Spatial Autoregressive Models

Daniel Reiff

November 8, 2018

Motivation

Standard Linear Regression

- $Y_i = X_i\beta + \epsilon_i$
- Each obsevation refers to a location/region
- X is a matrix of covariates
- β is an associated parameter with X
- An observation at one location is independent of an observation at another location
- But, what if our data is not iid . . .

Spatial Dependence

- Oberservations at one location depend on a neighboring location
- $i = 1, j = 2$
- i and j represent neightbors
- $y_i = \alpha_i y_j + x_i \beta + \epsilon_i$
- $y_j = \alpha_j y_i + x_j \beta + \epsilon_j$
- $\epsilon_i \sim N(0, \sigma^2)$
- $\epsilon_j \sim N(0, \sigma^2)$

Applications

- Financial contagion
- Similarities among different campaigns
- Travel times through different regions
- Price of homes in a neighborhood
- Economic decision making

Models for Applications

- Time-dependence motivation
- SAR (Spatial Autoregressive Model)
- Omitted variables motivation
- SDM (Spatial Durbin Model)
- Externalities-based motivation
- SDM
- Model uncertainty motivation
- SEM (Spatial Error Model)

The Models

An Example

- Consider a CBD (R4) surrounded by 3 districts on each side

Spacial Autoregressive Process

$$y_i = \rho \sum_{j=1}^{n} W_{ij} y_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

CBD Example

$$Y = \begin{bmatrix} 42 \\ 37 \\ 30 \\ 26 \\ 30 \\ 37 \\ 42 \end{bmatrix}, X = \begin{bmatrix} 10 & 30 \\ 20 & 20 \\ 50 & 10 \\ 30 & 10 \\ 20 & 20 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} R1 \\ R2 \\ R3 \\ R4 \\ R5 \\ R6 \\ R7 \end{bmatrix}$$

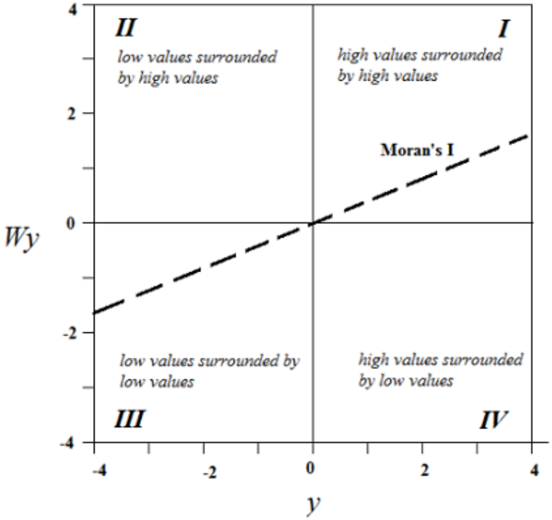
- This pattern violates independence

Adjacency Matrix

$$
W = \begin{array}{c c}
 & \begin{array}{ccccccc} R1 & R2 & R3 & R4 & R5 & R6 & R7 \end{array} \\
\begin{array}{c} R1 \\ R2 \\ R3 \\ R4 \\ R5 \\ R6 \\ R7 \end{array} &
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}
\end{array}, \quad Wy = \begin{bmatrix}
y_2 \\
(y_1 + y_3)/2 \\
(y_2 + y_4)/2 \\
(y_3 + y_5)/2 \\
(y_4 + y_6)/2 \\
(y_5 + y_7)/2 \\
y_6
\end{bmatrix}
$$

Moran Scatterplot

- ρ describes the strength of dependence



The Models

- SAR: $Y = \rho Wy + \alpha\iota_n + X\beta + \epsilon$
- SDM: $Y = \rho Wy + \alpha\iota_n + X\beta + WX\gamma + \epsilon$
- SEM: $Y = \alpha\iota_n + X\beta + u, u = \rho Wu + \epsilon$
- SAC: $Y = \alpha\iota_n + \rho W_1 y + X\beta + u, u = \theta W_2 u + \epsilon$
- SARMA: $Y = \alpha\iota_n + \rho W_1 y + X\beta + u, u = (I_n - \theta W_2)\epsilon$

Maximum Likelihood Estimation

- The Process:
- Concentrate the likelihood wrt β, σ^2, and ϵ
- Substitute closed-form solutions from first order conditions for β and σ^2
- Maximize the concentrated log likelihood wrt ρ

Calculation for SAR

$$lnL = C + ln \mid I_n - \rho W \mid -(\frac{n}{2})ln(e'e)$$

$$e = e_0 - \rho e_d$$

$$e_0 = y - X\beta_0$$

$$e_d = Wy - X\beta_d$$

$$\beta_0 = (X'X)^{-1}X'y$$

$$\beta_d = (X'X)^{-1}X'Wy$$

Calculation for SDM

$$X = [X \quad WX]$$

$$lnL = C + ln \mid I_n - \rho W \mid -(\frac{n}{2})ln(e'e)$$

$$e = e_0 - \rho e_d$$

$$e_0 = y - X\beta_0$$

$$e_d = Wy - X\beta_d$$

$$\beta_0 = (X'X)^{-1}X'y$$

$$\beta_d = (X'X)^{-1}X'Wy$$

Calculation for SEM

$$lnL = C + ln \mid I_n - \rho W \mid -(\frac{n}{2}) ln(e'e)$$

$$\tilde{x} = X - \rho WX$$

$$\tilde{y} = y - \rho Wy$$

$$\beta^\star = (\tilde{x}'\tilde{x})^{-1}\tilde{x}\tilde{y}$$

$$e = \tilde{y} - \tilde{x}\beta^\star$$

Using the MLE

- Using $\hat{\rho}$, we can find other parameters:

$$\hat{\beta} = \beta_0 - \hat{\rho}\beta_d$$

$$\hat{\sigma}^2 = n^{-1} e_0' e_0 - 2\hat{\rho} e_0' e_d + \hat{\rho}^2 e_d' e_d$$

Interpretation of Results

- Statistical Significance of $\hat{\rho}$

$$H_0 : \hat{\rho} = 0$$

$$H_1 : \hat{\rho} \in [0, 1]$$

Impacts

- CBD revisited

$$X = \begin{bmatrix} 10 & 30 \\ 20 & 40 \\ 50 & 10 \\ 30 & 10 \\ 20 & 20 \\ 10 & 30 \end{bmatrix}$$

- $X_{2,2}$ is doubled, how will each region's travel time change?

$$\hat{y}^{(1)} = (I_n - \hat{\rho} W) X \hat{\beta}$$

$$\hat{y}^{(2)} = (I_n - \hat{\rho} W) X \hat{\beta}$$

Impacts

- The impact

$$\begin{bmatrix} & \hat{y^{(1)}} - \hat{y^{(2)}} \\ R1 & 2.57 \\ R2 & 4 \\ R3 & 1.45 \\ R4 & 0.53 \\ R5 & 0.20 \\ R6 & 0.07 \\ R7 & 0.05 \end{bmatrix}$$

- Total Impact, Direct Impact, Indirect Impact

Impacts

- For a linear regression:

$$y = \sum x_r \beta_r + \epsilon$$

$$\frac{\partial y_i}{\partial x_{ir}} = \beta_r$$

$$\frac{\partial y_i}{\partial x_{jr}} = 0$$

- This is due to independence

Impacts

- SDM Model:

$$y = \sum S_r(W)X_r + V(W)\alpha + \epsilon$$

$$Sr(W) = V(Q)(I_n\beta_r + W\theta_r)$$

$$V(W) = (I_n - \rho W)^{-1}$$

- A change in the explanatory variable for a region can potentially affect DVs in other regions

$$\frac{\partial y_i}{\partial x_{ir}} = Sr(W)_{ij}$$

Impacts

- SAR Model:

$$\frac{\partial y}{\partial x_r'} = I_n\beta_r + W\rho\beta_r + [W^2\sigma^2\beta_r + W^3\sigma^3\beta_r + ...]$$

- Direct effects on DV

$$\bar{M}(r)_{direct} = n^{-1}tr(S_r(W))$$

- Average of the diagonal
- Total effects on DV

$$\bar{M}(r)_{total} = n^{-1}\iota_n' S_r(W)\iota_n$$

- Average of row sums
- Indirect effects on DV

$$\bar{M}(r)_{indirect} = \bar{M}(r)_{total} - \bar{M}(r)_{direct}$$

Issues

$$\frac{\partial y_i}{\partial x_{ir}} = Sr(W)_{ii}$$

- Effect of feedback loops - Magnitude of feedback error depends on: - Postion of regions in space - Degree of connectivity among region - The parameters

Application in R

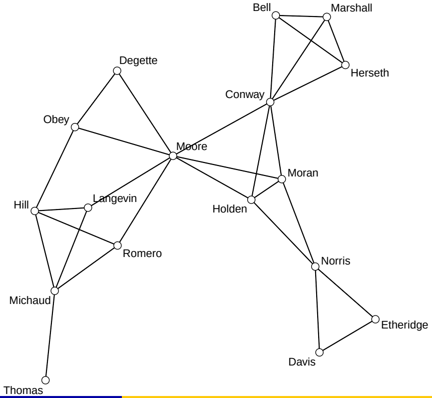
SAR Example from Professor Montgomery

- Are campaigns more likely to adopt startegies that are also being used by other campaigns with chich they share consultants?

$$Y = X\beta + \rho Wy + \epsilon$$

Adjacency Matrix

$$W_{c,d} = \sum_{j=1}^{s} I(e_{c,j} \geq 25,000) * I(e_{d,j} \geq 25,000)$$



Execution

```
selector <- c(!is.na(DruckData.Reduced$risk) &
                !is.na(DruckData.Reduced$newdist))
this.mod <- (risk ~ y2004 + y2006 + democrat +  open
+ chall + factor(newdistAlt)+ factor(region))
clean_data <- DruckData.Reduced[selector,]
adj.matrix=adjMatShared
startlist<-adj.matrix
startlist0<-startlist[selector, selector]
final.list<-mat2listw(startlist0, style="W")
summary(lagsarlm(formula = this.mod, data = clean_data,
                 listw = final.list, zero.policy = TRUE))
```

Results

```
Coefficients: (asymptotic standard errors)
                    Estimate Std. Error z value  Pr(>|z|)
(Intercept)        -0.397006   0.333518 -1.1904   0.23391
y2004               0.118521   0.185066  0.6404   0.52189
y2006               0.335079   0.178581  1.8763   0.06061
democrat            1.210873   0.141266  8.5716 < 2.2e-16
open                1.601629   0.224069  7.1479 8.811e-13
chall               2.716738   0.173431 15.6647 < 2.2e-16
factor(newdistAlt)1  0.066410   0.191578  0.3466   0.72886
factor(newdistAlt)2 -0.086343   0.208824 -0.4135   0.67926
factor(newdistAlt)3 -0.117252   0.231889 -0.5056   0.61311
factor(region)1     0.198431   0.311355  0.6373   0.52392
factor(region)2     0.728852   0.289136  2.5208   0.01171
factor(region)3     0.473485   0.327815  1.4444   0.14864
factor(region)4     0.446075   0.276697  1.6121   0.10693
factor(region)5     0.452145   0.306912  1.4732   0.14070
```

Results

```
Rho: 0.10517, LR test value: 5.1114, p-value: 0.02377
Asymptotic standard error: 0.047895
    z-value: 2.1958, p-value: 0.028105
Wald statistic: 4.8216, p-value: 0.028105

Log likelihood: -959.4322 for lag model
ML residual variance (sigma squared): 2.2766, (sigma: 1.5088)
Number of observations: 524
Number of parameters estimated: 18
AIC: 1954.9, (AIC for lm: 1958)
LM test for residual autocorrelation
test value: 2.3277, p-value: 0.12709
```