Learn More about Exploratory Regression

Finding a properly specified OLS model can be difficult, especially when there are lots of potential explanatory variables you think might be important contributing factors to the variable you are trying to model (your dependent variable). The Exploratory Regression tool can help. It is a data mining tool that will try all possible combinations of explanatory variables to see which models pass all of the necessary OLS diagnostics. By evaluating all possible combinations of the candidate explanatory variables, you greatly increase your chances of finding the best model to solve your problem or answer your question.

Cautions: The Wiggle Clause

Please be aware that, similar to using methods such as Stepwise Regression (found in traditional statistical packages), using the Exploratory Regression tool in ArcGIS is controversial. While an exaggeration, there are basically two schools of thought on this: the traditional statistician's viewpoint and the data miner's viewpoint.

The traditional statistician would strongly object to exploratory regression. From their perspective, you should formalize your hypotheses before exploring your data to avoid creating models that fit only your data, but don't reflect broader processes. Constructing models that over fit one particular data set may not be relevant to other datasets – sometimes, in fact, even adding new observations will cause an over fit model to become unstable (performance might decrease and/or explanatory variable coefficient significance may wane). When your model isn't robust, even to new observations, it certainly is not getting at the key processes for what you are trying to model. In addition, please realize that regression statistics are based on probability theory and when you run thousands of models, you strongly increase your chances of inappropriately rejecting the null hypothesis (a type 1 statistical error). When you select a 95% confidence level, for example, you are accepting a particular risk; if you could resample your data 100 times, probability indicates that as many as 5 out of those 100 samples will produce false positives. P-values are computed for each coefficient, for example. The null hypothesis is that the coefficient is actually zero and, consequently, the explanatory variable associated with that coefficient is not helping your model. Probability theory indicates that in as many as 5 out of 100 samples, the p-value might be statistically significant only because you just happened to select observations that falsely support that conclusion. When you are only running one model, a 95% confidence level seems conservative. As you increase the number of models you try, you diminish your ability to draw conclusions from your results.

Researchers from the data mining school of thought, on the other hand, would feel it is impossible to know a-priori all of the factors that contribute to any given real world outcome. Often the questions we are trying to answer are complex, and theory on our particular topic may not exist, or might be out of date. Data miners are big proponents of inductive analyses like

those provided by exploratory regression. They encourage thinking outside of the box and using exploratory regression methods for hypothesis development.

We feel that Exploratory Regression, when used with discretion, is a valuable data mining tool that can help you find a properly specified OLS model. Our recommendation is that you always select candidate explanatory regression variables that are supported by theory, guidance from experts, and common sense. Calibrate your regression models using a portion of your data, and validate it with the remainder, or validate your model on additional datasets. If you do plan to draw inferences from your results, at minimum you will want to perform a sensitivity analysis. Extract 9, 99, or 999 random samples from your data and run the OLS model returned by the Exploratory Regression on all of these samples. If you have found a robust model, you will see statistically significant coefficient p-values for all of your samples.

Using the Exploratory Regression tool does have advantages over using other methods that only assess model performance in terms of Adjusted R^2 values. The Exploratory Regression tool evaluates candidate models using a variety of required checks to ensure you have a properly specified OLS model. These checks get not only at model performance, they also assess multicollinearity, model bias, and residual dependency.

Using the Exploratory Regression Tool

When you run the Exploratory Regression tool, you specify a minimum and maximum number of explanatory variables each model should contain, along with threshold criteria for Adjusted R^2, coefficient p-values, VIF values, Jarque-Bera p-values, and spatial autocorrelation p-values. Exploratory Regression runs OLS on every model and assesses each one against your criteria. When it finds a model:

- that exceeds your specified Adjusted R^2 threshold,
- with coefficient p-values, for all explanatory variables, less than you specified,
- with coefficient VIF values, for all explanatory variables, less than your specified threshold and
- returning a Jarque-Bera p-value larger than you specified

it then runs the Spatial Autocorrelation (Global Moran's I) tool on that model's residuals. If the spatial autocorrelation p-value is also larger than you specified in the tool's search criteria, the model is listed as a "passing model". The Exploratory Regression tool will also test regression residuals using the Spatial Autocorrelation tool for models with the 3 highest Adjusted R^2 results. The 3 highest Adjusted R^2 models are saved for each set of models containing the Min Number of Explanatory Variables specified, up to the Max Number of Explanatory Variables specified. If you enter *2* for the Min Number of Explanatory Variables and *4* for the Max Number of Explanatory Variables, for example, Exploratory Regression will report 3 sets of

highest Adjusted R^2 and passing models: for all models with only 2 explanatory variables, for all models with 3 explanatory variables, and for all models with 4 explanatory variables.

Models listed under "Passing Models" meet your specified search criteria. If you take the default values for the Max Coefficient p-value, Max VIF value, Min Jarque-Bera p-value, and Min Spatial Autocorrelation p-value parameters, your passing models will also be properly specified OLS models. A properly specified OLS model has:

- explanatory variables where all of the coefficients are statistically significant,
- coefficients reflecting the expected, or at least justifiable, relationship
- explanatory variables that get at different facets of what you are trying to model (none are redundant so their VIF values are less than about 7.5),
- normally distributed residuals (the Jarque-Bera p-value is *not* statistically significant indicating your model is free from bias)
- randomly distributed over and under predictions (the spatial autocorrelation p-value is *not* statistically significant indicating model residuals are randomly distributed).

When you specify a folder for the optional Output Table Workspace parameter, models that meet the Max VIF Value and for which all explanatory variables meet the Max Coefficient p-value criteria will be written to a table. You can decide if want to just store models with the Max Number of Explanatory Variables, with the Max and Min Number of Explanatory Variables, or to create tables for all models tried. These tables are helpful when you want to examine more than just those models included in the text report file.

When None of the Models Pass

Even when the Exploratory Regression tool hasn't found any passing models, the report it produces is very useful. Most importantly, it can tell you why none of the models are passing and once you know where the difficulties lie, you can often begin to correct the problems. For a full description of the Exploratory Regression tool results see Interpreting Exploratory Regression Results.

Start with the Exploratory Regression Global Summary. It lists the key diagnostics and what proportion of the models tried, passed. If you see 0% for any of these diagnostics, you have identified a definite problem area. Common problem areas involve model bias (all of the models report statistically significant Jarque-Bera p-values), and missing key explanatory variables (all of the report statistically significant spatial autocorrelation p-values). The report sections titled "Summary of Residual Normality" and "Summary of Residual Autocorrelation", at the end of the report, will give you an idea of how close your models came to passing these diagnostics. The 3 models reporting the largest p-values for both the Jarque-Bera and Spatial Autocorrelation diagnostic tests are reported here. For these tests, you do not want statistically significant p-

values, so large p-values are good! P-values larger than 0.10 allow you to go with the null hypothesis, and that's what you are hoping to be able to do. The null hypothesis for the Jarque-Bera test is that the regression residuals are normally distributed; normally distributed residuals are a requirement for a properly specified OLS model. The null hypothesis for the Spatial Autocorrelation test states that model over/under predictions reflect random noise; this too, is an important requirement for a properly specified OLS model. If the largest p-values associated with any of the many models tried is still 0.00000, you have some work to do. Review the table in Regression Analysis Basics called "Common problems, consequences, and solutions" for some strategies.

When Lots of Models Pass

If you accept the search criteria default values, all of the passing models reported are, in fact, properly specified OLS models. Having lots of your models pass is a good position to be in. In fact, you may want to increase your minimum Adjusted R^2 criteria in order to reduce the number of passing models reported. Finding a properly specified OLS model for spatial data is not an easy task at all, so you are definitely in a position to celebrate! Your next task is to choose the best model from the set of all passing models. First consider the AICc values. The models with the lowest AICc values are your best models and differences of even 3 are important enough to choose one model over another. Sometimes you will have several models with very similar AICc values. In that case, you may want to choose the model that includes variables suggesting realistic remediation strategies. If, for example, one model includes poverty and another includes low education, you might want to consider if it is easier to implement economic incentive programs, or programs that would either encourage kids to stay in school or would increase educational opportunities. Sometimes you will have several models with very similar AICc values, but some will have fewer explanatory variables. In this case it is often best, for reasons of parsimony, to go with the model that has fewer explanatory variables.

When the Exploratory Regression tool Takes Forever to Finish AND/OR When the Exploratory Regression tool Crashes with a Memory Error

If you include LOTs (more than 50, for example) candidate explanatory variables and also elect to try models with many (more than 5, for example) variables per model, the Exploratory Regression tool will, without doubt, take a very long time to finish. We have seen cases where it runs for 17 hours and then fails with an out of memory error (always disheartening!). While we have tried our best to minimize the memory requirements (and are still working on this issue), it is unlikely that a tool that takes many days to finish will be helpful, especially since you will likely run the Exploratory Regression tool several times.

The best thing you can do to improve performance (shorten the time the tool takes to finish and avoid out of memory errors) is to reduce the number of candidate explanatory variables you are considering. To do this, run the Explanatory Regression tool with something like "3" for the Max Number of Explanatory variables, then:

1) Look at the Summary of Variable Significance section of the report file. Find variables that are rarely significant in the models tested and remove these from your list of candidate variables.
2) Check the Summary of Multicollinearity section of the report file and look for variables that are redundant. Remove redundant variables, using the Summary of Variable Significant to help you choose which variables to remove.
3) For every candidate explanatory variable, make a table stating the expected coefficient sign (are you expecting a positive or negative relationship?) and the reason you believe that the variable is an important predictor. If you cannot justify a variable (using theory, expert knowledge, or common-sense) – if you cannot clearly explain the expected relationship – remove the variable from your list of candidates.

Once you have whittled your candidate list as small as possible, gradually increase the Max Number of Explanatory Variables.

Another strategy to shorten the amount of time it takes for the Exploratory Regression tool to finish, especially when you have LOTs of features (more than 5000, for example), is to tighten your criteria for passing models. Only models that pass the Adjusted R^2, Jarque-Bera, VIF, and Coefficient significance diagnostics are tested for Spatial Autocorrelation of the residuals. For large datasets, the spatial autocorrelation test will take several minutes to run. If, when you set the Maximum Number of Explanatory Variables to 3, you are getting Adjusted R^2 values near 0.67, for example, you may want to increase the Min Adjusted R-Squared parameter from 0.5 to 0.6. Also, since you want a robust model, you may want to reduce your Max Coefficient p-value threshold from 0.05 to 0.01 requiring passing models (and those tested for spatial autocorrelation) to include explanatory variables where all of the coefficients are statistically significant at the 0.01 level (a 99% confidence level instead of the more lenient default 95% confidence level).

If these strategies don't solve the performance problems, using a Stepwise Regression tool found in a traditional statistical package may be your best option. Keep in mind, however, that the passing models Stepwise or other data mining tools find for you may not meet all of the requirements of OLS (Stepwise Regression tools do not assess all of the required OLS diagnostic checks).

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:
  - Interpreting Exploratory Regression Results
  - 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.