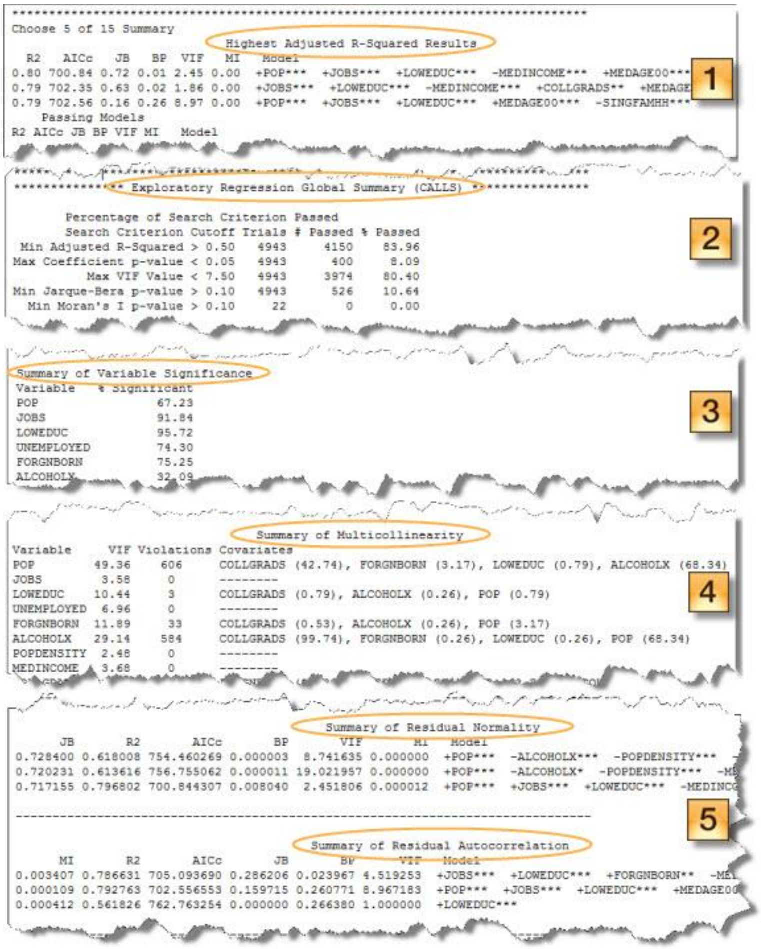
Interpreting Exploratory Regression Results

When you run the Exploratory Regression tool, the primary output is a report. The report can be seen in the geoprocessing messages window when you run the Exploratory Regression tool in the foreground, or found in the location where you chose to save the report file. Optionally, a set of tables will be also created that will help you further investigate the models that have been tested. The purpose of the report is to help you figure out whether or not the candidate explanatory variables you are considering yield any properly specified OLS models. In the event that there are no passing models (models that meet all of the criteria you specified when you launched the Exploratory Regression tool), the output will also help you determine which diagnostics are giving you problems. Strategies for addressing problems associated with each of the diagnostics are given in Regression Analysis Basics (see the table titled "Common regression problems, consequences, and solutions"). For more information about how to determine whether or not you have a properly specified OLS model, please see the documentation on Regression Analysis Basics and Intepreting OLS results.

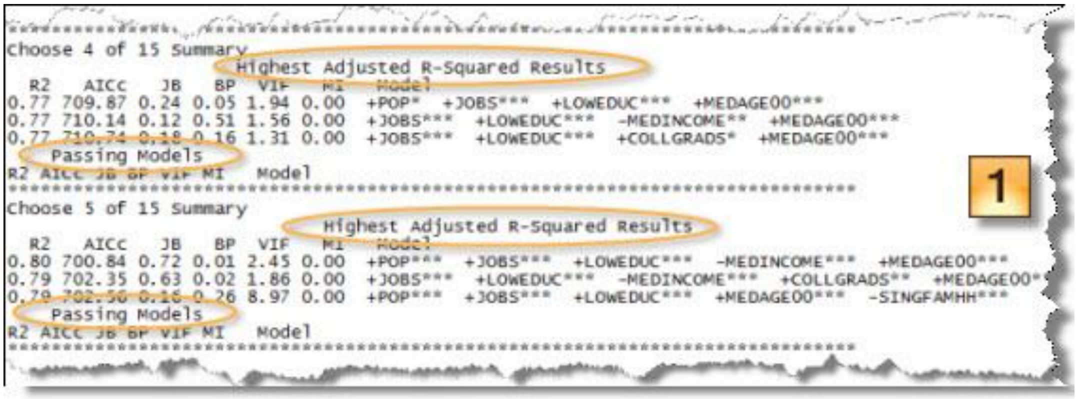
The Report

The Exploratory Regression report has 5 distinct sections. Each section is described below.



```
******************************************************************************
Choose 5 of 15 Summary
                          Highest Adjusted R-Squared Results
  R2   AICc   JB   BP   VIF   MI    Model
0.80 700.84 0.72 0.01 2.45 0.00  +POP***  +JOBS***  +LOWEDUC***  -MEDINCOME***  +MEDAGE00***
0.79 702.35 0.63 0.02 1.86 0.00  +JOBS***  +LOWEDUC***  -MEDINCOME***  +COLLGRADS**  +MEDAGE
0.79 702.56 0.16 0.26 8.97 0.00  +POP***  +JOBS***  +LOWEDUC***  +MEDAGE00***  -SINGFAMHH***
    Passing Models
R2 AICc JB BP VIF MI   Model

*************** Exploratory Regression Global Summary (CALLS) ****************

       Percentage of Search Criterion Passed
       Search Criterion Cutoff Trials # Passed % Passed
 Min Adjusted R-Squared > 0.50   4943     4150    83.96
Max Coefficient p-value < 0.05   4943      400     8.09
          Max VIF Value < 7.50   4943     3974    80.40
Min Jarque-Bera p-value > 0.10   4943      526    10.64
  Min Moran's I p-value > 0.10     22        0     0.00

Summary of Variable Significance
Variable   % Significant
POP                67.23
JOBS               91.84
LOWEDUC            95.72
UNEMPLOYED         74.30
FORGNBORN          75.25
ALCOHOLX           32.09

                          Summary of Multicollinearity
Variable     VIF Violations Covariates
POP        49.36    606     COLLGRADS (42.74), FORGNBORN (3.17), LOWEDUC (0.79), ALCOHOLX (68.34)
JOBS        3.58      0     --------
LOWEDUC    10.44      3     COLLGRADS (0.79), ALCOHOLX (0.26), POP (0.79)
UNEMPLOYED  6.96      0     --------
FORGNBORN  11.89     33     COLLGRADS (0.53), ALCOHOLX (0.26), POP (3.17)
ALCOHOLX   29.14    584     COLLGRADS (99.74), FORGNBORN (0.26), LOWEDUC (0.26), POP (68.34)
POPDENSITY  2.48      0     --------
MEDINCOME   3.68      0     --------

                                     Summary of Residual Normality
     JB        R2       AICc        BP        VIF        MI    Model
0.728400 0.618008 754.460269 0.000003  8.741635 0.000000  +POP***  -ALCOHOLX***  -POPDENSITY***  -
0.720231 0.613616 756.755062 0.000011 19.021957 0.000000  +POP***  -ALCOHOLX*  -POPDENSITY***  -M
0.717155 0.796802 700.844307 0.008040  2.451806 0.000012  +POP***  +JOBS***  +LOWEDUC***  -MEDINC

------------------------------------------------------------------------------

                                   Summary of Residual Autocorrelation
     MI        R2       AICc        JB        BP       VIF    Model
0.003407 0.786631 705.093690 0.286206 0.023967 4.519253  +JOBS***  +LOWEDUC***  +FORGNBORN**  -ME
0.000109 0.792763 702.556553 0.159715 0.260771 8.967183  +POP***  +JOBS***  +LOWEDUC***  +MEDAGE00
0.000412 0.561826 762.763254 0.000000 0.266380 1.000000  +LOWEDUC***
```

1. Best models by number of explanatory variables

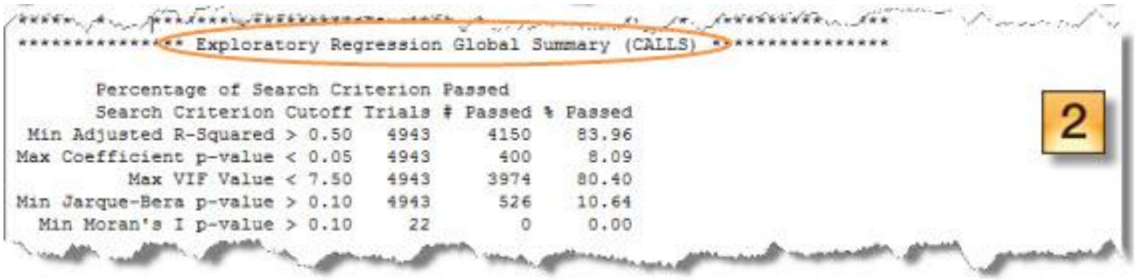


```
********************************************************************************
Choose 4 of 15 Summary
                         Highest Adjusted R-Squared Results
  R2    AICc    JB    BP   VIF    MI    Model
0.77 709.87 0.24 0.05 1.94 0.00  +POP*   +JOBS***   +LOWEDUC***   +MEDAGE00***
0.77 710.14 0.12 0.51 1.56 0.00  +JOBS***   +LOWEDUC***   -MEDINCOME**   +MEDAGE00***
0.77 710.74 0.18 0.16 1.31 0.00  +JOBS***   +LOWEDUC***   +COLLGRADS*   +MEDAGE00***
    Passing Models
R2 AICc JB BP VIF MI    Model
********************************************************************************
Choose 5 of 15 Summary
                               Highest Adjusted R-Squared Results
  R2    AICc    JB    BP   VIF    MI    Model
0.80 700.84 0.72 0.01 2.45 0.00  +POP***   +JOBS***   +LOWEDUC***   -MEDINCOME***   +MEDAGE00***
0.79 702.35 0.63 0.02 1.86 0.00  +JOBS***   +LOWEDUC***   -MEDINCOME***   +COLLGRADS**   +MEDAGE00*
0.79 702.56 0.16 0.26 8.97 0.00  +POP***   +JOBS***   +LOWEDUC***   +MEDAGE00***   -SINGFAMHH***
    Passing Models
R2 AICc JB BP VIF MI    Model
********************************************************************************
```

The first set of summaries in the output report is grouped by the number of explanatory variables in the models tested. If you specify a 1 for the Minimum Number of Explanatory Variables parameter, and a 5 for the Maximum Number of Explanatory Variables parameter, you will have 5 summary sections. Each section lists the 3 models with the highest adjusted R-Squared values, and all passing models. Each summary section also includes the diagnostic values for each model listed: corrected Akaike Information Criteria (AICc), Jarque-Bera p-value (JB), Koenker's studentized Breusch-Pagan p-value (BP), Variance Inflation Factor (VIF), and Global Moran's I p-value (MI). These summaries give you an idea of how well your models are predicting (R2), and if any models pass all of the diagnostic criteria you specified. If you accepted all of the default Search Criteria (Min Adj R-Squared, Max Coefficient p-value, Max VIF value, Min Jarque-Bera p-value, and Min Spatial Autocorrelation p-value parameters), any models included in the Passing Models list are *properly specified OLS models*.

If there aren't any passing models, the rest of the output report will help you make some important decisions about how to move forward.

2. Exploratory Regression Global Summary



```
*************** Exploratory Regression Global Summary (CALLS) ***************

       Percentage of Search Criterion Passed
       Search Criterion Cutoff Trials # Passed % Passed
  Min Adjusted R-Squared > 0.50   4943     4150    83.96
 Max Coefficient p-value < 0.05   4943      400     8.09
          Max VIF Value < 7.50   4943     3974    80.40
 Min Jarque-Bera p-value > 0.10   4943      526    10.64
   Min Moran's I p-value > 0.10     22        0     0.00
```

The Exploratory Regression Global Summary is an important place to start, especially if you have not found any passing models, because it shows you *why* none of the models passed your criteria. The Global Summary lists the 5 diagnostic tests and the percentage of models that passed each of those tests. If you don’t have any passing models, this summary will help you figure out which diagnostic test is giving you trouble.

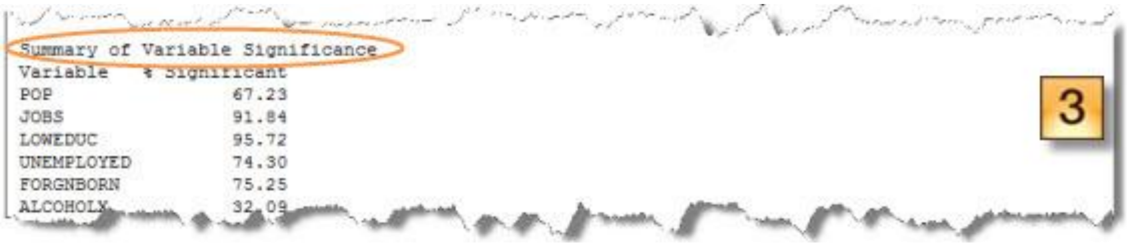
Often the diagnostic giving you problems will be the Global Moran’s I test for Spatial Autocorrelation (MI). When all of the models tested have spatially autocorrelated regression residuals, it most often indicates you are missing key explanatory variables. One of the best ways to find missing variables is to examine the map of the residuals output from the Ordinary Least Squares regression (OLS) tool. Choose one of the exploratory regression models that performed well for all of the other criteria (use the lists of highest adjusted R-Squared values, or select a model from those in the optional output tables), and run OLS using that model. You will then be able to examine the residuals to see if they provide any clues about what might be missing. Try to think of as many spatial variables as you can (distance to major highways, hospitals, or other key geographic features, for example). Consider trying spatial regime variables: if all of your under-predictions are in the rural areas, for example, create a dummy variable to see if it improves your exploratory regression results.

The other diagnostic that is commonly problematic is the Jarque-Bera test for normally distributed residuals. When none of your models pass the Jarque-Bera test, you are having a problem with model bias. Common sources of model bias include:

a) Non-linear relationships
b) Data outliers

Viewing a scatterplot matrix will show you if you have either of these problems. Additional strategies are outlined in Regression Analysis Basics (see the table titled "Common regression problems, consequences, and solutions"). If your models are failing the Spatial Autocorrelation test (MI), fix those issues first. The bias may be the result of missing key explanatory variables.

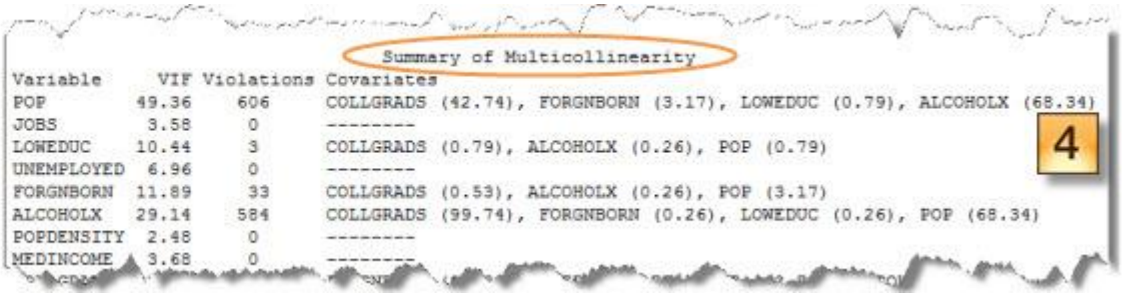
3. Summary of Variable Significance



```
Summary of Variable Significance
Variable    % Significant
POP                 67.23
JOBS                91.84
LOWEDUC             95.72
UNEMPLOYED          74.30
FORGNBORN           75.25
ALCOHOLX            32.09
```

The Summary of Variable Significance is there to help you determine your strongest and weakest explanatory variables. This is especially important when you are working with a lot of candidate explanatory variables (over 50), and want to try models with 5 or more predictors. When you have a large number of explanatory variables and are testing many combinations, the calculations can take a very long time. In some cases the tool won’t finish at all due to memory errors. A good approach is to gradually increase the number of models tested: start by setting both the Minimum and the Maximum Number of Explanatory Variables to 2, then 3, then 4, etc. With each run, remove the variables that are rarely statistically significant in the models tested. This Summary of Variable Significance report will help you find those variables. Even removing 1 candidate explanatory variable from your list can greatly reduce the amount of time it takes for the Exploratory Regression tool to complete.

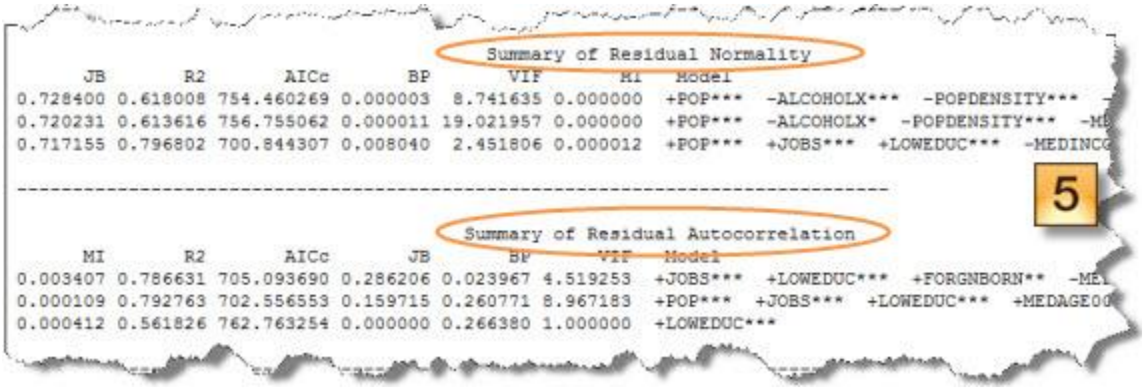
4. Summary of Multicollinearity



```
                                      Summary of Multicollinearity
Variable      VIF Violations Covariates
POP         49.36     606      COLLGRADS (42.74), FORGNBORN (3.17), LOWEDUC (0.79), ALCOHOLX (68.34)
JOBS         3.58       0      --------
LOWEDUC     10.44       3      COLLGRADS (0.79), ALCOHOLX (0.26), POP (0.79)
UNEMPLOYED   6.96       0      --------
FORGNBORN   11.89      33      COLLGRADS (0.53), ALCOHOLX (0.26), POP (3.17)
ALCOHOLX    29.14     584      COLLGRADS (99.74), FORGNBORN (0.26), LOWEDUC (0.26), POP (68.34)
POPDENSITY   2.48       0      --------
MEDINCOME    3.68       0      --------
```

The Summary of Multicollinearity can be used in conjunction with the Summary of Variable Significance to understand which candidate explanatory variables may be removed from your analysis. The Summary of Multicollinearity tells you how many times each explanatory variable was included in a model with high multicollinearity, and the other explanatory variables that were also included in those models. When two (or more) explanatory variables are frequently found together in models with high multicollinearity, it indicates that those variables may be telling the same story. Since you only want to include variables that are explaining a unique aspect of the dependent variable, you may want to choose only one of the redundant variables to include in further analysis. One approach is to use the strongest of the redundant variables based on the Summary of Variable Significance.

5. Additional Diagnostic Summaries

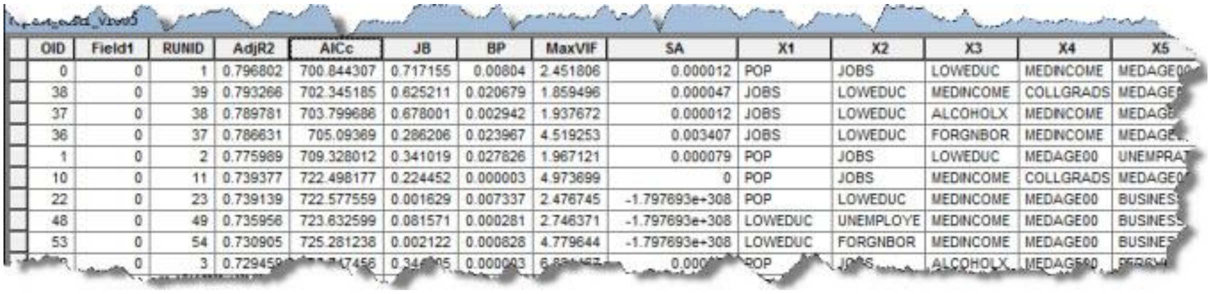


```
                                             Summary of Residual Normality
      JB       R2       AICc       BP        VIF        MI    Model
0.728400 0.618008 754.460269 0.000003  8.741635 0.000000  +POP***  -ALCOHOLX***  -POPDENSITY***  -
0.720231 0.613616 756.755062 0.000011 19.021957 0.000000  +POP***  -ALCOHOLX*  -POPDENSITY***  -M
0.717155 0.796802 700.844307 0.008040  2.451806 0.000012  +POP***  +JOBS***  +LOWEDUC***  -MEDINC

-----------------------------------------------------------------------------------

                                          Summary of Residual Autocorrelation
      MI       R2       AICc       JB       BP       VIF     Model
0.003407 0.786631 705.093690 0.286206 0.023967 4.519253  +JOBS***  +LOWEDUC***  +FORGNBORN**  -ME
0.000109 0.792763 702.556553 0.159715 0.260771 8.967183  +POP***  +JOBS***  +LOWEDUC***  +MEDAGE00
0.000412 0.561826 762.763254 0.000000 0.266380 1.000000  +LOWEDUC***
```

The final diagnostic summaries show the highest Jarque-Bera p-values (Summary of Residual Normality) and the highest Global Moran's I p-values (Summary of Residual Autocorrelation).

These summaries are not especially useful when your models are passing the Jarque-Bera and Global Moran's I test, because if your criterion for statistical significance is 0.1, all models with values above 0.1 are equally passing models. These summaries are useful, however, when you do not have any passing models and you want to see how far you are from having normally distributed residuals or residuals that are free from statistically significant spatially autocorrelation. For instance, if all of the p-values for the Jarque-Bera summary are 0.000000, you are pretty far from having normally distributed residuals. Alternatively, if the p-values are 0.092, then you are very close to having residuals that are normally distributed (in fact, depending on the level of significance that you chose a p-value of 0.092 might be passing). These summaries are there to demonstrate how serious the problem is and, when none of your models are passing, which variables are associated with the models that are at least getting close to passing.

The Tables



OID	Field1	RUNID	AdjR2	AICc	JB	BP	MaxVIF	SA	X1	X2	X3	X4	X5
0	0	1	0.796802	700.844307	0.717155	0.00804	2.451806	0.000012	POP	JOBS	LOWEDUC	MEDINCOME	MEDAGE0
38	0	39	0.793266	702.345185	0.625211	0.020679	1.859496	0.000047	JOBS	LOWEDUC	MEDINCOME	COLLGRADS	MEDAGE
37	0	38	0.789781	703.799686	0.678001	0.002942	1.937672	0.000012	JOBS	LOWEDUC	ALCOHOLX	MEDINCOME	MEDAGE
36	0	37	0.786631	705.09369	0.286206	0.023967	4.519253	0.003407	JOBS	LOWEDUC	FORGNBOR	MEDINCOME	MEDAGE
1	0	2	0.775989	709.328012	0.341019	0.027826	1.967121	0.000079	POP	JOBS	LOWEDUC	MEDAGE00	UNEMPRA
10	0	11	0.739377	722.498177	0.224452	0.000003	4.973699	0	POP	JOBS	MEDINCOME	COLLGRADS	MEDAGE0
22	0	23	0.739139	722.577559	0.001629	0.007337	2.476745	-1.797693e+308	POP	LOWEDUC	MEDINCOME	MEDAGE00	BUSINES
48	0	49	0.735956	723.632599	0.081571	0.000281	2.746371	-1.797693e+308	LOWEDUC	UNEMPLOYE	MEDINCOME	MEDAGE00	BUSINESS
53	0	54	0.730905	725.281238	0.002122	0.000828	4.779644	-1.797693e+308	LOWEDUC	FORGNBOR	MEDINCOME	MEDAGE00	BUSINE
[illegible]	0	3	0.72945	[illegible]	[illegible]	0.000093	[illegible]	[illegible]	POP	[illegible]	ALCOHOLX	[illegible]	[illegible]

The tables include diagnostics for all of the models where all of the explanatory variables met your Max Coefficient p-value and VIF value criteria. Even if you do not have any passing models, there is a good chance that you will have some models in the output tables. One way to evaluate the tables is to open them in ArcMap (which you can do by dragging them from the Catalog window to the Table of Contents). Each row in the table represents a model meeting your criteria for coefficient and VIF values. The columns in the table provide the model diagnostics and explanatory variables. The diagnostics listed are: Adjusted R-Squared (R2), corrected Akaike Information Criteria (AICc), Jarque-Bera p-value (JB), Koenker's studentized Breusch-Pagan p-value (BP), Variance Inflation Factor (VIF), and Global Moran's I p-value (MI). You may want to sort the models by their AICc values. The lower the AICc value, the better the model performed. You can sort the AICc values in ArcMap by double-clicking on the AICc column. If you are choosing a model to use in an OLS analysis (in order to examine the residuals), remember to choose a model with a low AICc value *and* passing values for as many of the other diagnostics as possible. For example, if you have looked at your output report and you know that Jarque-Bera was the diagnostic that gave you trouble, you would look for the model with the lowest AICc value that met all of the criteria except for Jarque-Bera.

Additional Resources

If you're new to regression analysis in ArcGIS, we strongly encourage that you run through the Regression Analysis tutorial before using Exploratory Regression.

- Downloadable tutorial: Regression Analysis in ArcGIS 10

- Additional tool documentation for Exploratory Regression:
  - Learn More about Exploratory Regression
  - What they don't tell you about regression analysis

- Online documentation that explains some basic terminology and best practices for regression analysis
  - Regression Analysis Basics
  - Ordinary Least Squares Regression

- Free Esri Training: Virtual Campus Training Seminar: Regression Analysis Basics

- Burnham, K.P. and D.R. Anderson. 2002. Model Selection and Multimodel Inference: a practical information-theoretic approach, 2nd Edition. New York: Springer. Section 1.5.

- Keep checking back at http://bit.ly/spatialstats for upcoming videos, tutorials and more.