面数据建模一些理论上的补充

金百锁

2017 年 3 月 26 日

1. 空间相关性

全局空间相关性

Moran指数I是最早应用于全局聚类检验的方法。它检验整个研究区中邻近地区间是相似、相异（空间正相关、负相关），还是相互独立的。Moran指数I的计算公式如下：

$$I = \frac{n\sum_i\sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i\neq j} w_{ij}\sum_i (Y_i - \bar{Y})^2}$$

这里n为研究区域内地区总数，w_{ij}是空间权重（例如当区域i和区域j相邻时$w_{ij} = 1$，否则$w_{ij} = 0$）。

I的取值一般在-1到1之间，大于0正相关，小于0负相关，接近0不存在空间自相关。采用delta方法，如果Y_i是独立同分布的，则当n很大时，I的均值近似为$-1/(n-1)$，方差近似为

$$Var(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}$$

这里$S_0 = \sum_{i \neq j} w_{ij}$，$S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2$，$S_2 = \sum_k (\sum_j w_{kj} + \sum_i w_{ik})^2$。

与Moran指数I相似，Geary指数C也是全局聚类检验的一个指数。计算Moran指数I时，用的是中值离差的叉乘，但是，Geary指数C强调的是观察值之间的离差，其公式为：

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{\sum_{i \neq j} w_{ij} \sum_i (Y_i - \bar{Y})^2}$$

Geary指数C的取值一般在0到2之间，大于1表示负相关，等于1表示不相关，而小于1表示正相关。

局部空间自相关

Anselin提出了一个局部Moran指数，或称LISA（local indicator of spatial association），用来检验局部地区是否存在相似或相异的观察值聚集在一起。区域i的局部Moran指数用来度量区域i和它邻域之间的关联程度，定义为：

$$I_i = \frac{n(Y_i - \bar{Y}) \sum_{j \neq i}^{n} w_{ij}(Y_j - \bar{Y})}{\sum_j (Y_j - \bar{Y})^2}$$

正的I_i表示高值被高值包围，或低值被低值包围。负的的I_i表示高值被低值包围，或低值被高值包围。

类似的，Getis和Ord开发了一个Geary指数的局部聚类检验，称之为*Gi*指数，用来检验局部地区是否存在统计显著的高值或低值，定义为：

$$G_i = \frac{\sum_{j \neq i}^{n} w_{ij} Y_j}{\sum_{j \neq i}^{n} Y_j}.$$

这个指数用来检验局部地区是否有高值或低值在空间上趋于集聚。高的*Gi*值表示高值的样本集中在一起，而低的*Gi*值表示低值的样本集中在一起。*Gi*指数还可用于回归分析中的空间滤值处理，解决空间自相关问题。

相邻距离

通过空间中的相对位置定义相邻，相邻用“1”表示，不相邻“0”表示。常用的相邻关系有如下几种：线性相邻（linear contiguity），区域*i*和区域*j*在左侧或右侧有共同的边；“车”相邻（rook contiguity），区域*i*和区域*j*有共同的边；“象”相邻（bishop contiguity），区域*i*和区域*j*有共同的顶点但没有共同的边；“后”相邻（queen contiguity），区域*i*和区域*j*有共同的顶点或共同的边。

另外一种以距离阀值定义权重（在阀值范围内定义1，在阀值范围外定义0）。具体设定方法：d_{ij}表示两个区域（不一定相邻）之间的欧式距离，$d_{\max i}$表示最大空间相关距离，对于区域i若：$d_i \leq d_{\max i}$，则$w_{ij} = 1$；否则$w_{ij} = 0$。除了阀值还可以选择k个最邻近区域作为相邻关系，例如$k = 1$是选择最小距离的邻近区域。

除了欧式距离还有一种叫做负指数距离，具体设定为$\exp(-\beta d_{ij})$，d_{ij}表示两个区域之间的欧氏距离，β为预先设定的参数。

空间权重矩阵的标准化，令$W=(w_{ij})$，$\sum_j w_{ij}=w_{i+}$，则$\tilde{W}=(w_{ij}/w_{i+})$为标准化后的空间权重矩阵，即行和为1，但$\tilde{W}$不对称。

2. 模型

空间自回归模型（Spatial autoregressive models,SAR）

令W为空间权重矩，空间自回归模型如下：

$$Y = \rho WY + X\beta + \epsilon.$$

采用极大似然估计方法估计参数。对数似然函数为

$$\frac{1}{2}\log|\sigma^{-1}(I-\rho W)| - \frac{1}{2\sigma^2}(Y-\rho WY-X\beta)^T(Y-\rho WY-X\beta).$$

R语言采用spdep包中的spautolm函数估计。

空间误差模型(Spatial error model,SEM)

令W为空间权重矩，同步自回归模型如下：

$Y = X\beta + u, u = \rho W u + \epsilon,$ 即 $Y = \rho WY + (I - \rho W)X\beta + \epsilon.$

如果$\epsilon \sim N(0, \sigma^2 I)$，则

$$Y \sim N(X\beta, \sigma^2 (I - \rho W)(I - \rho W)^T).$$

采用极大似然估计方法估计参数。对数似然函数为

$$\frac{1}{2}\log|\sigma^{-1}(I-\rho W)| - \frac{1}{2\sigma^2}(Y-X\beta)^T(I-\rho W)(I-\rho W)^T(Y-X\beta).$$

R语言采用spdep包中的spautolm函数估计。

空间杜宾模型（Spatial Dubin model,SDM）

空间杜宾模型是SAR和SDM的混合，令$\pi_a + \pi_b = 1$，令

$$
\begin{aligned}
Y_a &= (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\epsilon \\
Y_b &= X\beta + (I - \rho W)^{-1}\epsilon
\end{aligned}
$$

令$I-\rho W=R$，则

$$
\begin{aligned}
Y_c &= \pi_a y_a + \pi_b y_b \\
Y_c &= R^{-1} X \pi_a \beta + \pi_b X \beta + R^{-1}\epsilon \\
RY_c &= X\pi_a\beta + R\pi_b X\beta + \epsilon \\
Y_c &= X\beta + WX(-\rho\pi_b\beta) + \epsilon \\
Y_c &= \rho W y_c + X\beta + WX(-\rho\pi_b\beta) + \epsilon
\end{aligned}
$$

令$-\rho\pi_b\beta = \theta$，则SDM模型为

$$Y = \rho Wy + X\beta + WX\theta + \epsilon.$$

SDM模型也可以写为SAR模型

$$\begin{aligned} Y &= \rho Wy + X\beta + WX\theta + \epsilon \\ &= \rho Wy + Z\delta + \epsilon \end{aligned}$$

这里$Z = (X, WX)$，$\delta = (\beta^T, \theta^T)^T$。

SAR模型的Bayes估计方法

令$A = I - \rho W$有

$$p(D|\beta,\sigma,\rho) = (2\pi\sigma^2)^{-\frac{n}{2}}|A|\exp[-\frac{1}{2\sigma^2}(Ay-X\beta)^T(Ay-X\beta)]$$

令β,σ,ρ的先验为

$$\begin{aligned}\pi(\beta|\sigma^2) &= N(c,\sigma^2 T)\\ \pi(\sigma^2) &= IG(a,b) = \frac{b^a}{\Gamma(a)}\sigma^{-2(a+1)}\exp(-b/\sigma^2)\\ \pi(\rho) &= U(\lambda_{max}^{-1},\lambda_{min}^{-1})\end{aligned}$$

这里$\lambda_{max},\lambda_{min}$为空间权重矩阵的最大最小特征根。

采用MCMC算法估计参数

1.采用如下条件分布进行Gibbs抽样

$$
\begin{aligned}
p(\beta|\rho,\sigma^2) &= N((X^TX+T^{-1})^{-1}(X^TAy+T^{-1}c), \\
&\quad \sigma^2((X^TX+T^{-1})^{-1})) \\
p(\sigma^2|\beta,\rho) &= IG(a+n/2, b+(Ay-X\beta)^T(Ay-X\beta)/2)
\end{aligned}
$$

2.采用Metropolis-Hastings抽样方法抽取ρ，ρ后验分布为

$$p(\rho|\beta,\sigma) \propto |I-\rho W| \exp[-\frac{1}{2\sigma^2}(Ay-X\beta)^T(Ay-X\beta)]$$

具体做法如下：

1. 从提议分布(proposal distribution)$g(\cdot|\rho_t)$产生一个候选值ρ'
2. 计算接受概率

$$\alpha(\rho_t,\rho') = \min\{1, \frac{p(\rho'|\beta,\sigma)g(\rho_t|\rho')}{p(\rho_t|\beta,\sigma)g(\rho'|\rho_t)}\}$$

3. 依概率$\alpha(\rho_t,\rho')$接受$\rho_{t+1}=\rho'$，否则$\rho_{t+1}=\rho_t$。

提议分布(proposal distribution)$g(\cdot|\rho_t)$的选择要使得产生的马尔科夫链满足不可约、正常返、非周期性且具有平稳分布p等正则化条件。

3. 时空自回归模型（spatio-temporal autoregressive models，STAR）

$$
\begin{aligned}
y_t &= G y_{t-1} + X_t \beta + \nu_t \\
X_t &= \varphi^t X_0 \\
G &= \tau I + \rho W \\
d_t &= X_t \gamma \\
\nu_t &= r + d_t + \varepsilon_t
\end{aligned}
$$

这里τ代表每个区域在时间t和$t-1$之间的依赖，ρ代表每个区域在时间t和邻近区域$t-1$之间的依赖，φ表示解释变量以每期不变的速度φ增长。r服从$N(0,\sigma_r^2 I_n)$表示与X_t无关的遗漏变量，$d_t = X_t\gamma$表示和X_t相关的遗漏变量，γ反应相关强度，ε_t为随机误差服从$N(0,\sigma_\varepsilon^2 I_n)$ 独立于X_t和r。

通过递归关系

$$
\begin{aligned}
y_t &= (I_n\varphi^t + G\varphi^{t-1} + \cdots + G^{t-1}\varphi)X_o\beta + G^t y_0 + z \\
z &= z_1 + z_2 + z_3 \\
z_1 &= (I_n + G + \cdots + G^{t-1})r \\
z_2 &= (I_n\varphi^t + G\varphi^{t-1} + \cdots + G^{t-1}\varphi)X_o\gamma \\
z_3 &= \varepsilon_t + G\varepsilon_{t-1} + G^2\varepsilon_{t-2} + \cdots + G^{t-1}\varepsilon_1
\end{aligned}
$$

如果当$t \to \infty$时，$G^t \to 0, G^t\varphi^{-t} \to 0$则

$$
\begin{aligned}
E(y_t) &\approx (I_n\varphi^t + G\varphi^{t-1} + \cdots + G^{t-1}\varphi)X_o(\beta+\gamma) \\
&\approx (I_n - G\varphi^{-1})^{-1}X_t(\beta+\gamma) \\
&\approx (I_n - \frac{\rho}{\varphi-\tau}W)^{-1}\frac{\varphi(\beta+\gamma)}{\varphi-\tau}X_t
\end{aligned}
$$

$$
\begin{aligned}
Var(y_t) &\approx E(z_1 z_1^T) + E(z_3 z_3^T) \\
&\approx (I_n + G + G^2 + \cdots)^2 \sigma_r^2 \\
&\quad + (I_n + G^2 + G^4 + \cdots)\sigma_\varepsilon^2 \\
&\approx (I_n - G)^{-2}\sigma_r^2 + [(I - G)(I + G)]^{-1}\sigma_\varepsilon^2.
\end{aligned}
$$

时空模型和SAR模型的关系

时空模型和如下的基于时间t的截面空间自回归模型的期望相同，即

$$y_t = \rho^* W y_t + X_t \beta^* + \xi_t$$

这里ξ_t为干扰项，$\rho^* = \frac{\rho}{\varphi - \tau}$，$\beta^* = \frac{\varphi}{\varphi - \tau}(\beta + \gamma)$。即经过长时间迭代，把$\beta + \gamma$放大$\varphi(\varphi - \tau)^{-1}$倍得到$\beta^*$。$\tau$为自回归参数，$\varphi$为控制趋势增长趋势参数。如果$\varphi = 1$解释变量没有增长，将会得到经典的时间乘数$(1 - \tau)^{-1}$。

ρ为空间依赖参数，ρ^*为长期空间乘数，如果$\varphi > 1$，X的增长减少了系统的空间依赖，而$\varphi < 1$则给过去值更大的权重，允许更长时间来实现空间影响。

截面空间模型会导致较高的空间依赖估计值，而时空回归会产生较高的时间依赖和较低的空间依赖。截面空间模型是估计和推断长期均衡值。而时空模型是估计和推断时间动态上的参数值。

时空模型和误差模型（SEM模型）的关系

时空SEM模型：

$$
\begin{aligned}
y_t &= G(y_{t-1} - X_{t-1})\beta + X_t\beta + \nu_t \\
X_t &= \varphi^t X_0 \\
G &= \tau I + \rho W \\
d_t &= X_t\gamma \\
\nu_t &= r + d_t + \varepsilon_t
\end{aligned}
$$

通过递归关系

$$
\begin{aligned}
y_t &= \varphi^t X_o \beta + G^t (y_0 - X_0 \beta) + z \\
z &= z_1 + z_2 + z_3 \\
z_1 &= (I_n + G + \cdots + G^{t-1}) r \\
z_2 &= (I_n \varphi^t + G\varphi^{t-1} + \cdots + G^{t-1}\varphi) X_o \gamma \\
z_3 &= \varepsilon_t + G\varepsilon_{t-1} + G^2 \varepsilon_{t-2} + \cdots + G^{t-1}\varepsilon_1
\end{aligned}
$$

如果当$t \to \infty$时，$G^t \to 0, G^t\varphi^{-t} \to 0$则

$$E(y_t) \approx X_t\beta + (I_n - \frac{\rho}{\varphi - \tau}W)^{-1}\frac{\varphi\gamma}{\varphi - \tau}X_t.$$