University of North Texas
Department of Geography and the Environment

GEOG 5190

LAB 3: MULTIPLE REGRESSION

Scenario: You are a Community Data Analyst for the Planning Department of the City of Fort Worth and are working to understand the residential property development market in the city. You have collected statistics that represent the population density of Fort Worth's census tracts, along with other data fields that define the key characteristics of each census tract. Your task is to use multiple regression to provide you with insights into the factors that impact residential population densities in the city.

Part A: Forced Entry

Please use SPSS's <u>Linear Regression module</u> to perform a multiple regression on the urban data set provided. Your first step is to get the data set into SPSS (the data set print out is included with this lab, but the data set is also available for download from the course website). <u>Please see the final page of this lab</u> to clarify any questions you may have about the data set and its individual variables.

Once you have the data set in SPSS, use Multiple Regression to produce a model for Population Density (**Analyze > Regression > Linear …**). "Population Density" will be your dependent variable, and all the other variables (<u>except for "census tract ID"</u>) will be your independent variables. Under **Method...**, use the **Enter...** command to force all of the variables into the analysis simultaneously (again, <u>except for "census tract ID"</u>). The graphic below shows what your linear regression window should look like once these settings have been completed.
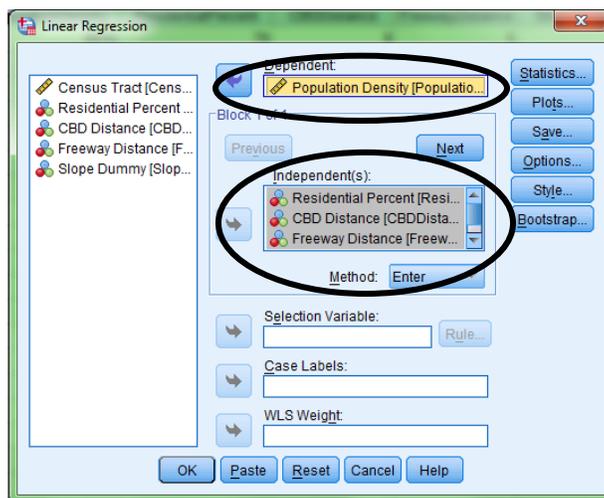
With all of these settings in place, please click **OK** to run the regression analysis.

With your regression analysis completed, please print out your results for inclusion in your report, indicating on the output which parts of the output relate to each part of the lab. Also, please answer the following questions based your output.

Part A, Question 1. Write out the regression equation (use *B* values, not *β* values).

Part A, Question 2. What is your multiple *R*? What does this multiple *R* value mean?

Part A, Question 3. What is the proportion of the variance explained?

Part B: Stepwise Entry

Step 1: Using the same data, run a Stepwise Multiple Regression (under the **Method** dropdown menu, choose **Stepwise**, as in the screen graphic below).
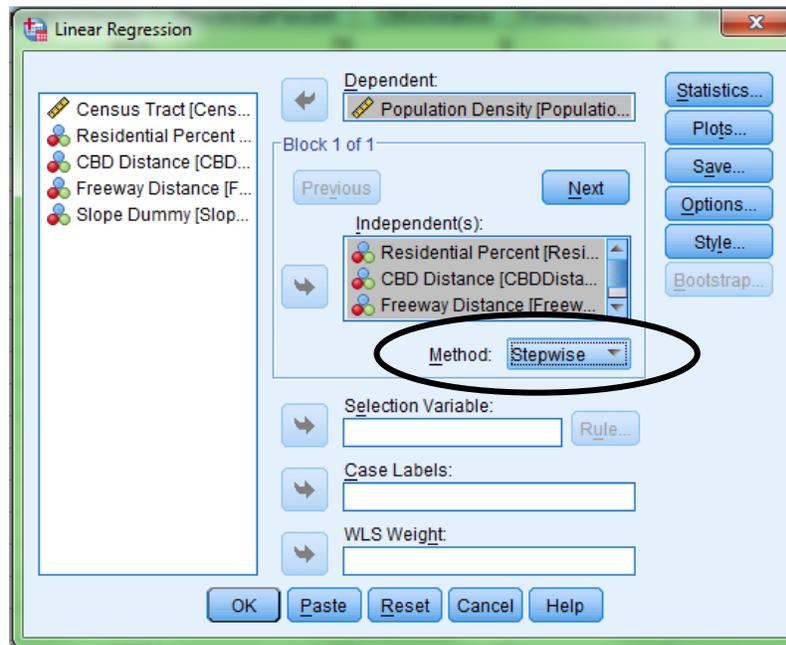

Once you have set this entry method, click **OK** to run the regression analysis.

Step 2: You will find that all of the variables will not enter the model.

Part B, Question 1: Which variables do not enter?

Running the multiple regression this way allows SPSS to evaluate which variables are of sufficient strength to be allowed into the model.

Step 3: Next, as in the screen shown below, go to **Options...** in the Linear Regression window and reset the probabilities for **Entry** and **Removal** to 0.90 and 0.91 respectively (the default values are 0.05 and 0.10). This will force ALL variables to be entered into your model using the Stepwise method – a helpful task to show us how SPSS deals with multiple variables and steps in regression modelling.
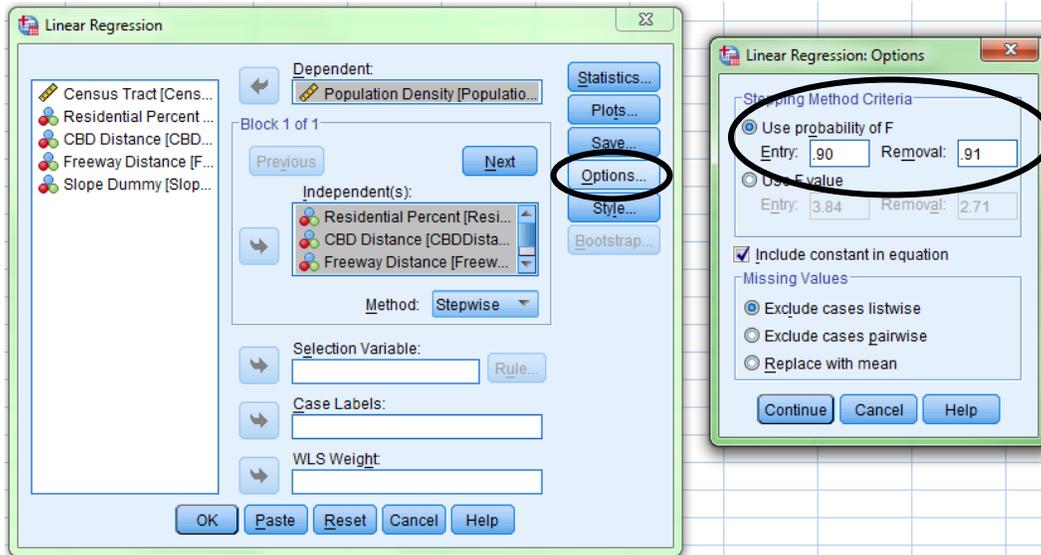
Once that these settings are complete, please answer the following questions:

Part B, Question 2. For each step of your second (stepwise) run, list the <u>name of the variable entered</u> at each step, the <u>total proportion of variation explained</u> at each step, and the <u>increment in this proportion</u> since the last step (remember, I showed you a table like this in our in-class discussion).

Part B, Question 3. Write out the <u>equation for the final model</u> (the one where all variables have been entered into the equation) using the B values from the SPSS output. Give the total proportion of variation explained by this final model. Use the final model from this question for all of your answers to Part B, Questions 4-8 below.

Part B, Question 4. Use this regression model (equation) to <u>predict the population density</u> for the following census tract:

    % land suitable = 100.0
    Miles from Central Business District (CBD) = 10.0
    Miles from nearest freeway interchange = 0.0
    Topography = flat land

Part B, Question 5. If you were looking for a census tract with the <u>highest possible population density</u>, what would you look for? Referring to your final model,
- Specify <u>what</u> you would look for, and
- Discuss <u>how</u> you came to that conclusion.

Part B, Question 6. Taking census tract 1, work out its residual (<u>remember</u>: residual = observed value - model value). Do the same for census tracts 3, 15, and 17. Census tracts 1 and 3 are on <u>serviced land</u> (meaning that utilities such as electrical, water, and telephone are present on site), while 15 and 17 are on <u>unserviced land</u> (meaning that utility infrastructure still needs to be completed for these census tracts). What, if anything, do the residual results suggest about the impact of servicing land for development?

Please note: as with any multivariate analysis <u>you may end up with results for question 6 that are not intuitive</u>; this is not necessarily an indication of a data or analysis error. When this happens (here or at any other time), you will need to invest some time in thinking about what a <u>counter-intuitive result</u> might mean.

Part B, Question 7. Using the values from Question 4 for "% land suitable", "miles from nearest freeway interchange", and "topography", <u>how far would you have to go from the CBD</u> to reach a point where the <u>population density is zero</u>? Be sure to include your equation.

Part B, Question 8. Using the variable values from Question 4 for "miles from CBD", "miles from freeway interchange", and "topography", what happens to your result if the "<u>% land suitable</u>" is <u>zero</u>? Comment on your answer.

Please remember to include a copy of your SPSS output in your written report.

Feel free to also integrate selected tables from the SPSS output alongside your answers above if you think this could be helpful.

Urban Data for Lab #3: Multiple Regression

This data set is available for you to download from the course website. There are 25 census tracts (observations) in the database. The variables, by column, are:

Column 1: ID number of the census tract
Column 2: Population density (persons/square mile)
Column 3: Percent of land suitable for residential use
Column 4: Miles from the CBD to the center of the census tract
Column 5: Miles from the nearest freeway interchange to the center of the census tract
Column 6: Topography: 1 for hilly, 0 for flat land (binary dummy variable)

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
1	8915	79	8	5	1
2	11492	68	7	2	0
3	10449	72	2	4	1
4	12311	91	7	2	1
5	12142	75	9	3	1
6	13020	93	12	0	0
7	15828	82	8	1	1
8	10025	60	17	4	0
9	15171	76	6	1	0
10	18325	58	5	2	1
11	16220	91	5	3	0
12	12511	94	12	3	1
13	13593	69	11	2	1
14	6880	57	18	3	0
15	11659	81	20	5	1
16	17484	72	1	5	1
17	16801	99	12	4	0
18	15820	90	3	4	1
19	12825	68	17	1	1
20	9368	69	16	1	1
21	13096	98	17	0	1
22	6134	57	19	4	0
23	21345	99	5	3	1
24	17296	58	7	0	0
25	5257	65	19	2	0