异质性与内生性同行效应模型及Stata应用

王群勇（南开大学数量经济研究所 & 经济行为与政策模拟实验室）

2024年8月20号，南开大学，天津

内容

同行效应

异质性同行效应: `snreghnet`

内生性同行效应: `snregenet`

同行效应

linear-in-means model：

$$y_i = \alpha + \gamma x_i + \delta \frac{1}{d_i} \sum_{j \in N_i} x_j + \lambda \frac{1}{d_i} \sum_{j \in N_i} y_j + \epsilon_i$$

其中，d_i为i的邻居数量。

γ：个体效应（individual effect）

δ: 情景效应（contexture effect）

λ：内生同行效应（endogenous peer effect）

设x_i包含k个变量，模型共包含$2k+2$个参数。

严格外生: $E(\epsilon_i|x, G) = 0$，模型中x与G是外生的。

内生: $E(\epsilon_i|x, G) \neq 0$?

关系矩阵

$$y = \alpha 1 + \gamma x + \delta G x + \lambda G y + \epsilon.$$

$$G = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & 0 & a_{13} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 0 \end{bmatrix}$$

权重矩阵的元素为0或1，表示是否存在连接。

如果权重矩阵做归一化（行和为1），即相当于同行的均值的影响（linear-in-means)。权重矩阵本身相当于同行的加总的影响（称之为Linear-in-sums模型）。

权重矩阵的不同的规范化得到不同的结果，具有不同的含义。比如，吸烟行为的同行效应。个体i有一个朋友，每天吸10支烟。个体j有两个朋友，每个朋友每天吸10支烟。

其他统计量 (linear-in-*)

可以采用其他指标（中位数、标准差、最小值、最大值、分位数等）：

$$y = \alpha 1 + \gamma x + \delta s(G, x) + \lambda s(G, y) + \epsilon.$$

其中，$s(G, x)$, $s(G, y)$表示x和y的统计指标（标准差、最大值、最小值、分位数等）。

同行效应

模型

$$y = \alpha 1 + \gamma x + \lambda G y + \epsilon, \tag{1}$$

$$y = (I - \lambda G)^{-1}(\alpha + \gamma x + \epsilon).$$

$$E(y|x) = \alpha(I - \lambda G)^{-1} 1 + \gamma(I - \lambda G)^{-1} x,$$

令$C_k = \gamma_k(I - \lambda G)^{-1}$，$C_{ij}$体现个体$j$的$x_k$对个体$i$的$E(y|x)$的边际效应；

第i行体现其他个体对个体i的边际影响，第j列体现个体j对其他个体的边际影响。i行的均值体现其他个体对个体i的平均边际效应，j列的均值体现个体j对其他个体的平均边际效应。

C_k的非对角线元素的均值即为x_k的平均间接效应。

当权重矩阵的行（列）和为1时，C_k的各行（列）之和为常数。

同行效应的工具变量估计

由$(I-\lambda G)^{-1}=I+\lambda G+\lambda^2G^2+\cdots$，可得，

$$\begin{aligned}E(y|x)&=\alpha(I-\lambda G)^{-1}1+\gamma(I-\lambda G)^{-1}x\\&=\tilde{\alpha}1+\tilde{\gamma}_0x+\tilde{\gamma}_1Gx+\tilde{\gamma}_2G^2x+\tilde{\gamma}_3G^3x+\cdots\end{aligned}$$

因此，$(Gx,G^2x,G^3x,\cdots)$是Gy的合适的工具变量。G^kx表示k阶同行的x的均值。

类似地，对于模型

$$y=\alpha1+\gamma x+\delta Gx+\lambda Gy+\epsilon, \tag{2}$$

$$E(y|x)=\tilde{\alpha}1+\tilde{\gamma}_0x+\tilde{\gamma}_1Gx+\tilde{\gamma}_2G^2x+\tilde{\gamma}_3G^3x+\cdots$$

因此，$(G^2x,G^3x,\cdots)$是Gy的合适的工具变量。G^kx表示k阶同行的x的均值。

定理：设网络G是连接的，且$\delta+\gamma\lambda\neq0$，那么模型（1）是可识别的当且仅当$(I,G,G^2)$线性独立。模型（2）是可识别的当且仅当$(I,G,G^2,G^3)$线性独立。

极大似然估计

其它有效的工具变量：二阶同行的x的和（或方差等统计量）。

极大似然估计:

$$ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Omega| - \frac{1}{2}\epsilon'\Omega^{-1}\epsilon.$$

其中，$\Omega = \sigma^2(I - \lambda G)^{-1}(I - \lambda G')^{-1}$.

只有外生变量的加权均值的情况，可以直接用OLS估计：

$$y = \alpha 1 + \gamma x + \delta Gx + \epsilon.$$

内容

同行效应

异质性同行效应: `snreghnet`

内生性同行效应: `snregenet`

异质性同行效应

Beugnot et al. (2019), (Bramoullé 2013; Arduini et al. 2019a,b)：

- 男性与女性对其他人的影响不同；
- 受到男同行的影响与女同行的影响不一样。

异质性同行效应（BP模型）

设$\Lambda = \mathrm{diag}(\lambda_0 + \lambda_1 z_i)$表示由向量构成的对角矩阵，第一种形式（列异质性）

$$y = \alpha[I - \Lambda G]^{-1}1 + \gamma x + \epsilon.$$

其中，ΛG是将向量$(\lambda_0 + \lambda_1 z_i)$分别与$G$的每一列进行元素相乘。$\lambda_1$衡量了个体对同行的外溢效应的大小。$\lambda_1$越大，表明个体对同行的影响越大（或者对同行越有用）。

第二种形式（行异质性）

$$y = \alpha[I - G\Lambda]^{-1}1 + \gamma x + \epsilon.$$

其中，$G\Lambda$是将向量$(\lambda_0 + \lambda_1 z_i)$分别与$G$的每一行进行元素相乘。$\lambda_1$衡量了个体受到同行的影响，$\lambda_1$越高，表明个体越容易受到同行的影响（或者越能充分利用同行效应）。

如果$\lambda = 0$，模型退化为同质模型。

异质性同行效应（BLP模型）

列异质性：

$$y = (I - \Lambda G)^{-1}(\alpha + \gamma x + \epsilon).$$

行异质性：

$$y = (I - G\Lambda)^{-1}(\alpha + \gamma x + \epsilon).$$

附：集中对数似然函数

线性回归模型：

$$y = x\beta + u, \epsilon \sim N(0, \sigma^2 I_n).$$

y_i的似然函数（概率密度函数）为：

$$f(y_i|\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - x_i\beta)'(y_i - x_i\beta)}{2\sigma^2}\right).$$

$(y_1,\cdots,y_n)$的对数似然函数为：

$$lnf(y|\beta,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{(y - x\beta)'(y - x\beta)}{2\sigma^2}.$$

将$\sigma^2 = \epsilon'\epsilon/n$带入,可得：

$$lnf(y|\beta) = -\frac{n}{2}\ln\left(2\pi\frac{\epsilon'\epsilon}{n}\right) - \frac{\epsilon'\epsilon}{2\epsilon'\epsilon/n} = -\frac{n}{2}\ln(2\pi+1) - \frac{n}{2}\ln(\sigma^2).$$

ML估计

一般的空间回归模型：

$$y = x\beta + W_x x\gamma + \lambda W_y y + u = X_a\xi + \lambda W_y y + u$$

$$u = \rho W_u + \epsilon, \epsilon \sim N(0, \sigma^2 I_n).$$

其中，$X_a = (x, W_x x)$，$\xi = (\beta', \gamma')'$.

简化模型为：

$$y = (I_n - \lambda W_y)^{-1} X_a\xi + (I_n - \lambda W_y)^{-1}(I_n - \rho W_u)^{-1}\epsilon.$$

$\epsilon = (I_n - \rho W_u)[(I_n - \lambda W_y)y - X_a\xi]$，似然函数为：

$$lnf(y|\xi, \lambda, \rho, \sigma^2) = -\frac{n}{2} - \frac{n}{2}\ln(\sigma^2) + \ln|I_n - \lambda W_y| + \ln|I_n - \rho W_u| - \frac{1}{2\sigma^2}\epsilon'\epsilon$$

ML估计

如果已知(λ,ρ),

$$(I_n-\rho W_u)(I_n-\lambda W_y)y=(I_n-\rho W_u)X_a\xi+\epsilon,$$

给定(λ,ρ), ξ的ML估计量为

$$\hat{\xi}=[X_a'(I_n-\rho W_u)'(I_n-\rho W_u)X_a]^{-1}X_a'(I_n-\rho W_u)'(I_n-\rho W_u)y,$$

$$\hat{\sigma}^2=\frac{\epsilon'\epsilon}{n},$$

将其带入似然函数，得到集中对数似然函数，

$$lnf_c(y|\lambda,\rho)=-\frac{n}{2}[\ln(2\pi)+1]-\frac{n}{2}\ln(\sigma^2(\lambda,\rho))+\ln|I_n-\lambda W_y|+\ln|I_n-\rho W_u|.$$

对lnf_c求极值，得到$(\hat{\lambda},\hat{\rho})$，进而得到$(\hat{\xi},\hat{\rho})$.

利用格点搜索法设置(λ,ρ)的初始值。

SAR的ML估计

对于SAR模型，给定λ, ξ的ML估计量为

$$\hat{\xi} = (X_a' X_a)^{-1} X_a' (I_n - \lambda W_y) y$$

$$\hat{\sigma}^2 = \frac{1}{n} [(I_n - \lambda W_y) y - X_a \hat{\xi}(\lambda, \rho)]' [(I_n - \lambda W_y) y - X_a \hat{\xi}(\lambda, \rho)]$$

将其带入似然函数，得到集中对数似然函数，

$$lnf_c(y|\lambda, \rho) = -\frac{n}{2} [\ln(2\pi) + 1] - \frac{n}{2} \ln(\sigma^2(\lambda)) + \ln |I_n - \lambda W_y|.$$

对lnf_c求极值，得到$\hat{\lambda}$，进而得到$\hat{\xi}$.

利用格点搜索法设置λ的初始值。

存在孤点情况下的SAR估计

如果存在部分孤点，设SAR的$W_y = G$，可以分解为

$$G = \begin{pmatrix} G_1 & 0 \\ 0 & 0 \end{pmatrix}, (I - \lambda G)^{-1} = \begin{pmatrix} (I - \lambda G_1)^{-1} & 0 \\ 0 & I \end{pmatrix}.$$

设对应的样本量分别为(n_1, n_0)，残差分别为(ϵ_1, ϵ_0)，似然函数为(n_1, n_0)两部分的似然函数之和，

$$\begin{aligned}\ln(L) = & -\frac{n_1}{2}[\ln(2\pi) + 1] - \frac{n_1}{2}\ln(\sigma^2(\lambda)) + \ln|I_{n_1} - \lambda G_1| \\ & - \frac{n_0}{2}[\ln(2\pi) + 1] - \frac{n_0}{2}\ln(\sigma^2).\end{aligned}$$

异质性同行效应模型: Stata

`snreghnet` *varlist* [*if*] [*in*] , [`rowx` (*varname*) `colx` (*varname*) `wmata` (*matrix*) `vce` (*string*) `bs` (50) `vbias` (*varname*) `nolog` `level` (95)]

```
. regeneth les gender nowhite seniority nchair, xin(gender nowhite) wmata(W) bs(0) nolog

Regression with heterogenous social network:

log-likelihood  =       -609.5367               Number of obs    =        416
AIC             =       1233.0734               R-squared        =     0.2331
BIC             =       1261.2882               Adj R-squared    =     0.2200
HQIC            =       1244.2295               Root MSE         =     1.0575
------------------------------------------------------------------------------
         les | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      gender |  -.3001238   .2022444    -1.48   0.138    -.6965157     .096268
     nowhite |  -.2601256   .2648866    -0.98   0.326    -.7792938    .2590426
   seniority |   .0210623   .0121537     1.73   0.083    -.0027586    .0448833
      nchair |   2.519157   .2527532     9.97   0.000      2.02377    3.014544
       _cons |    .877124    .101598     8.63   0.000     .6779955    1.076252
-------------+----------------------------------------------------------------
xin          |
      gender |    .083848   .1442002     0.58   0.561    -.1987791    .3664751
     nowhite |   -.091395   .1728883    -0.53   0.597    -.4302499    .2474599
       _cons |    .007637   .0597171     0.13   0.898    -.1094064    .1246803
------------------------------------------------------------------------------
```

内容

同行效应

异质性同行效应: `snreghnet`

内生性同行效应: `snregenet`

具有内生社会网络的回归模型

模型(Battaglini, M., Sciabolazza, V.L. and Patacchini, E., 2020)：

$$y = \alpha + \lambda G y + \gamma x + \epsilon, \epsilon_i = \sigma_\epsilon^2$$

其中，G是内生的。

内生网络的第一阶段估计

内生关系的简化模型（Fafchamps and Gubert, 2007)为

$$g_{ij} = \delta_0 + \delta_1 w_{ij} + \sum_k \delta_k |x_{k,i} - x_{k,j}| + \eta_{ij}, E(\eta_{ij}) = \sigma_\eta^2.$$

其中，$E(\epsilon_i \eta_{i,j}) = \sigma_{\epsilon\eta}$，$w_{ij}$表示外生网络（比如，同学关系）。

上述模型假定关系是独立的，忽略了网络结构特征。Fafchamps, Leij, and Goyal (2010)建议模型中加入(i, j)的最短距离，Graham (2015)建议加入(i, j)的共同邻居的数量:

$$g_{ij} = \delta_0 + \delta_1 w_{ij} + \sum_k \delta_k |x_{k,i} - x_{k,j}| + \delta_2 s_{ij} + \eta_{ij}$$

Graham(2016)提出另外一种改进，在模型中加入(i, j)的固定效应:

$$g_{ij} = \delta_0 + \delta_1 w_{ij} + \sum_k \delta_k |x_{k,i} - x_{k,j}| + \zeta_i + \zeta_j + \eta_{ij}$$

内生网络的第二阶段估计

$$y = (I - \lambda G)^{-1}(\alpha + \gamma x + \psi\xi + \epsilon), \xi_i = \sum_{j \neq i} \eta_{ij}.$$

其中，$\psi\xi$为选择效应。

采用非线性最小二乘估计。但在两步估计中，第二步中的标准差是错误的，采用自举法得到第二步估计量的分布。

假定模型的结构误差ϵ是i.i.d., 但简化误差$u = (I - \lambda G)^{-1}\epsilon$不是i.i.d., 因此不能直接对$u$做自举，而需要对$\epsilon$做自举。

由$u = \lambda G u + \epsilon$，$\epsilon = (I - \lambda G)u$。设模型的拟合值为$\hat{y}$，残差为$\hat{u}$，那么
$\hat{\epsilon} = (I - \lambda G)\hat{u}$,
对$\hat{\epsilon}$自举得到ϵ^*，进而得到u^*和自举的因变量$y^* = \hat{y} + u^*$。

自举R次，即得到估计量的近似分布和标准差。

Stata: 内生网络的第一阶段估计

`netregress` *mata-matrix*, [`gen` (*newvar*) `frame` (*string*) `framepbeta` (*strin*) `iters` (50) `equal` (*varlist*) `absdiff` (*varlist*) `diff` (*varlist*) `sum` (*varlist*) `product` (*varlist*) `ratio` (*varlist*) `logdiff` (*varlist*) `receiver` (*varlist*) `sender` (*varlist*) `partner` `dist`]

其中，*mata-matrix*为mata中的邻接矩阵（可以由 `snimport` 从Excel导入到mata）.

`gen` 为生成残差矩阵的行和（即 `snregenet` 内生网络回归中的偏差修正项）

`frame` 指定数据框保存由网络和变量转换过来的所有向量， `framepbeta` 指定数据框保存随机置换估计系数， `iters` 为随机置换次数。

`equal` ， `absdiff` 等设定变量转换的函数。在Fafchamps and Gubert(2007)全部采用 `absdiff` 函数。用户也可以定义自己的函数。

`partner` ， `dist` : 分别表示(i,j)共同邻居的数量(Graham, 2015)，(i,j)的最短距离(Fafchamps, Leij, and Goyal, 2010).

Stata: 内生网络的两阶段估计

`snregenet` *varlist* , `weight` (*matrix-list*) [`vbias` (*varname*) `bs` (50)]

其中， `weight` 为 `spmatrix` 定义的矩阵； `vbias` 为偏差修正项，由 `netregress` 生成的修正变量； `wmata` 为mata中与 `weight` 相同的权重矩阵，如果不设定，那么程序在mata复制 `weight` 矩阵。 `bs` 为自举次数。

如果没有设定 `vbias` ，或者 `bs(0)` ，那么 `snregenet` 执行空间自回归模型。

```
. netregress W alumni, equal(gender - min_leader) iters(0) gen(eta)  partner dist
Net regression (Quadratic assignment procedure with permutation s.e. Iters = 0):
```

W	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
partner	0	(omitted)				
dist	-.0046254	.0000201	-229.60	0.000	-.0046648	-.0045859
alumni	.0030896	.000373	8.28	0.000	.0023584	.0038207
gender	.0000101	.0000221	0.46	0.646	-.0000331	.0000534
nowhite	-.0000817	.0000262	-3.11	0.002	-.0001332	-.0000303
party	-.0000102	.0000203	-0.50	0.615	-.0000499	.0000295
seniority	-.0000669	.0000322	-2.08	0.038	-.00013	-3.74e-06
margin	-.0000988	.0007631	-0.13	0.897	-.0015945	.0013969
dw	-.0000525	.0002536	-0.21	0.836	-.0005496	.0004446
deleg_size	-.000069	.0000391	-1.76	0.078	-.0001456	7.66e-06
nchair	-.0000636	.0000358	-1.77	0.076	-.0001338	6.65e-06
maj_leader	1.44e-06	.0000529	0.03	0.978	-.0001023	.0001052
min_leader	-.00002	.0000424	-0.47	0.637	-.0001031	.0000631
_cons	.0094013	.0000838	112.16	0.000	.009237	.0095656

```
. regenet les gender - min_leader, weights(Wmat) nolog vbias(eta) bs(50)
Social (spatial) regression with endogenous weight matrix (bootstrap s.e. = 50)
..................................................+

------------------------------------------------------------------------------
         les | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ystar        |
      gender |   .0392018   .1363194     0.29   0.774    -.2279792    .3063829
    nowhite  |   .0262985   .1700493     0.15   0.877     -.306992     .359589
       party |  -.7929605   .1457345    -5.44   0.000    -1.078595   -.5073262
   seniority |   .0386345   .0114976     3.36   0.001     .0160997    .0611694
      margin |  -.3174639   .2437449    -1.30   0.193    -.7951952    .1602674
          dw |   -.174865   .2701505    -0.65   0.517    -.7043502    .3546202
  deleg_size |  -.0061762   .0034976    -1.77   0.077    -.0130315     .000679
      nchair |   2.159525   .2845548     7.59   0.000     1.601808    2.717242
  maj_leader |   .5116833   .3091746     1.65   0.098    -.0942877    1.117654
  min_leader |   -.081104   .3335814    -0.24   0.808    -.7349116    .5727036
         eta |   .0132462   .2254203     0.06   0.953    -.4285694    .4550618
       _cons |   1.403505   .2625516     5.35   0.000     .8889133    1.918096
-------------+----------------------------------------------------------------
Wmat         |
       ystar |  -.0107856   .1740404    -0.06   0.951    -.3518986    .3303273
------------------------------------------------------------------------------
```

异质性：权重网络分解

设个体分为两组（比如，两个党派、弱连接与强连接等），相应地G分为

$$G=\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}=\begin{bmatrix} G_{11} & 0 \\ 0 & G_{22} \end{bmatrix}+\begin{bmatrix} 0 & G_{12} \\ G_{21} & 0 \end{bmatrix}=G_1+G_2.$$

模型为

$$y=(I-(\lambda_1 G_1+\lambda_2 G_2))^{-1}(\lambda x+\epsilon).$$

其中，λ_1提现了组内的同行效应，λ_2提现了组间的同行效应。

Thank you !