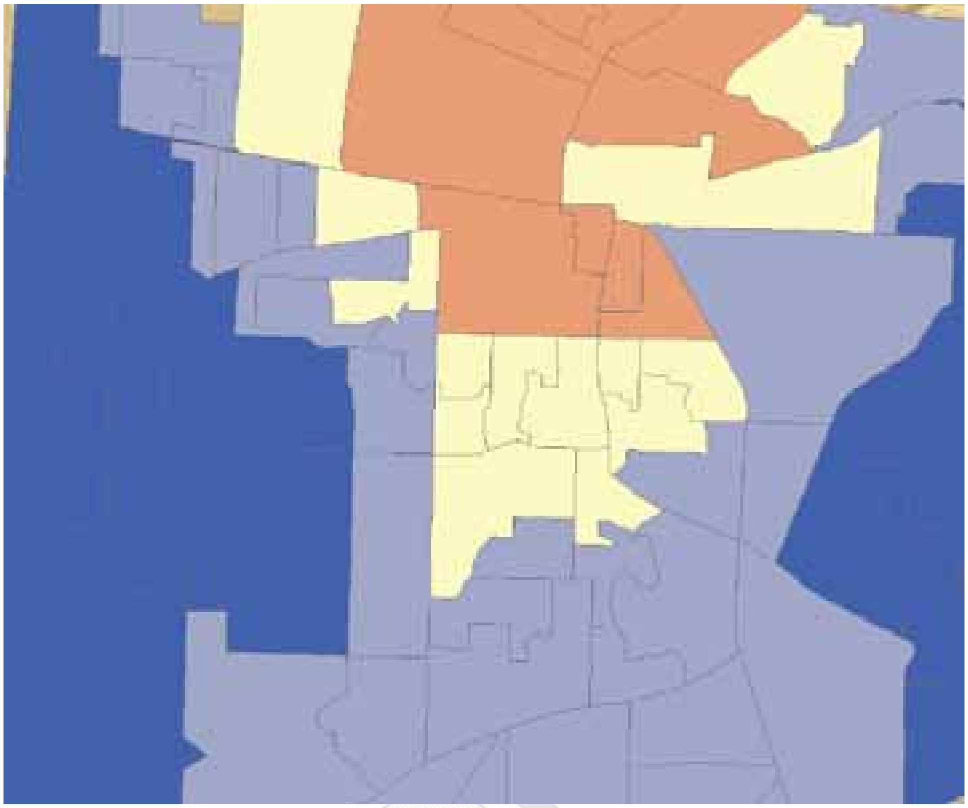
☐ Render the coefficients for C1_Pop, C2_Jobs, and C3_LowEduc.

- Use Standard Deviation class breaks.
- Use a cold-to-hot rendering scheme. You may need to Flip the color ramp.

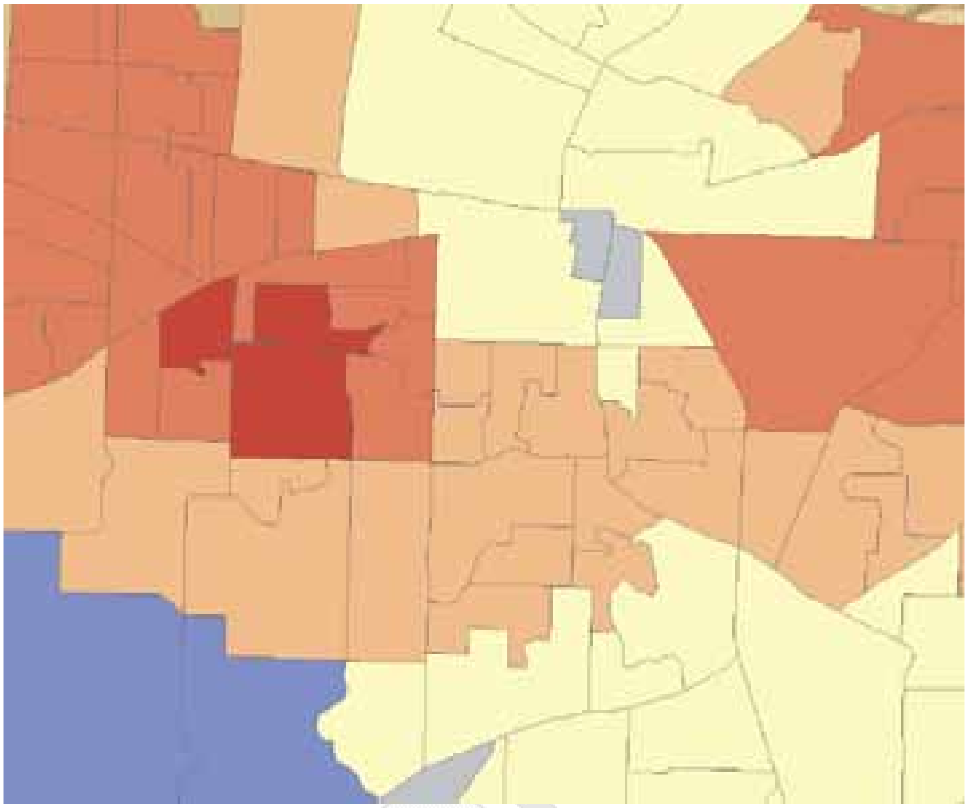


When the relationship is positive (as it is for Pop, Jobs, and LowEduc), the red areas show where the coefficients are large; these are the locations where the explanatory variable you are mapping is a strong predictor.

☐ Review this map of the coefficients for the Population (Pop) variable:



☐ Review this map of the coefficients for Low Education (LowEdu):



In the preceding map, 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. However, 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).

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

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.

> ***Note:*** To learn more about OLS and GWR see "Regression Analysis Basics" in the online help.

