Forensic Geography
Lab: Regression Part 2
Exploratory Techniques

Background

As you may recall, regression analyses allow us to test hypotheses about crime, to make predictions and to identify locations where crime may be more or less common than we would expect given the conditions in the neighborhood.

Luckily, software does a great deal of the calculations on our behalf. Excel, ArcMap, GeoDA and SPSS all have regression tools. Much of the hard work in regression involves preparing data for regression and identifying likely variables to put in a model.

For crime analyses, many of the causal variables have been identified by researchers over a number of years of intense testing. The first thing you should do is to know the research. You should familiarize yourself with the "usual suspects" that drive neighborhood crime.

Still, there are times when you need to figure out *new* variables that are may be causing crime (or some other phenomena), so you'll need to make some decisions on your own about the data so that you may build a useable model. Three major questions you'll have to answer:

1) Which variables should be included in a regression model?
2) How should the data be expressed so that it "plays nicely" within the model and with other variables?
3) How can I avoid including to many variables or variables that conflict with each other?

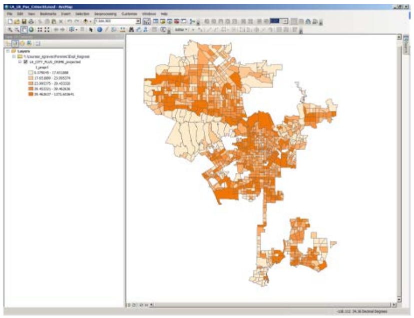
Skills

1. From a large list of variables, you will select a subset of variables that are likely to affect violent crime.
2. You will conduct exploratory regression analysis; and from it identify a "best possible" model.
3. You will transform one or more variables by re-expressing them on a different scale.
   a. You will demonstrate an understanding of the basic principles of regression analysis.
   b. You will demonstrate how to interpret the main "results" produced by a regression analysis.
   c. You will identify and interpret key terms, like, R-squared, confidence, multicollinearity, etc.

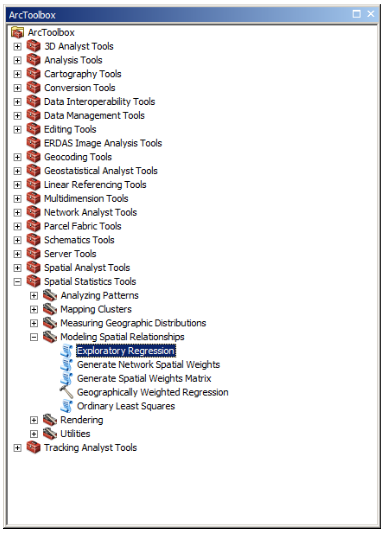
Your Task: Build a viable model of violent crime rates from a list of potential causal variables.

Step 1: Open the Data in ArcMap (and Microsoft Excel if you'd like)

1. The files you need are on the Y drive, under courses_sgraves in the folder Forensic in the subfolder Exp_Regress (Y:\courses_sgraves\Forensic\Expl_Regress)
2. Using ArcMap, open the map project file "LA_LB_PAS_Crime10.mxd" It includes basically only one shape file: "LA_CITY_PLUS_CRIME_projected". These files are maps of Los Angeles, Long Beach and Pasadena. There are loads of variables for you to possibly work with, including a large number of raw demographic variables, plus a long list of crime-related variables.
3. If you'd like to see the crime data just in Excel, including the variable explanations, then open the file "LA_crime_data.xlsx". This might be very useful when you're looking for variable to include in your model.

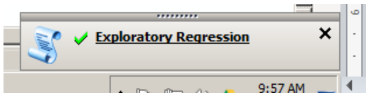


Step 2: Open the Exploratory Regression Tool and Select Promising Variables



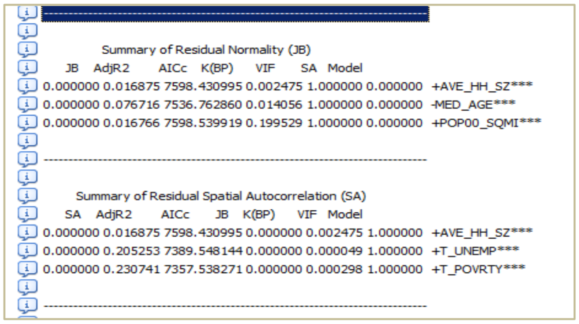
4. Open the Arc Toolbox and from the list of Spatial Statistic Tools, choose, from the Modeling Spatial Relationships and from that list chose “Exploratory Regression”
5. In the Exploratory Regression dialog box, you are presented with a number of input options. For the input feature, use the shape file on the map (remember the drop down menu works only some of the time – sometimes you have to use the open folder technique)
6. Select “t_violrt” for the dependent variable because you are going to try to model / predict the violent crime rate.
7. Chose 10-12 from the list of Candidate Explanatory Variables.
    - Consult the Excel file for the codes / headers
    - Do NOT choose murder rate, rape rate, assaults, etc. Those are PART OF the violent crime rate.
    - Try not to choose redundant variables (like population in 2000 and 2007)…just use one of those, otherwise you’ll get multicollinearity problems where variables explain each other rather than the crime rate.
    - Try to find those that are somewhat unrelated.
    - A few that I would suggest are include: Unemployment rate, Percent female-headed households, Percent high school dropouts, Poverty rate, Percent living in group quarters,
    - Percent of the total population who lives in group quarters, Residential instability, Percent White, Percent young males aged 15-24.
    - Pick a few others.
8. If you want a nicer output report file, you can include those. For 10.1 users, the “Output Report File” is very nice.
9. You may want to set the maximum number of Explantory Values (under Search Criteria) to 8…by default it is 5
10. The rest of the default settings are probably OK, but feel free to experiment with them.
11. Click OK.

Step 3: Interpret your results!



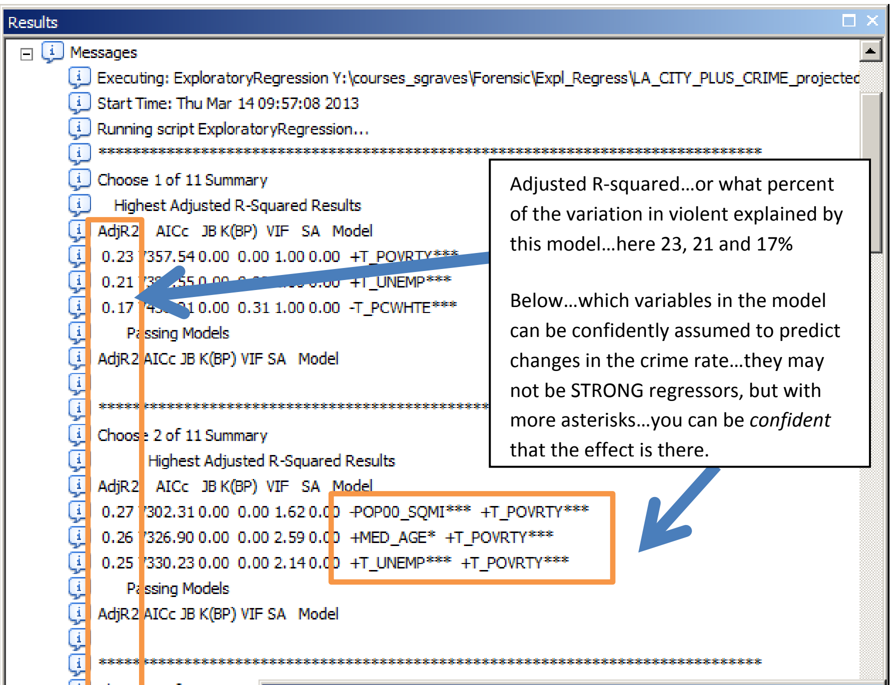
12. Hopefully it will run well, then you’ll need to click on the little “Exploratory Regression” box that will pop up in the lower right corner of your screen when the process has run.. Otherwise, you need to click on “Geoprocessing” from the top tool ribbon and from the drop down menu select “results”
13. You should see a “Results” Window. At the top you’ll see some ordinary stuff about your inputs and the processing time, etc. Below that are the results of the various models that the software tested, using different combinations of the variables that you chose.
14. At the top of the results window are models that used few or even just one variable to explain violent crime rates. Look first for the Adjusted R2 score. If it’s not above .50 then the model variables aren’t doing a great job of explaining violent crime rate…less than 50% accurate (strength). This area also, though, points out which variables that you can be confident are predicting some part of the variation in crime rate across the map. Make a note of those. (see image on next page)

15. Examine all the model possibilities
16. Scroll to the bottom until you see "Exploratory Regression Global Summary"
17. Check to see if any of the trials (models) passed the 50% test. (my first try didn't)
18. You can also check out the percent that passed the VIF test, the Jacque-Bera, and the Spatial Autocorrelation Test (mixed results for me)
19. Having trouble understanding all this? The HELP files, especially in 10.1 are pretty good. Press F1 and investigate.
20. You might also want to look at the Residual Normality and the Spatial Autocorrelation issues. You want to avoid variables that don't show "normalcy" for residuals. Those that are normal are those have a normal curve of "hits and misses" while predicting the violent crime rate (the JB score above is below .01; so throw those out).



```
----------------------------------------------------------------------

          Summary of Residual Normality (JB)
      JB    AdjR2      AICc    K(BP)     VIF      SA   Model
0.000000 0.016875 7598.430995 0.002475 1.000000 0.000000  +AVE_HH_SZ***
0.000000 0.076716 7536.762860 0.014056 1.000000 0.000000  -MED_AGE***
0.000000 0.016766 7598.539919 0.199529 1.000000 0.000000  +POP00_SQMI***

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

        Summary of Residual Spatial Autocorrelation (SA)
      SA    AdjR2      AICc      JB    K(BP)     VIF   Model
0.000000 0.016875 7598.430995 0.000000 0.002475 1.000000  +AVE_HH_SZ***
0.000000 0.205253 7389.548144 0.000000 0.000049 1.000000  +T_UNEMP***
0.000000 0.230741 7357.538271 0.000000 0.000298 1.000000  +T_POVRTY***

----------------------------------------------------------------------
```

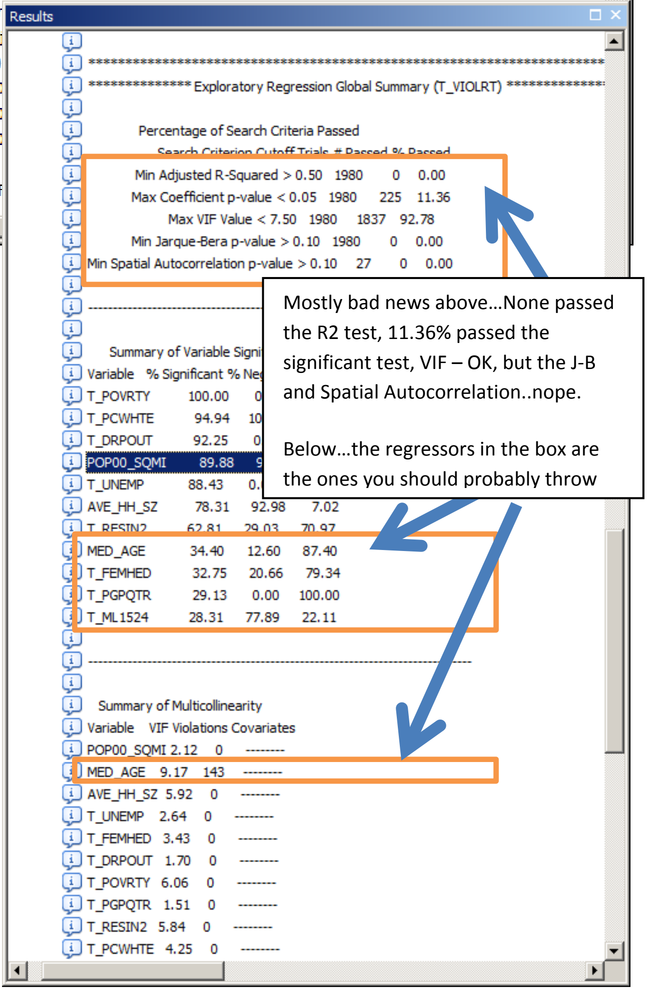
21. Spatial Autocorrelation is also not desirable because it's an indication that the effects of these variables are "spilling over" into neighboring census tracts. That can be fixed, but in the short term they should be noted



```
Messages
  Executing: ExploratoryRegression Y:\courses_sgraves\Forensic\Expl_Regress\LA_CITY_PLUS_CRIME_projectec
  Start Time: Thu Mar 14 09:57:08 2013
  Running script ExploratoryRegression...
  **********************************************************************
  Choose 1 of 11 Summary
     Highest Adjusted R-Squared Results
  AdjR2    AICc   JB K(BP)  VIF   SA   Model
   0.23 7357.54 0.00  0.00 1.00 0.00  +T_POVRTY***
   0.21 7389.55 0.00  0.00 1.00 0.00  +T_UNEMP***
   0.17 [illegible] 0.00  0.31 1.00 0.00  -T_PCWHTE***
        Passing Models
  AdjR2 AICc JB K(BP) VIF SA   Model

  **********************************************************************
  Choose 2 of 11 Summary
           Highest Adjusted R-Squared Results
  AdjR2    AICc   JB K(BP)  VIF   SA   Model
   0.27 7302.31 0.00  0.00 1.62 0.00  -POP00_SQMI***  +T_POVRTY***
   0.26 7326.90 0.00  0.00 2.59 0.00  +MED_AGE*  +T_POVRTY***
   0.25 7330.23 0.00  0.00 2.14 0.00  +T_UNEMP***  +T_POVRTY***
        Passing Models
  AdjR2 AICc JB K(BP) VIF SA   Model

  **********************************************************************
  Choose 3 of 11 Summar
               Highest Adj
  AdjR2    AICc   JB K(BP)
   0.30 7262.86 0.00  0.0
   0.30 7268.60 0.00  0.0
   0.29 7280.45 0.00  0.0
        Passing Models
  AdjR2 AICc JB K(BP) VIF
```

Adjusted R-squared…or what percent of the variation in violent explained by this model…here 23, 21 and 17%

Below…which variables in the model can be confidently assumed to predict changes in the crime rate…they may not be STRONG regressors, but with more asterisks…you can be *confident* that the effect is there.



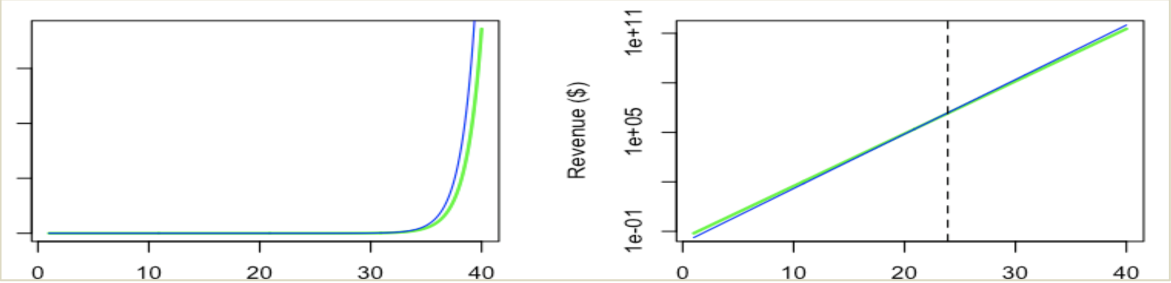
```
**********************************************************************
************** Exploratory Regression Global Summary (T_VIOLRT) **************

       Percentage of Search Criteria Passed
         Search Criterion Cutoff Trials # Passed % Passed
         Min Adjusted R-Squared > 0.50   1980      0    0.00
        Max Coefficient p-value < 0.05   1980    225   11.36
              Max VIF Value < 7.50   1980   1837   92.78
        Min Jarque-Bera p-value > 0.10   1980      0    0.00
Min Spatial Autocorrelation p-value > 0.10     27      0    0.00

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

     Summary of Variable Signi
Variable    % Significant % Ne
T_POVRTY       100.00      0
T_PCWHTE        94.94     10
T_DRPOUT        92.25      0
POP00_SQMI      89.88      9
T_UNEMP         88.43     0.
AVE_HH_SZ       78.31    92.98     7.02
T_RESIN2        62.81    29.03    70.97
MED_AGE         34.40    12.60    87.40
T_FEMHED        32.75    20.66    79.34
T_PGPQTR        29.13     0.00   100.00
T_ML1524        28.31    77.89    22.11

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

   Summary of Multicollinearity
Variable   VIF Violations Covariates
POP00_SQMI 2.12    0    --------
MED_AGE    9.17  143    --------
AVE_HH_SZ  5.92    0    --------
T_UNEMP    2.64    0    --------
T_FEMHED   3.43    0    --------
T_DRPOUT   1.70    0    --------
T_POVRTY   6.06    0    --------
T_PGPQTR   1.51    0    --------
T_RESIN2   5.84    0    --------
T_PCWHTE   4.25    0    --------
```

Mostly bad news above…None passed the R2 test, 11.36% passed the significant test, VIF – OK, but the J-B and Spatial Autocorrelation..nope.

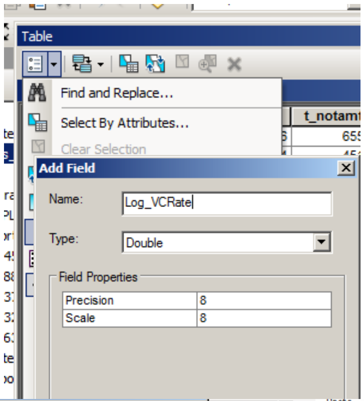
Below…the regressors in the box are the ones you should probably throw

Step 4: Try to Fix Your Data by Logging it.

22. One of the big problems in our data is that there isn't a nice linear increase in violent crime rate. Some census tracts have crime rates that are really way out there, so they have a curve like you see on the left (above). Still some of the variables (like poverty rates or female headed households) increase at a rate like the graph on the right (above). One way to get these two types of rates on to play nicely with each other is to put the skewed (above left) data on a logarithmic scale. Re-expressing a variable in a logarithmic fashion frequently helps the model work better, without having to throw out data.



23. How to put data on a log scale? You may need to "export" the Crime Map to your own drive. The provided map is in a "read only drive". Right click on the layer (crime map in the table of contents); choose Data, and chose Data Export. This is the same as "save as" with many other software packages. Add the new layer to your map.
24. Open the attribute table. Click on the icon in the upper left corner. From the drop down menu select "Add Field".
25. In the Add Field dialog box, name your new field something like "Log_VCRate" for log of Violent Crime Rate. Select Double as the type and enter 8 and 8 for precision and scale. Click OK.
26. Scoll to the far right of your attribute table and find the new column of data.
27. Right click on the header "Log_VCRate" and from the drop down menu select "Field Calculator"
28. Click Yes on the nag screen.
29. In the field calculator dialog box, it is prompting you to do a calculation. The calculation you want to do is to derive the log value for violent crime rate, so click on "Log( )" from the list of Functions; then note the cursor is in the middle of the parentheses in the formula window below.
30. Double click on t_violrt so that the formula window shows "Log ([t_violrt])" . Click OK. You should see your new column fill up with numbers.

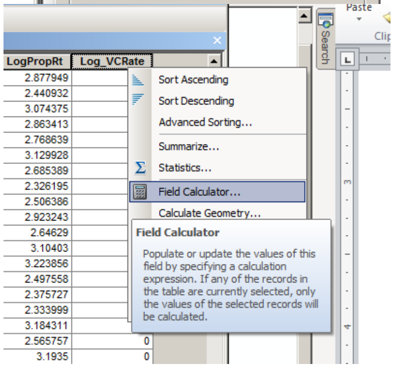


Name: Log_VCRate

Type: Double

Field Properties

Precision	8
Scale	8



LogPropRt	Log_VCRate
2.877949	
2.440932	
3.074375	
2.863413	
2.768639	
3.129928	
2.685389	
2.326195	
2.506386	
2.923243	
2.64629	
3.10403	
3.223856	
2.497558	
2.375727	
2.333999	
3.184311	
2.565757	0
3.1935	0

Field Calculator

Populate or update the values of this field by specifying a calculation expression. If any of the records in the table are currently selected, only the values of the selected records will be calculated.

Step 5: Run Exploratory Regession Again…with your new Logged version of the Violent Crime Rate.

31. Make a list of the variables you are using.
32. Run it a few times if you need. Get rid of variables that aren't working.
33. You might want to log a few of the predictor variables (regressors).
34. Add a few different ones.
35. See if you can get above 55% on the entire model adjusted R squared.
36. For credit capture a screen shot (Ctrl+Alt+Print Screen) of your best model results window; but mostly just the "Global Summary" part at the bottom.
37. Paste the screen capture (Ctrl V) into a word document and write a paragraph explaining what the results mean. For extra credit, run Ordinary Least Squares regression with the variables you discovered and paste that map into the documents with a paragraph explaining it. Email the word file as an attachment.