The Spatial Error Model

Another Approach for Areal Data Using Maximum Likelihood

Jamie Monogan

University of Georgia

Spatial Data Analysis

Objectives

By the end of this meeting, participants should be able to:

- Estimate, using maximum likelihood, a regression model with a spatial error term.
- Visually present areal data using maps.

Model Specification

- Again, start with: $y_i = \mathbf{x_i}\boldsymbol{\beta} + \varepsilon_i$.
- This time we say: $\varepsilon_i = \lambda \mathbf{w}_{i\cdot}\xi_i + \epsilon_i$.
- By substitution, the full spatial error model is: $y_i = \mathbf{x_i}\boldsymbol{\beta} + \lambda \mathbf{w}_{i\cdot}\xi_i + \epsilon_i$.
- In matrix notation, this gives us:
  - $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \lambda \mathbf{W}\xi + \boldsymbol{\epsilon}$,
  - $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

Estimating the Spatially Lagged y Model with Maximum Likelihood

- Recall: When estimating a linear model with MLE and the Gauss-Markov assumptions are true, our log likelihood function is: $\ln \mathcal{L}(\beta, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{(y-\mathbf{X}\beta)'(y-\mathbf{X}\beta)}{2\sigma^2}$.
- Our log likelihood function for the spatial error model is: $\ln \mathcal{L}(\beta, \sigma^2, \lambda) =$ $\ln|\mathbf{I} - \lambda\mathbf{W}| - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{(y-\mathbf{X}\beta)'(\mathbf{I}-\lambda\mathbf{W})'(\mathbf{I}-\lambda\mathbf{W})(y-\mathbf{X}\beta)}{2\sigma^2}$.
- Which simplifies to: $\ln \mathcal{L}(\beta, \sigma^2, \lambda) = \ln|\mathbf{I} - \lambda\mathbf{W}| - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{(y-\lambda\mathbf{W}y-\mathbf{X}\beta+\lambda\mathbf{W}\mathbf{X}\beta)'(y-\lambda\mathbf{W}y-\mathbf{X}\beta+\lambda\mathbf{W}\mathbf{X}\beta)}{2\sigma^2}$.
- We use Ord's (1975) trick again: Find the eigenvalues of $\mathbf{W}$, $(\omega_1, \ldots, \omega_n)$. This gives us the determinant we need: $|\mathbf{I} - \lambda\mathbf{W}| = \prod_{i=1}^{n}(1 - \lambda\omega_i)$.

Comparing Three Models of Democracy as a function of GDP

Source: Ward & Gleditsch 2008, Table 3.1

	Naïve OLS			Lagged DV			Spatial error		
	Est.	S.E.	t	Est.	S.E.	z	Est.	S.E.	z
Intercept	-9.69	2.43	-3.99	-6.20	2.08	-2.98	-7.49	3.07	-2.44
ln p.c. GDP	1.68	0.31	5.36	0.99	0.28	3.59	1.39	0.38	3.66
$\hat{\rho}$	—			0.56	0.08	7.43	—		
$\hat{\lambda}$	—			—			0.58	0.08	7.60
Log Lik.	-513.62			-491.10			-491.53		

$n = 158$

Choosing Between the Spatial Lag and Spatial Error Models

- Like time series, panel data, and others, there are many ways to fit a model to spatial data.
- Choosing which is a function of assumptions you believe.
- Which is more likely for your data?
  - Is one observation's value of the dependent variable shaped by neighbors' values of the dependent variable?
    - Likely answer: Spatially lagged dependent variable model.
  - Is there a lurking variable in the error term that is likely to be similar among neighbors?
    - Likely answer: Spatial error model.
- Empirically, a true spatial error process and a true autoregressive process are hard to untangle.
- Clarke's (2001) "Testing Nonnested Models of International Relations" (*American Journal of Political Science*) offers some guidance on empirical tests.
- Theory should dominate your view of which model is better, though.

Graphing Areal Data

- It is always good practice to visualize your data.
- Maps of point or lattice observations convey substantial information.
  - What patterns are apparent between variables? (See: V.O. Key.)
  - Are there trends in the data? Is there clustering?
- Usually shading or color saturation is how we convey values of variables on maps.
- Unfortunately, shading and color is pretty far down the list in humans' perceptual accuracy (Cleveland & McGill 1984).
- At least do the best you can with color:
  - Choose two colors, set white as the middle value, and create a bipolar color gradient.
  - You may have to redraw in grayscale for print, but many journals now allow color graphics for the online edition.
- Note: Accurate geographic location is important and conveys a lot of information, too. I personally do not like the idea of resizing areal units based on the value of an area.
  - Consider: Does doing this affect neighbor connections? Does it affect distance between two points?

For Next Time

- Read §2.1-2.3 from Banerjee, Carlin, & Gelfand.
- Download the county-level data set from: Bullock and Hood. 2006. "A Mile-Wide Gap: The Evolution of Hispanic Political Emergence in the Deep South." *Social Science Quarterly* 87(5):1117-1135.
- Subset the data only to the state of Georgia (**state**).
- Estimate a naïve OLS regression of Hispanic registration in 2004 (**hreg04**) as a function of Hispanic registration in 2002 (**hreg02**) and the Hispanic growth rate (**growthrate**).
- Plot a map of the residuals by county. Based on this graph, do the residuals appear to be spatially correlated? Why or why not?
- Choose a spatial weighting scheme and describe your choice.
- Report the results of Geary's $\mathcal{C}$ for the residuals. How do these results compare to your visual analysis?
- Re-estimate your model using a spatially lagged dependent variable and again using a spatial error term. Report all three models in a neat table.
- Which of these models do you believe to be the most sound theoretically? Why?
- Present your results in professional tables and figures. Attach your R code to the back of your final copy.