mwp-296

intro 1 - SAR モデルの紹介　【 評価版 】

本 whitepaper では空間自己回帰モデル (SAR: spacial autoregressive models) の概要について紹介します。

1. SP 系コマンド

[SP] マニュアルに記載されている SP 系コマンドは

1. 従属変数の空間的ラグ
2. 独立変数の空間的ラグ
3. 空間的に自己回帰な誤差

を伴うモデルをフィットさせることができます。これらの空間的特性に関する機能は任意の組合せで利用できます。

2. SAR モデルの概要

SAR モデルは空間単位 (spatial units) — 国とか地域、あるいは地理とは無関係な social network nodes 等 — に関する観測データを含むデータセットを対象にフィットが行われます。ここでは簡単のためにこれらの空間単位をエリアと称することにします。データセット中には少なくとも 1 つの連続的なアウトカム変数 — 疾病の発生数、企業の生産量、犯罪発生率、等 — と、そのアウトカムを予測するものとされるその他の変数が存在するものとします。そのようなデータセットに対しては次のような形の線形回帰をフィットさせることが可能です。

$$y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,k}\beta_k + \epsilon_i \tag{1}$$

ただしこの線形回帰は単なる出発点として提示されているものであって、SAR モデルと呼べるものではありません。しかしこの出発点に空間的な色合いを付加する意味で、観測データ (observations) のことをエリアと呼ぶことにします。また変数にはエリアの特性に関する情報が含まれているものとします。このようなコンテクストでこのモデル式に注釈を付加するなら次のようになります。

i	エリア（観測データ）の番号 $1, \ldots, N$
y_i	エリア i における従属（アウトカム）変数
$x_{i,1}$	エリア i における 1 番目の独立変数
$\vdots$	
$x_{i,j}$	エリア i における j 番目の独立変数
$\vdots$	
$x_{i,k}$	エリア i における最後の独立変数
ϵ_i	エリア i における誤差（残差）

この線形回帰モデルを列ベクトルを使った表記で表すと次のようになります。

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_k\mathbf{x}_k + \boldsymbol{\epsilon} \tag{2}$$

ただし太字で表記された変数は $N \times 1$ ベクトルを意味します。このモデルを `regress` コマンドでフィットさせようと思ったら次のように入力することになります。

```
. regress y x1 x2 ... xk
```

SAR モデルはこの線形回帰を拡張し、1 つのエリアのアウトカムが

1. 近隣エリアのアウトカム
2. 近隣エリアからの共変量
3. 近隣エリアからの誤差

によって影響されることを許容しようとするものです。専門用語を使って表現するなら、SAR モデルは

1. アウトカム変数の空間的ラグ
2. 共変量の空間的ラグ
3. 空間的な自己回帰誤差

を許容するものと言えます。

これらの用語は時系列の文献に由来するものです。時系列の分野において自己回帰的な (autoregressive) AR(1) 過程は

$$y_t = \gamma_0 + \gamma_1 y_{t-1} + \epsilon_t \tag{3}$$

のように表現されます。ここで y_{t-1} は y のラグ項と呼ばれます。ラグ演算子 $L.$ を用いたベクトル表記で表すなら

$$\mathbf{y} = \gamma_0 + \gamma_1 L.\mathbf{y} + \boldsymbol{\epsilon} \tag{4}$$

のようになります。

誤差項が自己回帰的なものであるときの AR(1) 過程は

$$\begin{aligned}&\mathbf{y} = \gamma_0 + \gamma_1 L.\mathbf{y} + \mathbf{u}\\ &\text{ただし } \mathbf{u} = \rho L.\mathbf{u} + \boldsymbol{\epsilon}\end{aligned} \tag{5}$$

のように表現されることになります。この場合、(5) 式は

$$\mathbf{y} = \gamma_0 + \gamma_1 L.\mathbf{y} + (\mathbf{I} - \rho L.)^{-1}\boldsymbol{\epsilon} \tag{6}$$

のように表現することができます。パラメータ ρ は誤差中における相関の度合いを示すもので、γ_0, γ_1 と共に推定の対象となります。

このような時系列分野における表記法や用語は空間的な分野にも移植できます。この場合、ラグ演算子は $N \times N$ 行列 $\mathbf{W}$ となります。すなわち $L.\mathbf{y}$ だったものは $\mathbf{Wy}$ — ベクトル $\mathbf{y}$ に行列 $\mathbf{W}$ をかけたもの — のようになるわけです。従って (4) 式に対応した SAR モデルは次のように表現されることになります。

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{Wy} + \boldsymbol{\epsilon} \tag{7}$$

また自己回帰誤差を伴う時系列のモデル式 (6) に対応する SAR モデルは

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{Wy} + (\mathbf{I} - \rho \mathbf{W})^{-1}\boldsymbol{\epsilon} \tag{8}$$

のように記述されます。

$\mathbf{W}$ は空間重み付け行列 (spatial weighting matrix) と呼ばれます。行列中の値はエリア間の空間的な関係を特徴付けます。$\mathbf{W}$ は時間と空間の違いはあるにしろ、$L.\mathbf{y}$ に類似したものと言えます。$L.\mathbf{y}$ は時点 $t-1$ から時点 t への潜在的波及 (potential spillover) の度合いを計測するものであるのに対し、要素 W_{i_1,i_2} はエリア i_2 からエリア i_1 への潜在的波及がどのくらいあるかを規定するものです。エリア i_2 が i_1 に対し何ら影響を及ぼさないのであれば W_{i_1,i_2} は 0 となります。潜在的波及が大きければ大きいほど W_{i_1,i_2} の値は大きくなります。なお、$\mathbf{W}$ の要素はモデルフィットに先立ち指定されるものとします。

評価版では割愛しています。

■