ASSIGNMENT 5

(1) In this study you will apply local G*-statistics to the Irish Blood Group data, which is displayed in the class directory file **T:\sys502\arcview\projects\eire\eire.mxd**. This study is discussed in Bailey and Gatrell (p.253,291) and more extensively in the Spatial Regression chapter by Upton and Singleton (pp.267-276), listed as **Item 18** in the class *Reference Materials*. The following additional materials have been also provided:

- First, the original study of Irish blood groups by Hackett and Dawson (1956) is provided as **Item 10** in the Reference Materials (**Irish_Blood_Groups.pdf**) [focus in particular on the Discussion starting on p.78]. Two more recent studies (**Blood_Group_paper_1.pdf** and **Blood_Group_paper_2.pdf**) are also included in the directory above, **T:\sys502\arcview\projects\eire**.

- Second, the key reference by Freeman (1969) in which the "Pale" counties were established is included in the file **Irish_Pale.pdf** listed as **Item 11** the Reference Material. Additional material can be obtained by Googling "Irish Pale", etc.

- Finally the original reference on *G-statistics* by Getis and Ord (1992) is included as **Item 6** in the Reference Material (**G_Statistics.pdf**). This should provide additional background material on the use of these statistics. (*NOTE:* Our present application of the *G** statistic is more general than their original application to radial distance matrices.)

The analysis of this data will be carried out in MATLAB and then exported to ARCMAP.

(i) First, reproduce this map document in your home directory by creating a directory, say **S:\home\eire**, and copying the three files **Eire(.shp,.dbf,.shx**) from the class directory **T:\sys502\arcview\projects\eire**. You can then display the blood group data as is done in **eire.mxd** and make a map document file of your own. Name the data frame as "Eire Data" and the blood group layer as "Blood Group A". Save this map document in your own directory as **eire.mxd**. (*NOTE*: The map scale is *unknown*, but 100 *km* is approximately equal to 160 of the map units shown.)

(j) Open MATLAB and import the file **T:\sys502\matlab\eire.txt**. You should now have a 26× 7 matrix, **eire**, in the workspace.

1. Define the *centroid locations*, **L**, and *blood-group percentages*, **z** , by:

```
» L = eire(:,1:2);
» z = eire(:,3);
```

2. Now make an exponential weight matrix by writing:

```
» info.type = [4,10,1];
» info.norm = 1;
» W = dist_wts(L,info);
```

In **info.type** the number **4** denotes the “exponential-weight-matrix” option, **10** denotes the exponent value, and **1** denotes the “include-diagonal-terms” option. Setting **info.norm = 1** yields a *row normalization* of the resulting matrix. [Please remember the convention used for exponential weights in expression (5.2.2) of the **NoteBook**.]

(k) Next, you will perform a **random-permutation test** using *local G*-statistics* and export the results to ARCMAP:

1. To perform this test in MATLAB using spatial data, **z**, and weight matrix, **W**, with 9999 random permutations, write:

```
» GP = g_perm_loc(z,W,9999);
```

2. The (26 x 2) output matrix, **GP**, contains the ***G*-statistics*** together with their associated **P-values** for each of the 26 counties in Eire.

3. You are going to join this data table to the attribute table for “Blood Group A”. This requires that the two tables have a common identifier. Since every attribute table in ARCMAP has an identifier “FID” that numbers all rows starting with the value “0”, the simplest procedure is to add a column to **GP** that is exactly of this form. To do so, simply write:

```
» GP(:,3) = [0:25]′ ;
```

Here the prime (′) transforms the row vector **[0:25]** to a column vector.

4. To export this data to your home directory as a file **G_stats**, write:

```
» save S:\home\G_stats.txt GP -ascii
```

5. To import this data into ARCMAP, you must first convert **G_stats.txt** into proper format using the procedure on pp.9-10 of Assignment 3. Here the three column titles of the new file, **G_stats.tab**, should be **G_star**, **P_Val**, and **ID**, respectively.

(l) With your map document **eire.mxd** open in ARCMAP, use **File** → **Add Data** to bring **G_stats.tab** into your document. It should now appear in the Table of Contents. To join this file to the attribute file for the layer “Blood Group A”:

1. Right click on “Blood Group A” and select **Joins and Relates** → **Join**.

2. In the **Join** window, set (1) = “FID”, (2) = “G_stats.tab”, (3) = “ID”, and click **OK** to finish.

3. If you open the attribute table for "Blood Group A" you will now see that new data appears in the last three fields.

4. To make this join permanent, right click on "Blood Group A" and select **Data → Export Data**. Now add this modified data to the map document as a new shape file, **eire_g.shp**. Before proceeding, it is a good idea to remove the join on "Blood Group A":

(i) Right click on "Blood Group A" and then select **Joins and Relates → Remove Join(s)**.

(j) Then select **Remove All Joins** and close. This will return the attribute table to its original state.

5. Rename the new layer from "eire_g.shp" to "P-Values". Open the properties of "P-values" and navigate to the **Symbology** window and select **Quantities**.

6. In the "Fields" box set **Value** = "P_Val", and in the "Classification" box, set **Classes** = 5 and click on **Classify**.

7. In the Classification window which opens, set **Method** = "Manual" and on the right hand side set the **Break Values** to be (0.01, 0.05, 0.10, 0.20). Click **OK**.

8. Set an appropriate **Color Ramp**, and exit with **Apply** and **OK**.

(m) Finally, using the procedure in part (1).(a).6 of Assignment 1, display the maps for layers "P-values" and "Blood Group A" side by side in WORD for comparison. (Include this display in your report.)

1. By comparing the distribution of Blood Group A percentages with the P-values obtained for each county, comment on whether these "concentration" results are what you would expect.

2. Look at the specific county of Waterford (FID 22), which is part of the original *Pale* defined by Freeman. How does its *Blood Group A percentage* compare with the mean value for Eire? What is its ranking among all counties in Eire? Now look at its associated *P-value*. How might you explain this apparent discrepancy?

3. Finally, take a look at the counties of Kildare (FID 8) and Meath (FID 16), both of which were also part of the Pale. First compare their *Blood Group A*

percentages, and then compare their *P-values*. How might you account for this apparent discrepancy?

(2) In this study you will lay the groundwork for a spatial regression analysis of **median housing values** in Philadelphia. This analysis will be continued in the final assignment (Assignment 6). A great deal of interesting information on the Philadelphia housing market can be found by Googling topics like "Philadelphia Housing Market". Just to get you started, I have included several items:

- The first is a set of summary profiles of trends in Philadelphia housing prices compiled by Kevin Gillen. This is **Item 17** in the Reference Materials (**Gillen_Housing_Indices.pdf**). The present data set is from the U.S. Census in **1990** (as I have marked on the first graph). So the historical trends shown in this reference are of most interest.
- In addition, I have included in the directory, **T:\sys502\arcview\projects\Phil_Housing**, an article, **Phila_Trends.pdf**, containing a brief discussion of housing market and demographic trends in Philadelphia (prior to the 2008 "market collapse"). See also the article on segregation, **Phila_Segregation.pdf**
- There is also a longer paper, **Philadelphia_Housing_Submarkets.pdf**, on the relation between traditional Philadelphia neighborhoods and housing submarkets.
- Finally, for those who may be interested, I have included a more detailed discussion, **Phila_Nbhd_Initiative.pdf**, of the *Neighborhood Transformation Initiative* (NTI) which is mentioned in the references above.

The relevant census tract data for Philadelphia is displayed in the ARCMAP file **T:\sys502\arcview\projects\Phil_Housing\Phil_Housing.mxd**. To analyze this data, first copy-and-paste all "**tracts90**" files into a directory of your own, say **S:\home\Phil_Housing**. Now open a new map document file (**.mxd** file) in ARCMAP and use **Add Data** to add the shape file **tracts90** to the data frame. You will observe that the default projection for Philadelphia looks wrong. Also, if you open the Attribute Table and scroll to the last two columns, you will see that the coordinates for this map are in decimal degrees. Hence to obtain a more natural projection in terms of Euclidean distance (in feet), your first task is to *project* this shapefile into State Plane Feet. This two-step procedure requires you to *define* and *construct* the projection. To do so:

(a) First open **ArcToolbox** on the main menu, and click:

Data Management Tools → Projections and Transformations
→ Define Projection

(b) In the window that opens, set **Input Data Set** = "tracts90". If the **Coordinate System** window now contains the coordinate system "GCS_North_American_1983", then it is already defined, and you may proceed to step (c) below. Otherwise, click the Browse icon to the right, and in the "Spatial Reference Properties" window that opens, click **Select…** and then proceed to

Geographic Coordinate Systems
→ **North America**
→ **USA and territories**
→ **NAD 1983** [NAD = "North American Datum"]

Click **Add** and **OK**, and the desired definition should now appear in the **Coordinate System** window. Click **OK**. When the process is completed, click **Close**.

(c) Next you will *construct* the projection. To do so return to **ArcToolbox** and click:

Data Management Tools → Projections and Transformations
→ Project

(d) In the **Project** window that opens, again set **Input Data Set** = "tracts90". A path will now appear in the **Output Data Set** window to a file called 'tracts90_Project'. Reset the path to your home directory, **S:\home\Phil_Housing**, and rename the output file as **tracts90_feet**. For the **Output Coordinate System** click the Browse icon, and in the "Spatial Reference Properties" window proceed to

Projected Coordinate Systems
→ **State Plane → NAD 1983 (US Feet)**
→ **NAD 1983 StatePlane Pennsylvania South FIPS 3702 (Feet)**

As before, click **Add** and **OK**, and the desired definition should now appear in the **Output Coordinate System** window. Click **OK**. When the process is completed, click **Close**.

(e) To see the new projection, first close this **.mxd** file (no need to save). Now open a new **.mxd** file and add your file, **tracts90_feet**. The projection should now look correct. Notice also that the default coordinates on the bottom of the map are now in *feet*.

(f) Your final task here is to recalculate the coordinates in terms of feet. To do so, open the Attribute Table for **tracts90_feet** and scroll to the last two columns, which list the coordinates in the old decimal degree units. You will now recalculate these in State Plane feet as follows:

- First right click on the **X_coord** column and select **Calculate Geometry…**

- In the "Calculate Geometry" window set **Property** = "X Coordinate of Centroid".

- Click **OK**, and when the calculation is complete, you should now see that the coordinate values are now in feet rather than decimal degrees.

Now repeat this procedure for the **Y_coord** column, using **Property** = "Y Coordinate of Centroid".

(g) Finally, rename the data frame as **Philadelphia Tract Data (1990)**, and save this map document in your home directory as **Phil_Housing.mxd**. You are now ready to begin!

(h) Our objective in this study is to develop a simple regression model to identify key predictors of **median housing value** in 1990 Philadelphia. From the above references, it should be clear that both **racial segregation** and **housing abandonmen**t were major issues in 1990, and can be expected to be negatively correlated with housing prices. As one positive predictor, it is equally clear that higher priced housing should be associated with higher **household incomes**. These three representative variables will form the starting point for our analysis.

(i) But before proceeding to the analysis, there is some additional "data cleaning" that needs to be done. Our primary variable of interest is Median Housing Value (MEDIANVAL). If you open the attribute table for the layer **tracts90_feet** and sort the column MEDIANVAL by first right clicking on the column label and then choosing the "Sort Ascending" option, you will see that there are 14 census tracts with MEDIANVAL = 0. These either involve missing data or zero housing units. All such tracts should be eliminated from the analysis. [You will also see a large number of tracts with MEDIANVAL = 14999. The reason for this is that the U.S. Census convention is to report all lower housing values as $14999. So be aware that this will necessarily create some bias in the later analysis. But since this Census reporting convention is standard, we shall keep these tracts in the analysis.] In addition, it turns out that Median Household Income (MEDHHINC) was not recorded for 14 tracts (not all the same as above). So it is desirable to eliminate these tracts as well.

1. To do so, use the "Select by Attributes" option [as in part (1).(a).5 of Assignment 1] to create a new layer containing only those census tracts with both positive housing values and household incomes. In the SELECT WHERE window, write

   ("MEDIANVAL" > 0) AND ("MEDHHINC" > 0)

   and click **Apply**.

2. Make a new shape file **Phil_Tracts_Final.shp** containing these selected tracts, and add it to the data frame. Before proceeding, open the attribute table of this new layer and check on the bottom to be sure that you now have **351** records. (You can also remove the original layer since it will no longer be used.)

3. Name this layer as "Median Housing Value", and in the **Symbology** window for this layer, display the MEDIANVAL data on the map, with an

appropriate color ramp. This will be the key dependent variable for the regression analysis.

(j) Next you will construct operational definitions of the desired explanatory variables, and compare them visually with MEDIANVAL as follows.

1. First, we characterize housing abandonment in terms of **percent of vacant housing** in each census tract. To do so, open the attribute table for "Median Housing Value" layer and observe that there are two columns "HOUSUNITS" and "VACANT".

   (a) To construct the desired quantity use **Options → Add Field** to create a new column (**Name** = "PCT_VACANT", **Type** = "double", **Precision** = "6", **Scale** = "3"). [Here **Precision** denotes the maximum number of digits to be used, and **Scale** denotes the number of decimal places.]

   (b) When the new field appears, right click on the column head and select **Calculate Values**.

   (c) Finally, in the **Field Calculator** window for "PCT_VACANT = " you should write: 100*[VACANT]/ [HOUSUNITS]. (The first value in the column should be 6.732.)

2. Next we characterize racial segregation in terms of the **percent of nonwhite population** in each census tract. To do so, observe that in the attribute table there are also columns "PERSONS" and "WHITE". So to make a new column "PCT_NONWT", proceed as part 1 above by writing 100*([PERSONS]-[WHITE])/[PERSONS] in the "PCT_NONWT = " window. (The first value in the column should be 3.792.)

3. Finally, since our third variable, **median household income**, is already included in the Table as MEDHHINC, you are ready to display these new variables as maps.

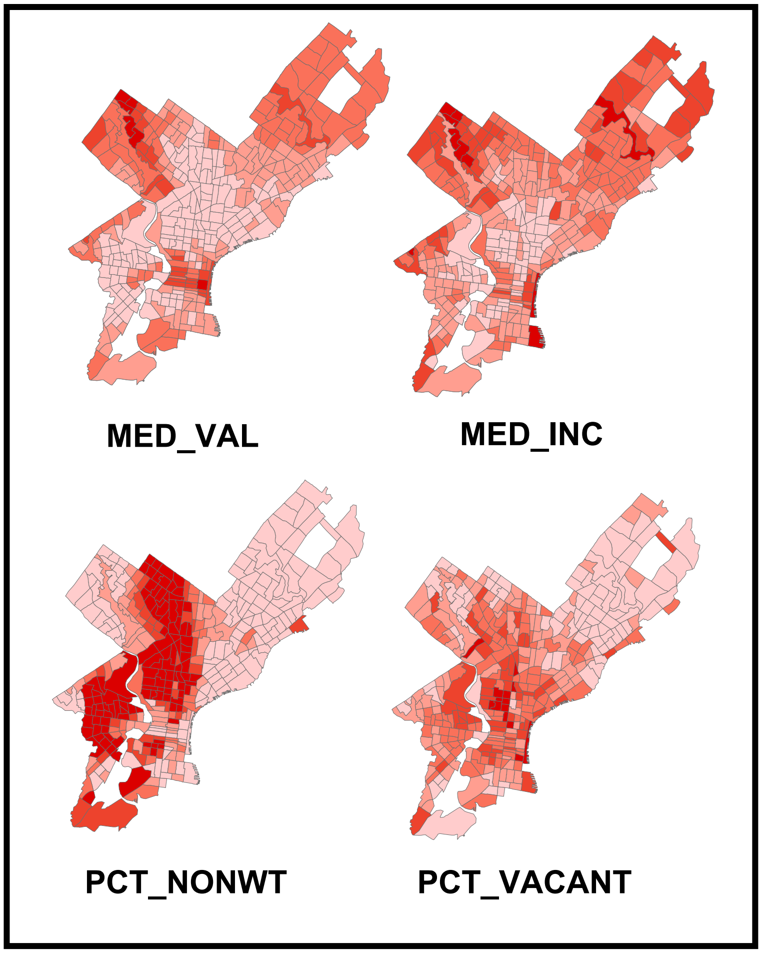
   (a) First, use **Copy** and **Paste Layer** [as in part (c).3.b of Problem 2 in Assignment 3] to make three copies of "Median Housing Prices" layer in the data frame.

   (b) Rename the three new layers as "Percent of Vacant Housing", "Percent NonWhite", and "Median Income".

   (c) In the Symbology window for "Percent of Vacant Housing", select the PCT_VACANT variable for display, and choose the same color ramp as for the "Median Housing Value" layer.

(d) Repeat this process for the “Percent NonWhite” and “Median Income” layers, and display the corresponding PCT_NONWT and MEDHHINC variables.

4. Next, construct a visual comparison of these four layers by successively copying each map to WORD, and resizing them so that they can be displayed together. Finally, label the variable displayed in each figure using **text boxes**. If you have done all steps correctly, you obtain a final display similar to the one shown below (which uses a red color ramp for all three maps):



MED_VAL

MED_INC

PCT_NONWT

PCT_VACANT

(k) By comparing these four figures visually:

1. What can you say about the expected signs of correlations between MED_VAL and each of the explanatory variables?

2. What can you say about possible spatial autocorrelation between the values of each variable across census tracts?

(l) Next you will export the data to JMP for analysis.

1. To do so, first open the data base file **Phil_Tracts_Final.dbf** in EXCEL, and remove all columns except "MEDHHINC", "MEDIANVAL", "X-COORD", "Y-COORD", , "PCT_VACANT", and "PCT_NONWT". Save the file as a tab delimited text file, **Phil_Housing.txt**.

2. Now open this file in JMP [using "Text Files (.txt)"], and save the file in your own directory as **Phil_Housing.jmp**.

(a) For convenience, rename the columns as **MI**, **MV**, **X**, **Y**, **%VAC**, and **%NW**, and for later use, use **Cols** → **Reorder Columns** → **Move Selected Columns** to reorder these columns with the dependent variable first as (**MV**, **MI**, **%NW**, **%VAC**, **X** and **Y**).

(b) Finally, check to be sure that you have exactly 351 rows of data. If you have extra rows (such as repeats of row 351 or zero-value rows) delete these from the table. [Sometimes data transfers create such errors.]

3. Display the frequency distribution of median values, **MV** (using **Analyze** → **Distribution** → **MV**), and observe that these values are highly skewed. To remove this effect, take logs:

(a) First make a new column **lnMV** just to the right of **MV** (using **Cols** → **New Columns**).

(b) Then right click on **lnMV** and use **Formula** → **Transcendental** → **Ln** to construct the natural log of **MV**.

(c) Comment on your result by comparing the Normal Quantile Plots of these two distributions.

4. Next you will examine the frequency distributions of the three explanatory variables, **MI**, **%VAC**, and **%NW**. [Before doing so, it should be re-emphasized that it is *not* essential for explanatory variables to be normally distributed. Consider for example a "dummy" variable, which of course can *never* be normal. What is important however is that the distributions of explanatory variables have sufficient spread to provide "leverage" for identifying regression coefficients. For example, a dummy variable with almost all values equal to zero may not provide sufficient information for estimation purposes.]

(a) Note first that like **MV** above, the variable **MI** is also highly skewed, which may affect its leverage for estimation purposes. More important for our present analysis is the fact that **MI** is in *monetary* units. So, to maintain compatibility with **MV**, it is appropriate to transform **MI** to **lnMI** in a manner paralleling **MV**.

(b) Observe next that **%VAC** also exhibits some degree of skewness. However, there are several tracts with zero vacancies, which preclude log transformations. Moreover, while this "zero problem" can in principle be avoided by modified transformations such as **ln(1 + %VAC)**, in the present case zero-vacancy values might be particularly informative. So we choose not to transform this variable.

(c) Turning finally to **%NW**, notice the dramatic *bimodal* nature of this distribution. How might you interpret this finding? This bimodal property is of potential importance for the present analysis, and we thus choose to leave this variable as is.

5. With these preliminary observations, you are now ready to perform the desired multiple regression.

(a) To regress **lnMV** on (**lnMI**, **%NW**, **%VAC**) in JMP, use **Analyze → Fit Model**. (Be sure to set **Emphasis** = "Minimal Report" to avoid unnecessary graphics.)

(b) Notice first that while **lnMI** and **%NW** are both significant (with the right signs), **%VAC** is completely insignificant. To gain further insight here:

- First carry out a simple regression of **lnMV** on **%VAC**.

- Now regress **%VAC** on the other two explanatory variables (**lnMI**, **%NW**).

- By comparing the **p-values** in these two regressions, explain what you think might be happening to **%VAC** in the original regression.

(c) Given the insignificance of **%VAC**, we now drop this variable from the analysis. To justify this decision, compare the **Adjusted R-Square** values in the original regression with this new regression (i.e., including only **lnMI** and **%NW**) and comment on the results.

(d) Also by comparing the signs and p-values of **lnMI** and **%NW** in both regressions, what else can you conclude about the influence of **%VAC** ?

(e) To analyze this new regression further, right click on the top bar, **Response lnMV**, and select **Row Diagnostics > Plot Residual by Predicted**. Notice in this plot that there is a concentration of **lnMV Predicted** values at both ends.

- To gain further insight, save these predicted values in JMP (by again clicking on **Response lnMV,** and selecting **Save Columns > Predicted Values**).

- Now plot these predicted values against **%NW** and discuss how this relation helps to explain the above pattern of **lnMV Predicted** values.

(f) Before proceeding, click once more on **Response lnMV** and select **Save Columns → Residuals** to save the regression residuals as a new column, which you should label as **Res**. Also delete the column, **Predicted lnMV**, which will not be used further.

(m) Next you will export the data to MATLAB and construct a **spatial weight matrix**:

1. First construct a new file from **Phil_Housing.jmp** by keeping only the columns (**lnMV**, **lnMI**, **%NW**, **X**, **Y**, **Res**) and saving this file in EXCEL as **Phil_Matrix.xlsx**

2. Next open MATLAB and import this EXCEL file file to the workspace. Here it is convenient to first import this file with **Output Type** = "Table", and rename it as **Phil_Table**. While this file is not directly useful for computations, if you click on it in the workspace you will see that it contains all column headers – which provides a useful reference.

3. With the Import Window still open, switch to **Output Type** = "Numerical Matrix", and import the EXCEL file once again as **Phil_Matrix** (which should appear as a 351×6 matrix). This is the file we will use for computations. Finally, save this workspace to your home directory as **Phil_Housing.mat**.

4. The next task is to make an appropriate weight matrix for spatial regression. To do so, observe from **Phil_Table** that the spatial coordinates (**X**,**Y**) are in columns 4 and 5 of Phil_Matrix. So we can make a matrix, **L**, of locations by using the command,

```
>> L = Phil_Matrix(:,4:5);
```

(i) The desired weight matrix, **W**, to be constructed is a *near-neighbor* matrix consisting of the first **4** nearest neighbors to each location. This can be accomplished by using option 1 in the program, **dist_wts.m**, with the following set of commands:

```
>> info.type = [1,4];
>> info.norm = 1;
>> W = dist_wts(L,info);
```

The first command sets the nearest-neighbor option with 4 nearest neighbors, and the second command sets specifies *row normalization*, so that the 4 positive values in each row will equal .25 (and thus sum to one).

(ii) To see that this program is working properly, set **Wt** equal to the transpose of **W** by writing

```
>> Wt = transpose(W);
```

a. Now display **Wt** by double clicking on it in the workspace.

b. You will then see a column of pairs, **(j,i)** , in which **j** is one of the first four nearest neighbors of **i**. So for example, the first four nearest neighbors of tract **1** are tracts (**2,3,4,6**). This is a convenient "trick" to identify nearest neighbors easily. [The column of .25 values are the cell values resulting from the row normalization above.] Now scroll down to the nearest neighbors of tract **49**, which should be (**40,43,57,64**).

c. To identify these neighbors in ARCMAP, recall that that FID numbers in ARCMAP attribute tables start counting from "0" rather than "1", so that row **49** corresponds to tract FID number **48** in ARCMAP. Similarly by subtracting "1" from each neighbor above, you will obtain the FID numbers of these tracts in ARMAP.

d. Now open ARCMAP and examine these tracts. Do the results of MATLAB make sense in terms of these tract locations? Comment.

e. Notice also that along with these four neighbors, tract **31** is also contiguous to tract **48**. So it might be desirable to include this tract as well. You will do so in Assignment 6, where *boundary-share* weights will be constructed for this example.

(n) Next recalling that the regression residuals, **Res**, were in the last column of **Phil_Matrix**, construct a combination of the four nearest-neighbor residuals, **nn4_res**, by writing:

```
>> res = Phil_Matrix(:,6);
>> nn4_res = W*res;
```

Before leaving MATLAB, test for spatial autocorrelation in **res** by using **sac_perm**:

```
>> sac_perm(res,W,999);
```

You will compare these results with a JMP analysis below.

(o) Finally, export these weighted residuals back to JMP by writing:

```
>> save “S:\home\nn4_res.txt -ascii
```

(Be sure to TYPE the last expression, **-ascii**, using MATLAB keys to be sure that the font is correct. Otherwise, you may get an error message.)

(p) With **Phil_Housing.jmp** open in JMP, import **nn4_res.txt** to JMP. Copy-Paste this data into **Phil_Housing.jmp** as a new column, **nn4_Res**.

(q) Finally, use **Analyze → Fit Y by X** to regress **Res** on **nn4_Res**.

(1) What does the **p-value** on **nn4_Res** tell you about the multiple regression above?

(2) How do these results relate to the **sac_perm** test done in part (n) above? Do they support one another?

(3) How does the value of **rho** in the output of **sac_perm** relate to the present slope estimate for **nn4_Res**? Why? (You might also try using **Analyze → Fit Model**, and repeat the regression of **Res** on **nn4_Res** using the **No Intercept** option at the bottom of the **Fit Model** window.)

(4) What additional information about this regression result is added by the output of **sac_perm**? (Compare the value of **rho** with the simulated *range* of values for **rho** reported by **sac_perm**).

(5) What do you expect to happen if you redo this analysis using **spatial regression models**? [You will do so in the final assignment for this course.]