Testing Spatial Autocorrelation with the Bootstrap

Anna Shchiptsova

September 30, 2016

International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria

This document outlines the bootstrap algorithm of testing spatial autocorrelation in regression model. The Moran's I (Moran, 1950) and Geary's C (Geary, 1954) indicators are used as statistics of spatial autocorrelation.

CONTENTS

1 Software

No installation needed. All packages are standalone java applications.
Requires JRE 1.8 installed on the target machine.

URL for download:
http://www.iiasa.ac.at/web/home/research/researchPrograms/AdvancedSystemsAnalysis/land-use-spatial-analysis.html

Name	regression-tests-1.0.0-standalone.jar
Type	jar package
Summary	Library for spatial statistical analysis with resampling
Version	1.0.0
License	MIT, http://opensource.org/licenses/MIT
Imports	Clojure 1.8.0, https://clojure.org/; Incanter 1.5.7, http://incanter.org/
Command line options	-t, --trace Print stack trace
	-h, --help Print command help
Author and maintainer	Anna Shchiptsova, shchipts@iiasa.ac.at

2 Model

Let us consider a geographic region consisting of n administrative units. Suppose that we have panel data (X, y) collected in every administrative unit of the region. Here, X is a $n \times (p+1)$ matrix of the explanatory variables and y is a $n \times 1$ observable vector of the response. Each column X^i consists of the sample observations on a single explanatory variable.

In general, we want to relate the response variable to available explanatory factors in an administrative unit based on the reported spatial panel data. For this purpose, we put forward a multiple regression model in the following form

$$\begin{gathered} y = X\beta + \varepsilon \\ \varepsilon_1, \dots, \varepsilon_n \sim F(0, \sigma^2) \end{gathered} \tag{1}$$

where $\beta = \left(\beta_0, \beta_1 \dots, \beta_p\right)^T$ is a $(p+1) \times 1$ vector of the unknown model parameters to be estimated from the data using the ordinary least squares method. By assumption, X^1 is identically 1, so that the regression equation has an intercept β_0. The error term ε is a $n \times 1$ vector of independent identically distributed errors with common distribution F having mean 0 and finite variance σ^2. Both F and σ^2 are unknown.

3 Statistics of spatial autocorrelation

We run the procedure of bootstrap testing for the Moran's I (Moran, 1950) and Geary's C (Geary, 1954) statistics of spatial autocorrelation. For model (1), their values are calculated as

$$I = \frac{\sum_{i,j=1\ldots n} w_{ij} r_i r_j / \sum_{i,j=1\ldots n} w_{ij}}{\sum_{i=1\ldots n} r_i^2 / n}, \quad (2)$$

$$C = \frac{\sum_{i,j=1\ldots n} w_{ij} (r_i - r_j)^2 / \left(2 \sum_{i,j=1\ldots n} w_{ij}\right)}{\sum_{i=1\ldots n} r_i^2 / n}, \quad (3)$$

where r_i and r_j are residuals in administrative units i and j ($i, j = 1 \ldots n$) and w_{ij} is an element of the $n \times n$ matrix of spatial proximity W. By convention, $w_{ii} = 0$ for any $i = 1 \ldots n$. In the first case, we check whether values of the neighboring residuals are linearly dependent. That is, the Moran's I statistic represents a spatial correlation coefficient; by definition, it equals the covariance of the residuals with themselves taken at the neighboring locations normalized by the estimated population variance in the residuals. Alternatively, Geary's C is a spatial generalization of the von Neumann ratio (von Neumann et al., 1941; Geary, 1954). It equals a ratio between the variance estimated from the differences in the neighboring residuals and the variance measured independently of spatial location. Consequently, the Geary's C statistic determines whether residuals are independent or whether a significant trend (which need not be linear) exists in their values.

3.1 Examples

Examples should be executed using 'regression-tests' library, which can be compiled from the source clojure code using leiningen (http://leiningen.org/).

Example 1: Moran's I calculation (from Paradis (2015) in 2.1 Phylogenetic Distances, pp. 2-4)

REPL

```
(require '[regression-tests.sample-tests :refer :all])

(def values [4.09434 3.61092 2.37024 2.02815 -1.46968])
(def proximity-matrix    {[0 1] 0.505744983336052 [0 2] 0.216747850001166
                          [0 3] 0.171300720162211 [0 4] 0.106206446500571
                          [1 0] 0.505744983336052 [1 2] 0.216747850001166
                          [1 3] 0.171300720162211 [1 4] 0.106206446500571
                          [2 1] 0.304848067656604 [2 0] 0.304848067656604
                          [2 3] 0.240928311535057 [2 4] 0.149375553151735
                          [3 1] 0.276243093922652 [3 2] 0.276243093922652
                          [3 0] 0.276243093922652 [3 4] 0.171270718232044
                          [4 1] 0.25 [4 2] 0.25
                          [4 3] 0.25 [4 0] 0.25})
```

```
(moran-i-test values proximity-matrix)
=> -0.07312179438450675
```

Example 2: Geary's C calculation (from Goodchild (1986) in 1.3.1 Geary's index (area objects, interval attributes), p. 14)

REPL

```
(require '[regression-tests.sample-tests :refer :all])

(def values [3 2 2 1])
(def proximity-matrix    {[0 1] 1 [0 2] 1 [0 3] 1 [1 0] 1 [1 3] 1
                          [2 0] 1 [2 3] 1 [3 0] 1 [3 1] 1 [3 2] 1})

(geary-c-test values proximity-matrix)
=> 6/5
```

4 Hypothesis testing with the bootstrap

4.1 Theoretical background

Generally, the presence of spatial autocorrelation can cause potentially misleading results and consequent misinterpretation of the regression model output (Fotheringham and Rogerson, 1993; Overmars et al., 2003). We examine spatial autocorrelation in the error terms of model (1). That is, we check whether the errors are determined and assigned to the neighboring administrative units independently and at random. Since distribution F is unknown, we use the bootstrap testing (Efron and Tibshirani, 1993) with the null hypothesis of no spatial autocorrelation. The alternative test hypothesis states that the chance of receiving the particular error value in an administrative unit depends on the error values in that unit's neighbors.

We use the observed regression residuals to estimate the true unobserved errors in model (1). Suppose that $\hat{s}$ is a residuals-based test statistic for measuring spatial autocorrelation. At first, we compute $\hat{s}$ in model (1) based on the original data (y, X, W). After that, the data is resampled with replacement k times to get the reference test distribution. That is, in every bootstrap replication γ we draw a random sequence of indexes $(j_1, \dots, j_n)$ from the set $\{1, \dots, n\}$ and compose a $n \times 1$ vector y' and a $n \times (p+1)$ matrix X' by taking the selected pairs $\{(y_{j_1}, X_{j_1}), \dots, (y_{j_n}, X_{j_n})\}$. For the data (y', X', W), we calculate the bootstrap statistic $\hat{s}_\gamma$. In the two-tailed test, we define a probability of obtaining a result equal to or more extreme than the original test statistic $\hat{s}$ as a probability of obtaining a result outside of the equal-tailed interval, where one of the end points coincides with $\hat{s}$. An equal-tailed property means that the probability of a value to be from the left side of an interval is the same as the probability of a value to be from the right side of an interval (Efron and Tibshirani, 1993). In fact, we do both one-tailed

tests and double the lowest p-value from these trials. Formally, the approximate equal-tail p-value in the two-tailed test is given by the formula (Lin et al., 2011)

$$\tilde{p}\text{-value}(\hat{s}) = 2\min\left(\frac{\#\{\gamma = 1 \dots k \mid \hat{s}_\gamma \le \hat{s}\}}{k}, \frac{\#\{\gamma = 1 \dots k \mid \hat{s}_\gamma > \hat{s}\}}{k}\right). \quad (4)$$

After k replications, we arrange a sequence $\hat{s}^*$ by taking bootstrap values $\{\hat{s}_\gamma\}_{\gamma=1\dots k}$ in ascending order. For the given level of confidence α, we find the $\lfloor (1-\alpha)/2\, k \rfloor$ and $\lceil (1+\alpha)/2\, k \rceil$ quantiles in $\hat{s}^*$ and set them as the lower and upper borders of the $100\times\alpha$-% percentile confidence interval respectively. Here, $\lfloor (1-\alpha)/2\, k \rfloor$ denotes the largest integer not greater than $(1-\alpha)/2\, k$ and $\lceil (1+\alpha)/2\, k \rceil$ stands for the smallest integer not less than $(1+\alpha)/2\, k$.

4.2 Implementation in 'regression-tests-1.0.0-standalone.jar'

Usage

$ java -jar regression-tests-1.0.0-standalone.jar [options] path n-replications "iid" path2

Arguments:

path	Path to the csv file with sample data
n-replications	Number of replications in bootstrapping
path2	Path to the additional csv file

Options:

-t, --trace	Print stack trace
-h, --help	Print command help

Input

;;;;; 'sample-x1-x2-x3.csv'

density.asc	land_use.asc	distance_roads.	distance_indus
7.157735	9	0	0.03761
6.326746	9	0	0.032535
-1.977347	0.019694	0.154481	0.244743
5.305789	6.25	0	0.008174
...	...	...	...

;;;;; 'sections-neighbours.csv'

id	n1	n2	n3	n4	n5	n6	n7	...
1	2	361	383	535	541			...
2	1							...
3	4	391						
4	3							...
5	531	534	669					...
7	9	13	16					...
9	7	10	11	13				...
10	9	11	13	14	29	43		...
...	...	...	...	...	...	...	...	...

The 'path' argument defines a file with sample values, e.g., 'sample-x1-x2-x3.csv'. It is expected that the first row contains variable labels. The first column should contain values of the response y.

The 'path2' argument defines a file with the matrix of spatial proximity. It is expected that the first column contains identifiers of administrative units, e.g., ids of sections. The row values contain identifiers of the contiguous administrative units (e.g., ids of sections) to the administrative unit in this row.

NB: Original matrix of spatial proximity will be row-normalized.

Output

Results are saved to the 'regression-tests' folder in the root execution directory. The 'regression-tests/independence-tests-bootstrap.csv' file contains results of the bootstrap hypothesis testing. The generated bootstrap samples are saved to 'regression-tests/morans-i-test-sample.csv' and 'regression-tests/geary-c-test-sample.csv'.

;;;;; 'independence-tests-bootstrap.csv'

statistics	95-percent-ci-1	95-percent-ci-2	mean	p-value
morans-i-test	-0.045375	0.044406	-0.001074	0.323568
geary-c-test	0.941274	1.065335	1.000543	0.741526

;;;;; 'morans-i-test-sample.csv'

value
0.033601
-0.0177
0.022376
0.016133
...

;;;;; 'geary-c-test-sample.csv'

value
0.980722
1.013114
0.960475
0.985381
...

Acknowledgments

The author would like to acknowledge DG research for funding through the FP7-funded COMPLEX project #308601, www.complex.ac.uk.

Views or opinions expressed herein do not necessarily represent those of the International Institute for Applied Systems Analysis, its National Member Organizations, or other organizations supporting the work.

References

[1] Efron, B., & Tibshirani, R. (1993). An Introduction to the Bootstrap. New York: Chapman and Hall.

[2] Fotheringham, S., & Rogerson, P. (1993). GIS and Spatial Analytical Problems. International Journal of Geographical Information Systems, 7(1): 3-19. DOI: 10.1080/02693799308901936.

[3] Geary, R. (1954). The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician, 5(3): 115-145. DOI: 10.2307/2986645

[4] Goodchild, M. (1986). Spatial Autocorrelation, CATMOG 47. Norwich, UK: Geo Books.

[5] Lin, K.-P., Long, Z.-H., & Ou, B. (2011). The Size and Power of Bootstrap Tests for Spatial Dependence in a Linear Regression Model. Computational Economics, 38(2): 153-171. DOI: 10.1007/s10614-010-9224-0

[6] Moran, P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika, 37(1-2): 17-23. DOI: 10.2307/2332142

[7] Paradis, E. (2015) Moran's Autocorrelation Coefficient in Comparative Methods. https://cran.r-project.org/web/packages/ape/vignettes/MoranI.pdf. Accessed 2016.

[8] Overmars, K., de Koning, G., & Veldkamp, A. (2003). Spatial Autocorrelation in Multi-scale Land Use Models. Ecological Modelling, 164: 257-270. DOI: 10.1016/S0304-3800(03)00070-X

[9] von Neumann, J., Kent, R. H., Bellinson, H. R., & Hart, B. I. (1941). The Mean Square Successive Difference. The Annals of Mathematical Statistics, 12(2): 153-162.