Testing Regression Residuals for Spatial Autocorrelation using SAS: A Technical Note

John Paul Jones, III
Univesity of Kentucky, USA*

Stuart A. Foster
Western Kentucky University, USA**

Spatial autocorrelation of regression residuals is a violation of an assumption of the general linear model. Yet despite the widespread use of spatial data in applied work, many researchers fail to examine residuals for spatial autocorrelation, or they rely on subjective interpretation of residual maps. In part, this problem stems from a lack of easy-to-use programs that can be incorporated into existing research procedures. In this note we present a short program for testing spatial association among regression residuals. Based on Moran's I statistic, the program employs SAS's PROC MATRIX language, and can easily augment regression analyses run using the SAS package.

An important assumption in regression analysis is the lack of association among error terms. This assumption may be easily violated in data analyses which employ temporal or spatial series. In the case of time series, the independence assumption is readily assessed by the Durbin-Watson test, which is routinely reported in most statistical packages. For the user of spatial data, however, the assumption of error term independence is more problematic. Tests for spatially autocorrelated residuals have been developed by Cliff and Ord (1973; 1980), but the formulae are cumbersome and are not incorporated into standard packages. Equally disconcerting is the selection of weights. The two-dimensional and non-regular nature of spatial data leads to a variety of possible weight specifications, and autocorrelation tests may be sensitive to the particular configuration chosen.

These problems have no doubt inhibited many researchers from performing tests on residuals as a means of assessing conformance to the assumptions of regression. Where formal tests are not performed, there may be a reliance upon maps of residuals, but this

* Department of Geography, University of Kentucky, Lexington, KY 40506, USA.
** Department of Geography and Geology, Western Kentucky University, Bowling Green, KY 42101, USA.

Geography Research Forum • Vol. 11 • 1991:78-83

procedure is subjective and should not be used as a definitive guide concerning the level of autocorrelation among residuals.

As a step in simplifying the assessment of spatial independence in residuals, this paper reports a program for performing spatial autocorrelation tests using the I statistic (Cliff and Ord, 1973; 1980). Written in SAS's PROC MATRIX language, it enables us to reduce to a few lines what would, in a FORTRAN program, be a large number of statements. The program can be merged with SAS regression programs, or it may be used independently, with residuals and other data supplied by the user. We also discuss alternative weight matrix configurations and their specification in SAS.

THE I STATISTIC

In the ordinary least squares regression model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e} \tag{1}$$

the estimates of regression parameters are best linear unbiased (BLUE) if a number of assumptions are met (Draper and Smith, 1966; Poole and O'Farrell, 1971). Our concern is with the assumption of error term independence. Residuals are found by

$$\mathbf{e} = \mathbf{Y} - \mathbf{XB} \tag{2}$$

When these are pairwise (positively) correlated over space, the standard errors associated with the estimates are downward biased, and inflation of R^2 will result. The problem is indicative of a nonlinear relationship, missing variables, or the need for an autoregressive structure.

Cliff and Ord (1980) developed a test for spatial independence of regression residuals. The distribution for the test statistic, I, has been examined under the assumption of normality (N) and randomness (R). They conclude that assumption N is the preferred distribution, and this is the form of the test presented here.

The test statistic, I, is given by (Cliff and Ord, 1980):

$$I = \frac{n}{S_0}\left(\frac{\mathbf{e'We}}{\mathbf{e'e}}\right) \tag{3}$$

where $\mathbf{W}$ is a modifiable n-by-n weight matrix and S_0 is the sum of all w_{ij} elements. The w_{ij}'s specify the relationship between ith and jth observations, with main diagonal elements set to zero.

The expected value of I, E(I), is

$$E(I) = -\frac{n^* tr(\mathbf{A})}{(n-k)S_0} \tag{4}$$

where tr denotes the trace of matrix $\mathbf{A}$, and where

$$\mathbf{A} = (\mathbf{X'X})^{-1}\mathbf{X'WX} \tag{5}$$

The variance of I is found by

$$\text{Var (I)} = \frac{n^2}{S_0^2(n-k)(n-k+2)} \left\{ S_1 + 2\text{tr}(A)^2 - \text{tr}(B) - \frac{2[\text{tr}(A)]^2}{n-k} \right\} \quad (6)$$

where S_1 is one half of the sum of the squared elements of $W + W'$, and where

$$B = 4(X'X)^{-1}X'U^2X \quad (7)$$

$$U = 0.5(W + W') \quad (8)$$

The z-variate for I,

$$z = \frac{I-E(I)}{\sqrt{\text{Var(I)}}} \quad (9)$$

may be tested for significance in the usual manner, commonly under the one tailed alternative hypothesis that $z>0$. In the section below we translate equations (3) through (9) into the SAS PROC MATRIX language.

SAS PROGRAM FOR SPATIAL AUTOCORRELATION TEST

PROC MATRIX (*SAS User's Guide*, 1982b) is a programming language which allows operations to be performed on entire matrices, vectors, and scalars. The language follows matrix algebra notation, with special operators and functions enabling a reduction in programming lines. For example, matrix multiplication is performed by the operator, *, while matrix inversion employs a function, INV.

Table 1 presents the SAS statements required to evaluate spatial autocorrelation using the I statistic under the assumption of normality. Three DATA steps input the required values of the independent variables, the residuals, and the weight matrix. DATA IVAR specifies the matrix of independent variables used in the regression. "k" in line 2 is the userspecified number of columns of X, where the first column, X1, must be set to unity. The records containing these values follow the CARDS statement; their placement in the program is denoted by the matrix insert. Using "list" input requires that each observation begins on a new record and that the ordered variables be separated by at least one blank space. Additional information on input modes can be found in *SAS User's Guide* (1982a). Lines 4 through 6 read a vector of regression residuals, E. Lines 7 through 9 specify the weight matrix, W. It can be read as shown provided the elements of the matrix are separated by blanks and that each row begins on a new record. "n" (line 8) is a user-specified number of observations.

Table 1: SAS PROC MATRIX program for testing regression residuals for spatial autocorrelation.

```
1       DATA IVAR;
2           INPUT XI-Xk;
3           CARDS;
            1     x2_1   x3_1   ...   xk_1
            1     x2_2   x3_2   ...   xk_2
            1     ...    ...    ...   ...
            1     ...    ...    ...   ...
            1     x2_n   x3_n . ..    xk_n
4       DATA RESID;
5           INPUT E;
6           CARDS;
            e_1
            e_2
            ...
            ...
            e_n
7       DATA WEIGHTS;
8           INPUT W1-Wn;
9           CARDS;
            0     w_12   w_13   ...   w_1n
            w_21  0      w_23   ...   w_2n
            ...   ...    ...    ...   ...
            ...   ...    ...    ...   ...
            w_n1  w_n2   w_n3   ...   0
10      PROC MATRIX;
11      FETCH X DATA=IVAR;
12      FETCH E DATA=RESID;
13      FETCH W DATA=WEIGHTS;
14      N=NROW(X);
15      K=NCOL(X);
16      S0=SUM(W);
17      S1=0.5#SUM((W+W')##2);
18      U=0.5*(W+W');
19      A=INV(X'*X)*X'*W*X;
20      B=4#INV(X'*X)*X'*(U**2)*X;
21      I=(N#/S0)#((E'*W*E)#/(E'*E));
22      EI=-(N#TRACE(A))#/((N-K)#S0);
23      VARI=((N##2)#/((S0##2)#(N-K)#(N-K+2))) #
24          (S1+2#TRACE(A**2)-TRACE(B)-2#(TRACE(A)##2)#/(N-K));
25      SDI= VARI##0.5;
26      Z=(I-EI)#/SDI;
27      PRINT I EI VARI SDI Z;
```

Notes:

1) k in line 2 is a user specified number of columns of X (number of independent variables plus one).
2) X's first column is set to unity.
3) n in line 8 is a user specified number of observations.

PROC MATRIX is invoked in line 10. Lines 11 to 13 assemble the data, while lines 14 and 15 compute the number of observations and columns of x (note that k is one more than the number of independent variables since the first column of X is set to unity). Lines 16 through 26 replicate, in PROC MATRIX notation, equations (3) through (9). I (line 21) is the calculated value of the test statistic; EI (line 22) is its expected value; VARI (lines 23-24) is the variance of I; SDI (line 25) is the standard deviation of I; Z (line 26) is the computed z-variate. These are printed using the PRINT statement (line 27). Alternatively, all the information including the original data may be printed using the statement: PRINT;.

WEIGHT MATRIX SPECIFICATION IN SAS

The most common configuration of the weight matrix is a contiguity relationship where $w_{ij} = 1$ when observations share a boundary and $w_{ij} = 0$, otherwise. Tests performed using this configuration of the matrix are equivalent to asking whether or not contiguous areal units tend to have similar residual values. The contiguity matrix involves no additional programming lines, as W is read directly by the statements in lines 7 through 9 of Table 1. Another measure of association is the percentage of an observation's boundary shared with a neighboring observation (Cliff and Ord, 1980). This has the advantage of retaining more information than a binary matrix, and can partially compensate for differences in the size of areal units. Alternatively, the w_{ij}'s may be measures of interaction among the observations (Gatrell, 1979). In this case it may be justified to consider nonlinear effects. This can be implemented by associating a parameter to the weight, as in w_{ij}^{b}. Values of b greater than unity enhance weights for any given measure, while values less than unity de-emphasize the w_{ij}'s. An analysis of the sensitivity of spatial autocorrelation tests to these modifications can be performed by running the program with different values of b. The PROC MATRIX statement required to implement this operation is

15a W = W##b; This should be inserted in Table 1 after line 15.

A common measure of (negative) association is the distance, d_{ij}, between observations or their centroids. This is a useful form of weighting when the observations are points (e.g., weather stations, cities). Let the weight matrix contain the d_{ij} values; a positive measure of association results when the weights are raised to a negative power, as in d_{ij}^{-b}. Higher values of b amount to testing for more local effects, while lower values provide for more spatially dispersed, or regional, tests. Sensitivity analyses may be performed by altering the b values. The PROC MATRIX statements required when weights are raised to a negative power are:

15a W = W + I(N) ;

15b W = (W##-b) - I(N) ;

Line 15a adds the identity matrix to W, ensuring that all elements are nonzero. Line 15b raises the w_{ij} values to the user-specified value of b, and then subtracts the identity matrix to the result, thus returning the diagonal elements to zero.

CONCLUSION

Autocorrelation tests on regression residuals from spatial data analyses may never be as routine as the equivalent tests in time series analysis. The complications that arise when using irregularly-spaced two-dimensional data (with a variety of possible weight matrix configurations) will doubtless leave many statistical packages without such capabilities in the near future. Nevertheless, the possibility of violating an assumption of the general linear model should not be ignored by those who use spatial series in regression analyses. The consequences of spatial interdependence are deflated standard errors and inflated R^2. Ultimately, spatial autocorrelation limits our ability to draw meaningful conclusions regarding the parameters of functional relations.

The program reported here takes advantage of the matrix manipulations capabilities of SAS[1]. Requiring only a few statements, the program can be easily employed to test for residual autocorrelation. If detected, autocorrelation can be addressed by examining the data for nonlinear relationships, a missing variable, or by directly incorporating a spatially autoregressive structure.

NOTES

1. In the absence of SAS, one may adapt the program shown in Table 1 to virtually any matrix language.

REFERENCES

Cliff, A.D. and Ord, J.K. (1973). *Spatial Autocorrelation.* London: Pion.

Cliff, A.D. and Ord, J.K. (1980). *Spatial Processes: Models and Applications.* London: Pion.

Draper, N.R. and Smith, H. (1966). *Applied Regression Analysis.* New York: Wiley.

Gatrell, A.C. (1979). Autocorrelation in space. *Environment and Planning,* 11:507-516.

Poole, M.A. and O'Farrell, P.N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers,* 52:145-156.

SAS User's Guide: Basics (1982a). Cary, NC: SAS Institute.

SAS User's Guide: Statistics (1982b). Cary, NC: SAS Institute.