Forensic Geography
Lab: Regression Part 1
Payday Lending and Crime
Seattle, Washington

Background

Regression analyses are in many ways the "Gold Standard" among analytic techniques for undergraduates (and for the rest of us). To do them very well, takes some measure of competency with statistics and the many rules that govern stats. However, if one waits until one knows all the possible violations of statistical laws; then one would rarely try to use these valuable tools. We don't learn a language by learning all the mistakes first, so we'll apply that strategy in this class.

In essence, regression analyses is statistical modeling of a phenomena to understand why it happens. We do this by making educated guesses about the causes that create an effect that we wish to understand. The reasons why something happens, or causal variables are the "independent variables", also "X" variables. The thing that is affected, or outcome is the "dependent variable" or "Y" variable…for which you seek the answer to "Why".

In this class our "Y – dependent - effect" variable is generally some sort of crime rate, and the "X – causal-independent" variables are those things that we think are root causes of crime, such as age, income, ethnicity, residential stability, etc.

A regression analysis involves the building of a "model". The model is basically a list of X and Y variables in which the X variables are supposed to do a good job of predicting the Y variable. How well the model predicts the Y variable is the "strength" of the model. Of course no model can accurately predict crime rates 100%, but good models do a pretty good job of it. When that happens, then the model is can be used to identify locations where crime is too high (according to the model) or unusually low.

Skills

1. You will run a regression analysis or two.
2. You will demonstrate an understanding of the basic principles of regression analysis.
3. You will demonstrate how to interpret the basic "results" produced by a regression analysis.
4. You will identify and interpret key terms, like outlier, R-squared, confidence, etc.
5. You will make a map of regression residuals to identify locations that have more/less crime that the model predicts.

Your Task: Find a ZIP code in Seattle where the crime rate is worse than you would expect…and better than you would expect.

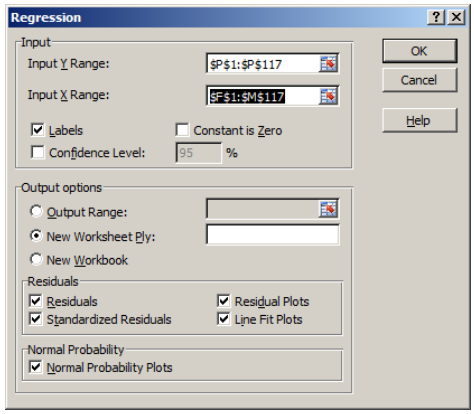
Step 1: Open the Data in Microsoft Excel

1. Open the file "SEATTLE_CRIME_PD_homework.xls" using Excel. It's on the Y drive in the Regression Folder: (Forensic\Regression\).
2. Examine the data. Note that there is a DATA tab, containing data already prepared for you to use in this exercise; a "Variable Names" tab, that has a key to understanding the column headers on the data page
3. The columns in **Green** represent the independent, causal variables. The variables that are in other colors and begin with LNT are the DEPENDENT or Y / effect variables…the crime rates.
4. Note that most of the data has been put on a logarithmic scale to even out the skewness that was in the data. You'll learn something about that later.

Step 2: Run the Regression Tool in Excel

5. Select the "Homework" tab to find the data that you need to use for this assignment.
6. Click on the Data tab above the tool ribbon.

7. Click on Data Analysis tool in the tool ribbon. If it's not there, refer to the exercise on correlation to see how to add this "add in" tool to Excel.
8. From the list of Data Analysis options, select "Regression" and click "OK".
9. The Regression tool dialog box will appear. For the dependent (Y) variable range, highlight the column with LNPROPRT_06_07 in it. That column contains the logarithmic version of the property crime rate in 2006-2007 by census tracts.
10. For Input Range X, select all the data in the green columns, plus the antecedent Property Crime rate from 1999-2001, "LNT_PROPRT". Be sure to include the column headers.
11. Check off all the boxes for Residuals and Residual Plots…if you want.
12. Use "New Worksheet Ply"
13. Click OK…and examine your results or the "Summary Output" in the new worksheet.
14. Examine the new worksheet that was created. The "in class version" is below. Yours is the "homework version" and it is different than the sample below.
15. Read and interpret the results using the image below as a guide.



1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.938329517							
5	R Square	0.880462282							
6	Adjusted R Square	0.871524883							
7	Standard Error	0.410959781							
8	Observations	116							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	8	133.1031144	16.6378893	98.51437076	8.67928E-46	0.000000000000000		
13	Residual	107	18.07100975	0.168887942					
14	Total	115	151.1741241						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	0.427007859	0.271797235	1.571052989	0.119122835	-0.11179846	0.965814177	-0.1118	0.965814
18	T_POP00	-1.13586E-05	2.20552E-05	-0.515010401	0.607608051	-5.50804E-05	3.23631E-05	-5.5E-05	3.24E-05
19	T_FEMHED	0.00392009	0.012128996	0.323199859	0.747174667	-0.02012423	0.02796441	-0.02012	0.027964
20	T_RESIN2	0.082054152	0.081477611	1.007076068	0.31617022	-0.079465709	0.243574014	-0.07947	0.243574
21	LNCOUNT_05	0.061244002	0.020167424	3.036778561	0.003003876	0.021264435	0.101223568	0.021264	0.101224
22	T_DISAD3	0.190949356	0.07109404	2.685870091	0.008387752	0.050013713	0.331884998	0.050014	0.331885
23	CBD	0.239059496	0.190972867	1.251798227	0.213372993	-0.139521939	0.617640931	-0.13952	0.617641
24	LNT_ML1524	0.049323703	0.113724301	0.433712958	0.665370085	-0.17612146	0.274768865	-0.17612	0.274769
25	LNT_VIOLRT	0.707955153	0.081002076	8.739963096	3.57119E-14	0.547377985	0.868532321	0.547378	0.868532
26									
27									

Strength/Accuracy of the model

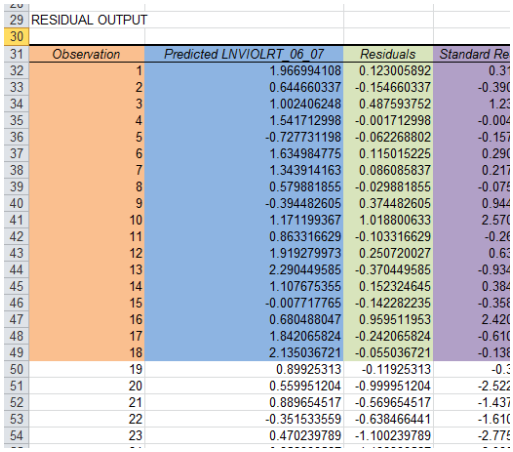
relationship is due to random chance

Variables **Strength** Chance of randomness

16. To quickly read this summary here's what you need to focus upon:
    - GREEN – The R squared and the adjusted R-squared indicate the predictive accuracy and/or power of the model. This one is at .88/.87. So, you would say that about 87% of the variability of the crime in Seattle is explained by this model….or by the variables in the model.
    - ORANGE – Significance or How likely is it that the relationship found by the model is due to just random chance?....In this case, you would say "Not likely. It is 8.67 *10 $^{-46}$ …or .0000 (46 zeroes) 867…in other words. Only a super, super tiny chance that this is random. That number is low because the strength is good and we have large N of 115 census tracts)
    - BLUE – Variables: These are the names of eXplanatory variable in the model.
    - PINK – These are in this instant roughly the change in crime one could expect with a unit change in the one of these variables. Because many of the variables in this model are on a logged scale and using a rate per 1,000 persons in each census tract, it's very difficult to easily translate this easily. However, you can think of it like this for LN_COUNT_05: When there is an extra .06 payday lenders, then the Violent Crime Rate (our dependent variable) tends to go up by 1 unit. When the scale is normal (not logged) and the variables are integers, this is easier to interpret.
    - BEIGE AND BROWN: Easier to read…again these answer the question, "How sure of this are you?" As the t-Statistic goes beyond --/+ 1.96, the chances for randomness to be at play with this particular variable's effect on the crime rate is less than 5%. If you want the exact percentage chance of randomness…look at the P Value. The smaller the P-value, the smaller the chance that

randomness can explain the relationship. So several of these variables (red square box) have VERY small possibility that they have a random relationship with the crime rate.

17. Read and understand the Residual Output. Here's How:

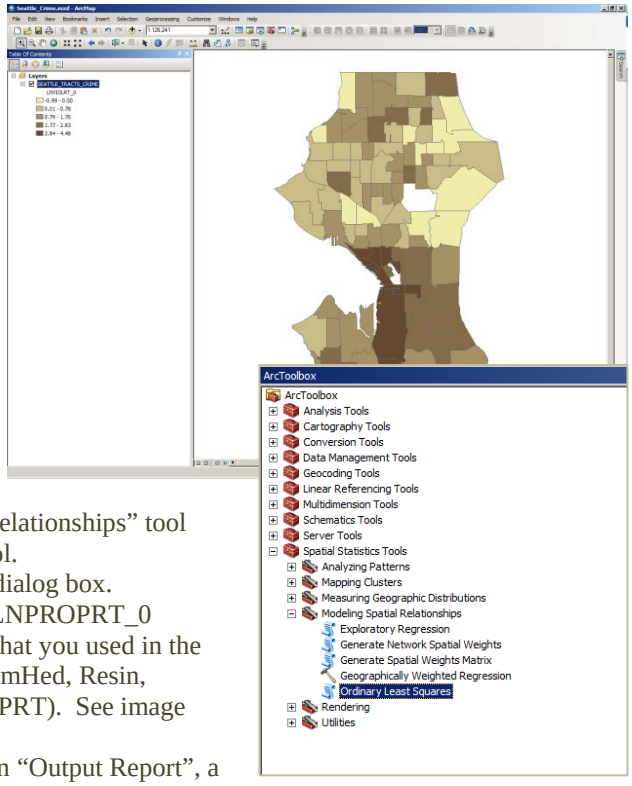


RESIDUAL OUTPUT

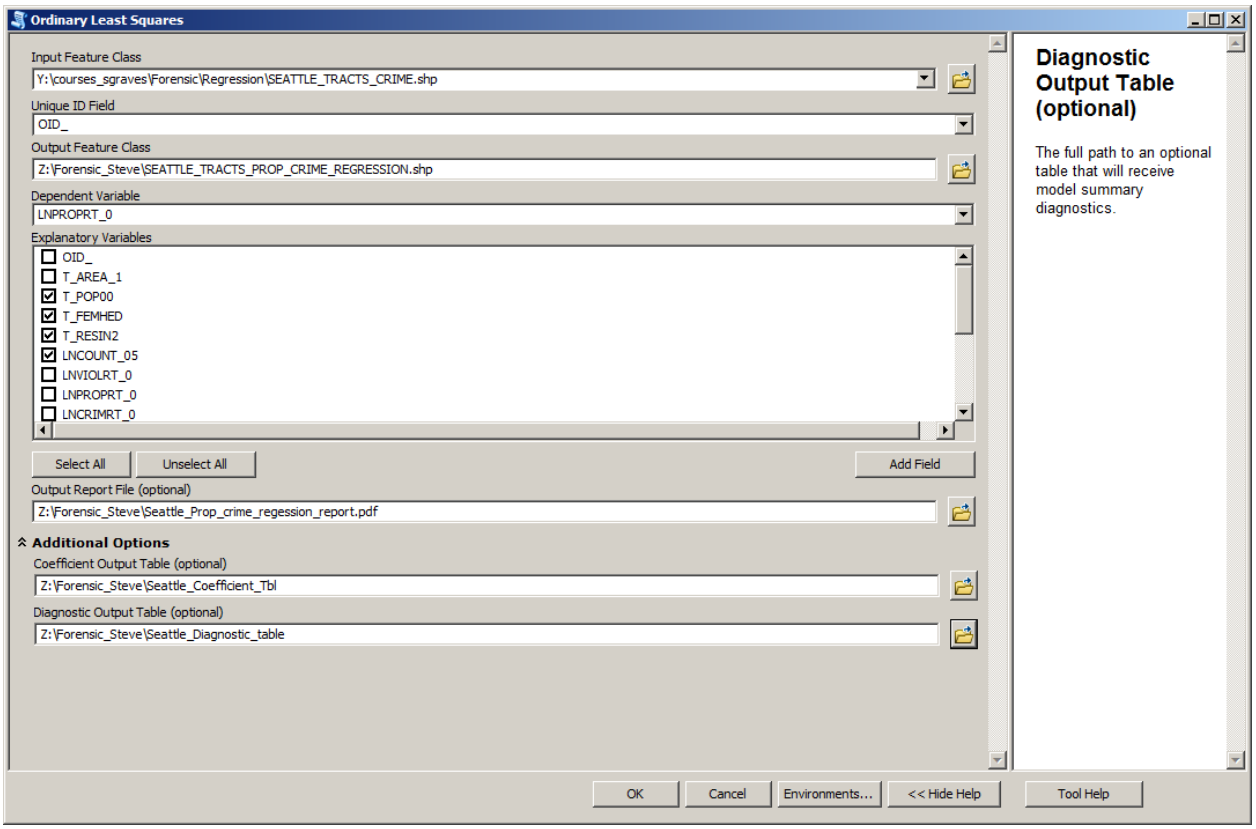
Observation	Predicted LNVIOLRT_06_07	Residuals	Standard Re
1	1.966994108	0.123005892	0.3
2	0.644660337	-0.154660337	-0.39
3	1.002406248	0.487593752	1.2
4	1.541712998	-0.001712998	-0.004
5	-0.727731198	-0.062268802	-0.157
6	1.634984775	0.115015225	0.29
7	1.343914163	0.086085837	0.217
8	0.579881855	-0.029881855	-0.075
9	-0.394482605	0.374482605	0.944
10	1.171199367	1.018800633	2.57
11	0.863316629	-0.103316629	-0.26
12	1.919279973	0.250720027	0.63
13	2.290449585	-0.370449585	-0.934
14	1.107675355	0.152324645	0.384
15	-0.007717765	-0.142282235	-0.358
16	0.680488047	0.959511953	2.42
17	1.842065824	-0.242065824	-0.61
18	2.135036721	-0.055036721	-0.138
19	0.89925313	-0.11925313	-0.3
20	0.559951204	-0.999951204	-2.522
21	0.889654517	-0.569654517	-1.437
22	-0.351533559	-0.638466441	-1.61
23	0.470239789	-1.100239789	-2.775

- ORANGE: These are the census tracts. It's too bad that Excel gives us only an ordered list of census tracts rather than the FIPS codes.
- BLUE: This is what the model predicted the Y/dependent variable should be…in the image to the right it's Violent Crime Rate. Sometimes this is the "expected" variable.
- GREEN: Residuals – this is how far off the model was from the actual, "observed" crime rate.
- PURPLE: This is the standardized residual, which means it's the residual's expressed as a standard deviation. This is useful to see just how "far off" individual residuals are compared to the rest. Residuals that are +/- 2 are above/below 95% of the rest of the residuals.

Step 3: Compare the results produced by Excel with the Regression Tool in ArcMap.

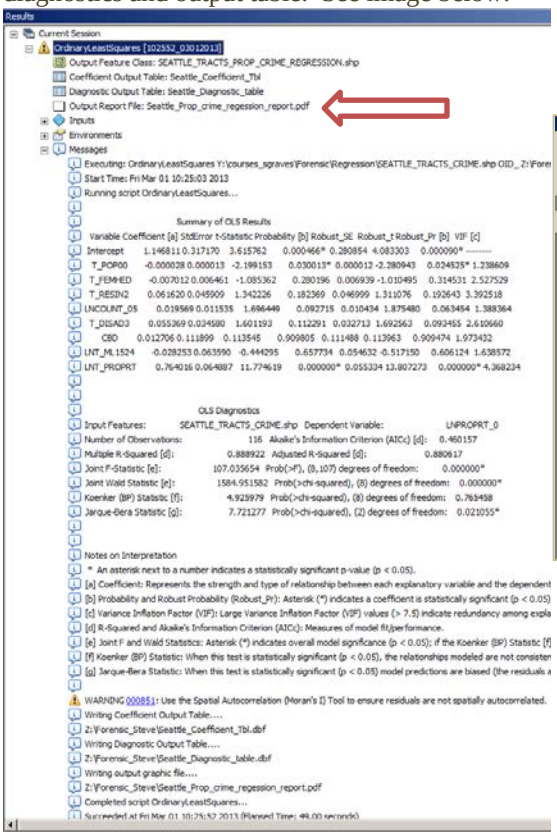


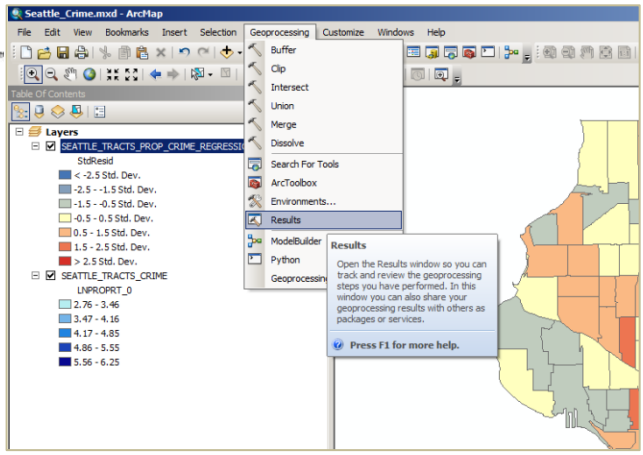
18. Open the map document Seattle_Crime.mxd. It's on the Y drive in the Forensic/Regression Folder. If it won't open, then just add the single layer file "SEATTLE_TRACTS_CRIME.shp".
19. In the image on the right, a thematic map of violent crime is visible. This is the data presented in class. You should map instead LNPROPR_0 for this assignment.
20. You may want to set the classification scheme to "equal interval".
21. Open the Arc Toolbox and from the list of Toolboxes, expand "Spatial Statistics Tools", and from the sublist, chose "Modeling Spatial Relationships" tool shelf, and pick the "Ordinary Least Squares" tool.
22. Fill out the Ordinary Least Squares Regression dialog box.
23. Make sure that your Dependent (Y) variable is LNPROPRT_0
24. Your explanatory variables should be the same that you used in the Excel component of the exercise above (Pop, FemHed, Resin, Count05, Disadv, CBD, ML_15_24, LNT_PROPRT). See image below.
25. You might also want to get fancy and produce an "Output Report", a Coefficient Output Table and a Diagnostic Output Table (the last two are under "Additional Options"). These tables just take the results and make them easier to read and place them in more permanent files. (see image below)
26. Click OK and wait a moment. Note that this map is already projected…which is a necessary precursor to running regression because the software will also check for "spatial autocorrelation"…which is a problem to avoid.



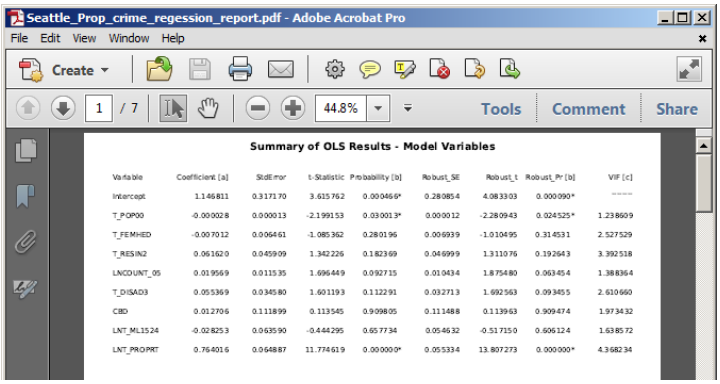
27. By default, the map that appears displays the Standardized Residuals, and if you click on it fast enough, the little “pop up” notification in the lower right corner of the map window, then you can see a ‘quick’

diagnostics and output table. See image below.





28. If you didn't click the little pop up window quickly enough, you can access the table of results of from your regression test by clicking on Geoprocessing/Results from the top tool ribbon tabs (see image above right)
29. You can read this results page (see image above) as is, or you can open the .pdf file that was generated by it by clicking on the link to the right of "Output Report File" (see red arrow above left). This .pdf file contains the same information, but it's just easier to read. (see below)
30. Note that the .pdf is 7 pages long, and should resemble reasonably closely the results you got using Excel. (see right)
31. Examine the results.



For Credit:

1. You are to identify TWO census tracts. For one, you are going to commit more resources to fight property crime. For the other you are going to give a bonus check to each officer in that district for having fewer property crimes than one would expect.

2. In a word processor, write a paragraph or two in which you: 1) identify the two tracts using FIPS code; 2) explain WHY you chose the two census tracts 3) Be sure to use NUMBERS (not just the map) to explain why these two tracts are deserving of help/recognition. 4) Introduce your short essay by writing a sentence or two in which you explain to a skeptical boss the strength and reliability of this model…before he asks. Be sure to include a screen capture of a map and of any other image to support your argument.
Email your report to your instructor.